SOFIA UNIVERSITY "ST. KLIMENT OF OHRID"

FACULTY OF MATHEMATICS AND INFORMATICS

PREDICTING THE SUCCESS OF TECHNOLOGY STARTUPS

Models, Empirical Evidence, and Software Tools

English Edition, 2026

Original dissertation title (2015): Methodology for Bulgarian Technological Start-ups Success Prediction

Models and a Software Application for Success Prediction of Start-up Companies

DOCTORAL DISSERTATION

submitted for the award of the educational and scientific degree of Doctor
in the academic specialty of Informatics,
professional field: Technological Entrepreneurship

Doctoral Candidate: Boyan Borislavov Yankov

Research Supervisor: Assoc. Prof. Petko Ruskov Ruskov, PhD

Sofia, 2015

PREFACE TO THE ENGLISH EDITION (2026)

This volume is the 2026 English edition of research conducted, submitted, and published in Bulgarian as a doctoral dissertation in Sofia in 2015. It is an English-language research monograph—not a new study undertaken in 2026 and not merely a literal translation.

Its edition title—Predicting the Success of Technology Startups: Models, Empirical Evidence, and Software Tools—expresses the work's integrated contribution for an international readership. The original title remains on the title page and in the front matter: Methodology for Bulgarian Technological Start-ups Success Prediction: Models and a Software Application for Success Prediction of Start-up Companies.

All interviews, survey data, literature, models, software implementations, technical references, and conclusions belong to the 2015 research context. They have not been updated to describe the startup ecosystem, technologies, markets, or scholarship of 2026. Time-sensitive statements should therefore be read historically. This edition changes the language and editorial framing, not the provenance of the evidence or findings.

The monograph brings together a startup-success prediction methodology, theoretical and empirically tested models, evidence from Bulgarian companies and entrepreneurs, and a software tool implementing the approach. Bulgaria is the empirical setting; readers may consider the findings comparatively in other entrepreneurial contexts.

The translation was prepared in 2026 with artificial-intelligence assistance with language transfer, terminology consistency, and translation of text embedded in figures. The author reviewed and edited the English text. AI did not conduct the original research, collect or generate evidence, formulate the 2015 findings, or independently revise the scholarly claims.

For citation and provenance, this volume should be identified as the 2026 English edition based on the doctoral dissertation originally submitted and published in Bulgarian in 2015. The Bulgarian dissertation remains the authoritative record of the original work.

Boyan Yankov

Sofia, 2026

CONTENTS

PREFACE TO THE ENGLISH EDITION (2026) **2**
CONTENTS **3**
LIST OF THE FIGURES **7**
LIST OF TABLES **10**
TERMS, ABBREVIATIONS, AND CONVENTIONS **11**
CHAPTER 1: INTRODUCTION **12**
1.1 Problem Definition 12
1.2 Aim and Objectives of the Dissertation 13
1.3 Structure of the Dissertation 13
CHAPTER 2: REVIEW OF THE PROBLEM AND PREVIOUS SOLUTION ATTEMPTS **15**
2.1 Facts, Trends, Causes, and Relevance 15
2.2 Definitions of Key Concepts 17
2.2.1 Definition of technological entrepreneurship 18
2.2.2 Definition of a start-up company 20
2.2.3 Definition of a successful start-up company 20
2.3 Selection of the Research Methodology 23
2.4 Proposed Solutions, Achievements, and Gaps 24
2.4.1 Comparison of existing theoretical models for predicting the success of start-up companies 25
2.4.2 Existing theoretical models for predicting the success of start-up companies based on Sandberg's model 29
2.4.3 Research and analysis of software products for predicting the success of start-up companies 31
2.4.3.1 StartupNation 32
2.4.3.2 Compass for New Companies 33
2.4.3.3 start-up Checker by Blueprint 35
2.4.3.4 Conclusions 35
2.5 Review of Data-Mining Software 37
2.5.1 Overview of IBM SPSS Statistics 37
2.5.2 Overview of IBM SPSS Modeler 37
2.5.3 Overview of Weka 37
2.5.4 Overview of HP Vertica 38
2.5.5 Overview of Microsoft SQL Server 38
2.5.6 Findings and Conclusions 39
2.6 Selection of Technologies and Tools for Information-System Development 39
2.6.1 Selection of application type 39
2.6.2 Selection of an operating system 39
2.6.3 Choosing a Web Server Application 40
2.6.4 Selection of server-side development tools 41

2.6.5 Customer Selection of Development Software 41
2.6.6 Selection of databases 42
2.6.7 Selection of Development Environment 42
2.7 Findings and Conclusions 43
CHAPTER 3: DEVELOPMENT OF A SUCCESS-PREDICTION MODEL 44
3.1 Model of the Company-Formation Process 44
3.2 Abstract Model for Predicting the Success of Start-up Companies 45
3.3 Detailed Model for Predicting the Success of Start-up Companies in Bulgaria 47
3.4 Findings and Conclusions 51
CHAPTER 4: RESEARCH, APPLICATIONS AND EXPERIMENTS 53
4.1 Qualitative study of the success factors of start-up companies 53
4.2 Quantitative Study of the Success Factors of Start-up Companies 56
4.2.1 Analysis of the Quantitative-Study Data 58
4.2.1.1 Size of companies 58
4.2.1.2 Age of the companies 59
4.2.1.3 Industrial sector of the companies 59
4.2.1.4 Gender of respondents 60
4.2.1.5 Age of the respondents 61
4.2.2 Preparation of the Quantitative-Study Data 61
4.2.2.1 Preparation of Survey Data in Excel 62
4.2.2.2 Preparation of survey data in IBM SPSS Statistics 65
4.2.2.3 Preparing survey data in Weka 66
4.2.3 Grouping success factors 66
4.2.4 Data Mining and Success-Model Development with IBM SPSS Modeler 72
4.2.4.1 Synthesizing models for predicting success 72
4.2.4.2 Development of a Cross-Validated C5.0 Model 74
4.2.4.3 Comparison of algorithms for obtaining models 80
4.2.4.4 Comparison of the obtained success prediction models 84
4.2.5 Data Mining and Success-Model Development with Weka 84
4.2.5.1 Synthesizing success prediction models 84
4.2.5.2 Comparison of Classification Algorithms 87
4.2.5.3 Comparison of the resulting success prediction models 88
4.2.6 Development of Success-Prediction Models for Technology Companies 90
4.2.7 Comparison of the Resulting Models 95
4.3 Results, Findings, and Conclusions 96
CHAPTER 5: IMPLEMENTATION OF SUCCESS-PREDICTION SOFTWARE 97
5.1 Requirements for Start-up Success-Prediction Software 97
5.2 Design of the Start-up Success-Prediction Software 97
5.2.1 General application architecture 98
5.2.2 Application modules 100
5.2.2.1 Login module 101

5.2.2.2 Administration module 101
5.2.2.3 Setup Module 101
5.2.2.4 Survey Data Module 101
5.2.2.5 Analysis and Forecast Module 102
5.3 Development of the Success-Prediction Software 102
5.3.1 Prediction Algorithm 103
5.3.1.1 Dictionary file 105
5.3.1.2 Regular Expressions 106
5.3.2 Success-Prediction Results Graph 106
5.3.3 Google APIs 108
5.3.3.1 Google Drive API 108
5.3.3.2 Google Sheets API 108
5.3.3.3 Google Charts API 109
5.3.4 Data Storage 109
5.3.4.1 Database structure 109
5.3.4.2 Local files 111
5.3.4.3 Files Located on Google Drive 111
5.3.5 Functionality Using Google APIs 112
5.3.5.1 Authentication 112
5.3.5.2 Retrieve User Profile Information 113
5.3.5.3 Working with Google Drive 113
5.3.6 Design templates 115
5.3.7 Security and reliability 115
5.3.7.1 Compliance with good practices 115
5.3.7.2 Data Validation, Filtering and Preparation 116
5.4 Testing the Success-Prediction Software 116
5.4.1 System login 116
5.4.2 Application Settings 119
5.4.3 Viewing the contents of a survey file 120
5.4.4 Success prediction 121
5.4.5 Administration Module 123
5.4.6 Results, Findings, and Conclusions 125
5.5 Evaluation and Validation of the Software Product 125
5.6 Results, Findings, and Conclusions 127
CONCLUSION 128
MAIN CONTRIBUTIONS OF THE DISSERTATION 129
PUBLICATIONS RELATED TO THE DISSERTATION TOPIC 131
REFERENCES 132
DECLARATION OF ORIGINALITY 139
APPENDICES 140
APPENDIX 1: GLOSSARY 140

APPENDIX 2: RESULTS OF ANALYSIS WITH DATA MINING SOFTWARE APPLICATIONS 144
Results of factor analysis with the software product IBM SPSS Statistics 144
Results of synthesizing success prediction models with the software product IBM SPSS Modeler 153
start-up Success Prediction Model 153
A model for predicting the success of technology start-up companies 156
Program code of the algorithm C5.0 159
Model synthesis results for predicting success with the software product Weka 165
APPENDIX 3: PROGRAM CODE OF THE INFORMATION SYSTEM I3SP 168
GoogleService class code: 168
AnalyseService class code: 171
UserService class code: 178
Code of frontend template "\templates\admin_manageusers.tpl": 183
Program code of frontend template "\templates\survey_data_analyse.tpl": 184
Exported information about the structure of the database in the format . SQL: 187

LIST OF THE FIGURES

Fig. 1 Number of publications and interest on the topic of entrepreneurship (created by the author) 16
Fig. 2 Time frame of Lean start-up methodology development 18
Fig. 3 The relationship between technological change and entrepreneurship (created by the author as described by Tony Bailletti) 19
Fig. 4 New product entry into the market - jumping the "gap" between the early and the main market (adapted from Tony Bailletti) 21
Fig. 5 Cash Flow J-Curve of a technology start-up (Adapted from Grabenwarter) 22
Fig. 6 Four-level decomposition of the CRISP-DM methodology (translated by the author from Chapman's figure) 23
Fig. 7 Life cycle of a data mining project using the CRISP-DM methodology (translated by the author from Chapman's figure) 24
Fig. 8 Results of the Risk Calculator for start-ups prediction tool (created by the author by implementing the tool) 33
Fig. 9 Fragment of results for prediction of success of the tool "start-up Compass" (created by the author by applying the tool) 34
Fig. 10 Success prediction result of the вЂњThink BlueprintвЂќ tool (created by the author by applying the tool) 35
Fig. 11Selecting an application from the graphical user interface of Weka (screenshot created by the author) 38
Fig. 12 Model of the process of creating a new firm (translated by the author according to Joan and James Carland) 44
Fig. 13 Supplemented model of the process of creating a new company (created by the author) 45
Fig. 14 A model for predicting the success of start-up companies (created by the author) 46
Fig. 15 Part of the survey for the validation of the model for predicting the success of start-up companies (created by the author) 58
Fig. 16 Size of companies (created by the author) 59
Fig. 17 Age of companies (created by the author) 59
Fig. 18 Industrial sector of companies (created by the author) 60
Fig. 19 Gender of respondents (created by the author) 61
Fig. 20 Age of respondents (created by the author) 61
Fig. 21 Sample of survey data viewed in Google Drive (author created) 62
Fig. 22 Sample data from the survey, after processing in Microsoft Excel (created by the author) 64
Fig. 23 Sample survey data entered into IBM SPSS Statistics (author created) 65
Fig. 24 Sample of survey data related variables in IBM SPSS Statistics (created by the author) 65
Fig. 25 Sample data loaded into the Weka software product (created by the author) 66
Fig. 26 Sequence of actions in IMB SPSS Modeler for obtaining classification models - graphical representation (created by the author) 72
Fig. 27 Selection of fields for the automatic classifier - input data and objectives (created by the author) 73
Fig. 28 Best obtained models for predicting success in automatic classification (created by the author) 74
Fig. 29 Accuracy assessment of a model obtained by the C5.0 algorithm and applying 75

cross-validation (created by the author)
Fig. 30 The distribution of companies correctly and incorrectly classified by the model (created by the author) 76
Fig. 31 Overall graphical representation (map) of the classification tree (created by the author) 78
Fig. 32 Classification tree - first level: Existence of a clear competitive advantage (created by the author) 79
Fig. 33 Classification tree - branch 2: the external environment is a key factor for the company's success (created by the author) 79
Fig. 34 Classification tree - branch 2.2: intangible assets - established business reputation of the company (created by the author) 80
Fig. 35 Classification tree - branch 2.2.2: way of entering the start-up company in the market (created by the author) 80
Fig. 36 Graphic representation of a classification tree (translated by the author) 81
Fig. 37 Classification algorithms in Weka (created by the author) 85
Fig. 38 Setting up the J48 algorithm in Weka and creating a classification model (created by the author) 86
Fig. 39 Schematic visualization of the classification tree generated with the J48 algorithm in Weka (created by the author) 89
Fig. 40 Best received models for predicting success of technology companies in automatic classification (created by the author) 90
Fig. 41 Accuracy assessment of a model for predicting the success of technology companies obtained by the C5.0 algorithm and applying cross-validation (created by the author) 91
Fig. 42 Overall graphical representation (map) of the classification tree for technology companies (created by the author) 92
Fig. 43 Classification tree - first level: Most companies in the industry have a good profit (created by the author) 93
Fig. 44 Classification Tree - Branch 1: Key Success Factors: Price (created by author) 93
Fig. 45 Classification Tree - Branch 1.2: Key Success Factors: Marketing and Advertising (created by author) 94
Fig. 46 Classification tree - branch 1.2.2: the chosen strategy is aggressive. (created by the author) 94
Fig. 47 Data flow diagram in I3SP (created by the author) 98
Fig. 48 Three-layer model Model - View - Controller for a web-based application implemented with the PHP programming language and MySQL database (adapted by the author from a figure from betterexplained.com) 99
Fig. 49 Schematic representation of the architecture and functionalities of the application (created by the author) 100
Fig. 50 UML diagram of actions in predicting the success of a start-up company with I3SP (created by the author) 103
Fig. 51 Schematic of the application prediction algorithm (created by the author) 104
Fig. 52 Fragment of the dictionary file used in the application (created by the author) 105
Fig. 53 Regular expression for extracting data from a file containing a classification tree (created by the author) 106
Fig. 54 Fragment of the classification tree presented in the analysis page (created by the author) 107
Fig. 55 Application database structure diagram (created by the author) 110
Fig. 56 Fragment of an IBM SPSS Modeler source file (created by the author) 111
Fig. 57 Fragment of Questionnaire file (created by the author) 112

Fig. 58 Code snippet of a design template in Smarty (created by the author) 115
Fig. 59 Login to the application with a Google account (created by the author) 117
Fig. 60 Dialog box for allowing access to Google Drive (created by the author) 118
Fig. 61 Waiting for approval after sending a request to access the system (created by the author) 118
Fig. 62 Settings page in the application (created by the author) 119
Fig. 63 Dialog window for selecting a survey file from Google Drive (created by the author) 119
Fig. 64 Page with a preview of the contents of a survey file (created by the author) 120
Fig. 65 Fragment of the results page of a company success forecast (created by the author) 121
Fig. 66 Saved results of analyzes (created by the author) 123
Fig. 67 User Management (created by author) 124
Fig. 68 Uploading a new algorithm (classification tree) to the system (created by the author) 124

LIST OF TABLES

Table 1 Definitions of used terms and abbreviations 11
Table 2 Search queries and results in scientific databases 15
Table 3 Number of publications on entrepreneurship and predicting the success of start-up companies 16
Table 4 Comparison of models for predicting the success of start-up companies 25
Table 5 Comparison of models for predicting success based on the model of Sandberg 29
Table 6 Summary comparison of analyzed success prediction tools 35
Table 7 Comparison of web server operating systems 40
Table 8 Summary results of the qualitative survey of opinions regarding the proposed prediction model success 54
Table 9 Kaiser-Meyer-Olkin and Barlett test 67
Table 10 Correlation matrix, after applying rotation 67
Table 11 Comparison of categories from the theoretical model for predicting success and the resulting categories from the factorial analysis 70
Table 12 Significance of components from the correlation matrix 71
Table 13 Comparison of classification models obtained with Weka 86
Table 14 Comparison of models for predicting start-up success companies 95
Table 15 Testing results of the software product I3SP 125
Table 16 Main contributions of the dissertation 129

TERMS, ABBREVIATIONS, AND CONVENTIONS

Table 1. Definitions of Terms and Abbreviations

Term / Abbreviation	Definition
I3SP (Information System for Start-ups Success Prediction)	In translation "information system for predicting the success of start-up companies", software an information system developed for this dissertation that serves to predict the probability of success of start-up companies and visualize a classification tree with the success factors.
Lean start-up	Methodology for the development of start-up companies, introduced by Eric Ries, which aims to emphasise the processes in the company, through the iterative development of products and meeting the needs of different consumer segments
NCIIA (The National Collegiate Inventors and Innovators Alliance)	Translated from "National University Association of Inventors and Innovators", a US organization that trains entrepreneurs and funds student companies.
NVP (New Venture Performance)	Translated from "start-up productivity", the end result of the process of creating and organizing a new business that develops, creates and markets products and services to satisfy market needs for profit and growth.

Frequently used popular terms and abbreviations that occur in the dissertation for submission in Appendix **1: Glossary.**

CHAPTER 1: INTRODUCTION

In the transition economy conditions of Central and Eastern Europe, start-ups are heavily relied upon to generate the conditions for economic growth. In the priorities of the program period of the European Union from 2014 to 2020, "Horizon 2020", the need for the development of innovative small and medium-sized businesses, with the aim of generating economic growth and job positions (European Commission, 2014), is laid down. Entrepreneurs and their start-ups increase the competitiveness of transition economies through their innovative strategies and business models, sometimes controversial practices, and their unrelenting energy to succeed (Peng M. W., 2001). Business also needs start-ups that offer innovative solutions to support the development of the various industrial sectors in the Bulgarian economy. Moreover, they are a good investment opportunity with potentially high returns.

The presence of small technology companies, as well as larger corporate entrepreneurial teams, in the market is associated with technological change (Jelinek, 1996). Both small companies, which are the predominant number and strive for efficiency, and medium and large companies, developing corporate entrepreneurship and developing large and ambitious projects, contribute to this.

At the moment, several business incubators, venture capital funds, as well as programs, financed by the government and the European Union to support small and medium enterprises. This trend shows the importance and focus of the government, the European Union and business towards the development of start-up companies in Bulgaria. The results of this activity are still on a small scale and most entrepreneurs start on their own, often without having the necessary experience, information and support, which sometimes leads to failures and is often the cause of inefficiency, missed opportunities and loss of investments made (Avramov, 2012). In Bulgaria, the image of entrepreneurs is distorted, they are expected to create jobs, and the development of a valuable and competitive product or service remains in the background. Only one in five wants to become an entrepreneur, which is twice less than the average for Europe (Stoycheva, 2013).

1.1 Problem Definition

Predicting the success of start-ups is an opportunity to increase the efficiency of the business start-up process and to reduce risk and resource consumption. The success of a company is tied to many factors, such as the economy, employment, education, innovation and added value of the new products or services that the company will develop, the company strategy, the qualities of the team and others. Models have been created to predict the success of start-up companies, which are specific to different markets, countries, economic features, as well as sometimes to the momentary situation of their creation. Unfortunately, in Bulgaria there are not many developed models for predicting the success of start-up companies and almost no software to apply these models and make them available to interested parties. Bulgaria is a small ecosystem and the business climate is specific with its stage of development, regulations, location, psychology, etc. This gives rise to the need to create a prediction model adapted specifically for the local market, which will serve as the basis of software for predicting the success of start-up companies in Bulgaria.

A model for predicting the success of Bulgarian technology start-upswould benefit entrepreneurs, business owners, business incubators, schools and universities, training students in entrepreneurship, business consultants, business analysts, investors and funds for venture capital. The model can be used both to predict the success of start-up companies and to identify the factors that led to this result, to identify possible strengths and weaknesses of the company, to improve the general knowledge of the interested parties about the company,

to measure its progress over a period of time, and to make comparisons between individual companies.

Based on a model for predicting the success of start-up companies, one can develop **software to implement the model.** Such software would be beneficial both to the entrepreneurs themselves, starting their new business, and to all parties interested in the process. With the use of the software, it is expected that the risks for investors in start-up companies arising from the loss of capital in the event of failure will be less, as it will be easier to analyze the state of a particular start-up business. Based on company data obtained from a special survey, the software will provide an early stage prediction of the chances of success and failure. It could also be used to analyze a start-up company or an existing one, giving indications of where and what needs to be changed in the company in order to obtain higher predictive results, which at a later stage will lead to a better business result.

1.2 Aim and Objectives of the Dissertation

The main goal of this dissertation work is to create a model, methodology and prototype of an information system for predicting the success of start-up companies and technological start-up companies from Bulgaria, which will help make management and investment decisions.

To achieve the set goal defines the **following** tasks:

1. To study, research and analyze the current state in our country and the world in libraries, scientific databases, scientific works and the Internet of the models for predicting the success of start-up companies, the success factors of start-up companies, the software products for extracting knowledge from data and their application in classification models for predicting success.
2. To prepare a theoretical model for predicting the success of start-up companies from Bulgaria, which will be adapted for the environment and for Bulgarian entrepreneurs through qualitative research.
3. To validate the theoretical model through quantitative research, and through software products for extracting knowledge from data to prepare models for predicting success, adapted for companies from Bulgaria. To prepare a model for predicting the success of technological start-up companies.
4. To design and implement a prototype of an information system for predicting the success of start-up companies, which will allow applying the models obtained in the research and future upgrading and improvement of the algorithm based on the collected data.

1.3 Structure of the Dissertation

The text of this dissertation is organized into 5 chapters, list of figures, list of tables, main terms used, conclusion, bibliography and 3 appendices.

Chapter 1 briefly introduces the problem area and provides a general overview of the problem of the lack of models for predicting the success of start-up companies adapted for the Bulgarian market. This chapter clarifies the importance of the problem. The goals and objectives of this dissertation are determined - creation of a methodology, models and software tool for predicting the success of start-up technological companies from Bulgaria.

Chapter 2is a detailed overview of the problem area and the already known methods, approaches and models for solving tasks similar to the task of this dissertation. This chapter defines the terms technology entrepreneurship, start-up and successful start-up. Software products for predicting the success of start-up companies, which are the motivation for

developing the software for this dissertation, are reviewed and compared. Data mining software products to be implemented in the study were explored and selected. The development technologies that will be used in the implementation of a software system have been analyzed and selected.

Chapter 3presents the methodology and models developed by the author for predicting the success of companies from Bulgaria. Theoretical abstract models of the process of creating a company and of a successful company are proposed. The model of a successful company is developed in detail, describing the specific success factors. The proposed models represent hypotheses that have been tested through qualitative and quantitative research described in Chapter 4.

Chapter 4 presents research and experiments based on the developed theoretical model of start-up company success from Chapter 3. The model is improved by conducting qualitative research - interviews with entrepreneurs and company owners. Then, it is validated using quantitative method research and application of data mining methodology using various software products. A factor analysis was carried out, which confirmed the proposed categories of factors for the success of start-up companies. Success prediction models were created and their accuracy was verified using the cross-validation method. The resulting models are compared and the algorithms used to synthesize them are discussed. The key success factors of start-up companies from Bulgaria have been identified.

Chapter 5is a detailed description of the implemented software for predicting the success of start-up companies. The requirements for the software product were reviewed and the development technologies were selected. The overall architecture of the application is presented and its main modules and used external libraries are described in detail. Results of testing and application of the software in practice are presented.

CHAPTER 2: REVIEW OF THE PROBLEM AND PREVIOUS SOLUTION ATTEMPTS

This chapter presents a detailed literature review of the problem area and of the already known methods, approaches and models for solving the task of predicting the success of start-up companies. The concepts of technological entrepreneurship, start-up company and successful start-up company are defined. Software products for predicting the success of start-ups are reviewed and compared. Data mining software applications to be used in the study were explored and selected. The development technologies that will be applied in the implementation of a software system have been analyzed and selected.

2.1 Facts, Trends, Causes, and Relevance

The present study begins with a literature review. Literature sources were studied through searches in scientific databases (Emerald, 2014), (Elsevier, 2014), (ProQuest, 2014), (Google Inc., 2014), as well as through specific searches in Internet search engines. A total of over 1500 documents matching the keywords and search queries were reviewed (Table 2).

Table 2Search queries and results in scientific databases

Search query	**emeraldinsight.com**	**www.sciencedirect.com**	**dissexpress.umi.com/dxweb**	**scholar.google.bg**
technology entrepreneurship prediction	425 result	3 146 the score	40+ the score	36 200 the score
entrepreneurship success prediction	426 the result	2 765 the result	40+ the result	41 900 result
In the title: entrepreneurship success	0 the result	7 the result	14 the result	-
In the title: entrepreneurship prediction	0 the result	1 the result	2 the result	-

Based on the reviewed literature sources, hypotheses, models, methodologies, research and tools for predicting the success of start-up companies were analyzed and compared.

Based on the research done, I found that **entrepreneurship** is a popular applied science topic and research area with over 40,000 publications annually worldwide (Google Inc., n.d.). It keeps its popularity stable in the period 2005-2014, which can be seen from the upper curve of the attached graph - Fig. 1generated through the Google Trends service. The lower curve shows the volume of demand for entrepreneurship as a keyword in the period 2005-2014. The curve shows that public interest in the topic is stable.

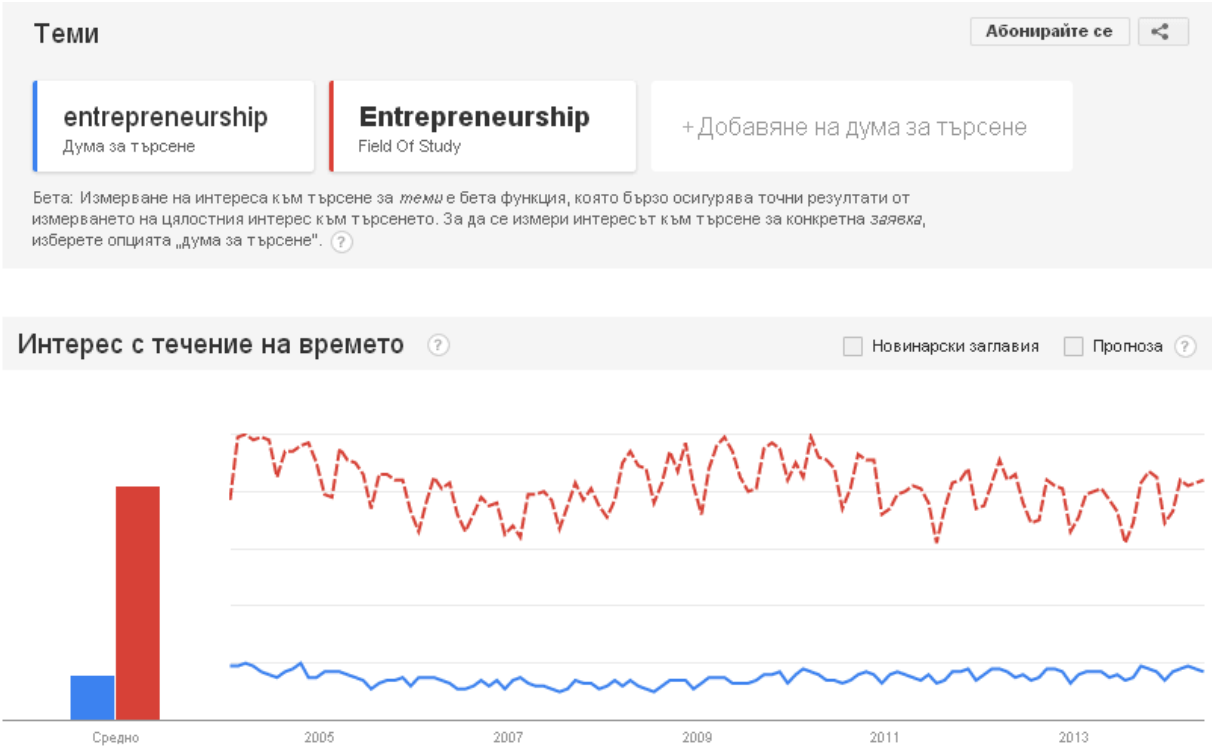


Fig. 1 Number of publications and interest on the topic of entrepreneurship (created by the author)

The annual number of publications on the topic of entrepreneurship on a global scale grew rapidly in the period 1990-2007, increasing more than 10 times (Table 3). In the period 2008-2013, the number of publications on the topic of entrepreneurship stabilized between 40 and 50 thousand per year, and between 2011-2013 it marked a small decline, which is probably due to saturation already reached, since interest in the topic remains relatively constant. The topic of entrepreneurship will continue to be relevant in Europe, as it lies in the priorities of the European Union for the program period from 2014 to 2020 "Horizon 2020". The emphasis will be on innovation within small and medium-sized businesses.

Table 3Number of publications on entrepreneurship and predicting the success of start-ups

Year	Number of publications on entrepreneurship	Number of articles on predicting the success of start-up companies
1990	3290 ↑	349 ↑
1991	3500 ↑	425 ↑
1992	3870 ↑	499 ↑
1993	4310 ↑	512 ↑
1994	4760 ↑	621 ↑
1995	4970 ↑	670 ↑
1996	5740 ↑	785 ↑
1997	6440 ↑	832 ↑
1998	7340 ↑	951 ↑

1999	8520 ↑	1140 ↑
2000	10700 ↑	1590 ↑
2001	12400 ↑	1510 ↓
2002	15400 ↑	1570 ↑
2003	18200 ↑	1860 ↑
2004	22500 ↑	1940 ↑
2005	27100 ↑	2150 ↑
2006	33400 ↑	2460 ↑
2007	37800 ↑	2670 ↑
2008	42600 ↑	2940 ↑
2009	46600 ↑	3150 ↑
2010	48500 ↑	3650 ↑
2011	47100 ↓	4050 ↑
2012	43600 ↓	4440 ↑
2013	41600 ↓	4630 ↑

The topic of predicting **the success of start-up companies** is not new, as there are publications from the 1970s (Wells, 1974). However, it is rapidly developing and gaining more and more popularity. From 1990 to 2013, the number of publications on an annual basis increased more than 10 times. The trend of increasing interest in the topic of predicting the success of start-ups has remained stable in recent years. In the period from 2008 to 2013, the number of publications on the topic grew one and a half times, compared to the relatively constant number of publications on the topic of entrepreneurship.

Models for predicting the success of start-ups evolve along with business and economic developments. There are different types of models depending on the application: general ones for evaluating start-ups; to evaluate companies participating in a business plan competition; to evaluate companies operating in a specific local market; to evaluate technology start-ups and others. Depending on the specific environment and conditions, the models have different characteristics.

This dissertation aims to create a comprehensive methodology and model for predicting the success of start-up companies in Bulgaria, taking into account the technological factor. The problem of creating this type of model is topical, because despite the interest in the subject of predicting success, no models have been prepared for Bulgaria. Such a model is needed by interested parties - owners of start-up companies, business analysts, venture capital funds and investors, educational institutions - training students to start companies, organizers of business plan competitions. Based on this model, software can be created that predicts the success of start-up companies from Bulgaria, and gives indications of what needs to be changed in the company in order to obtain better forecast results.

2.2 Definitions of Key Concepts

In this point, the definitions of basic concepts used in the dissertation are discussed labor, such as technology entrepreneurship, start-up, successful start-up. Definitions proposed by various authors are reviewed and the meanings that will be adopted in the dissertation are selected.

2.2.1 Definition of technological entrepreneurship

NCIIA defines entrepreneurship (NCIIA, 2007) as the process of organizing and taking risks in starting a business. Technology entrepreneurship is a combination of starting a business and applying technology and research.

The roots of technology entrepreneurship can be traced back to the Ancient Greek civilization (1450 BC) (Nelson, 1999). The ancient Greeks were among the first to build bridges and complex water and irrigation systems. Through technological advancements, they have demonstrated their advanced engineering skills. They are believed to be the first whose agricultural activity generated more produce than they could utilize, and the remainder was used for export.

The concept of technological entrepreneurship was first mentioned at a symposium in 1970 (Blank, The Four Steps to the Epiphany Steve Blank, 2005), (Bailetti, 2012) and has become a hot topic with an increasing number of published articles over the years. Technology entrepreneurship encompasses several areas, and one of them is the creation of new businesses. One of the most influential trends in technological entrepreneurship at the moment is "Lean start-up" (Ries, The Lean start-up - The Movement That is Transforming How New Products Are Built and Launched, 2014), (Ries, The Lean start-up: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses, 2011) (Fig. 2) - a methodology for the development of start-up companies, which aims to emphasize the processes in the company, through iterative construction of products and meeting the needs of different consumer segments.

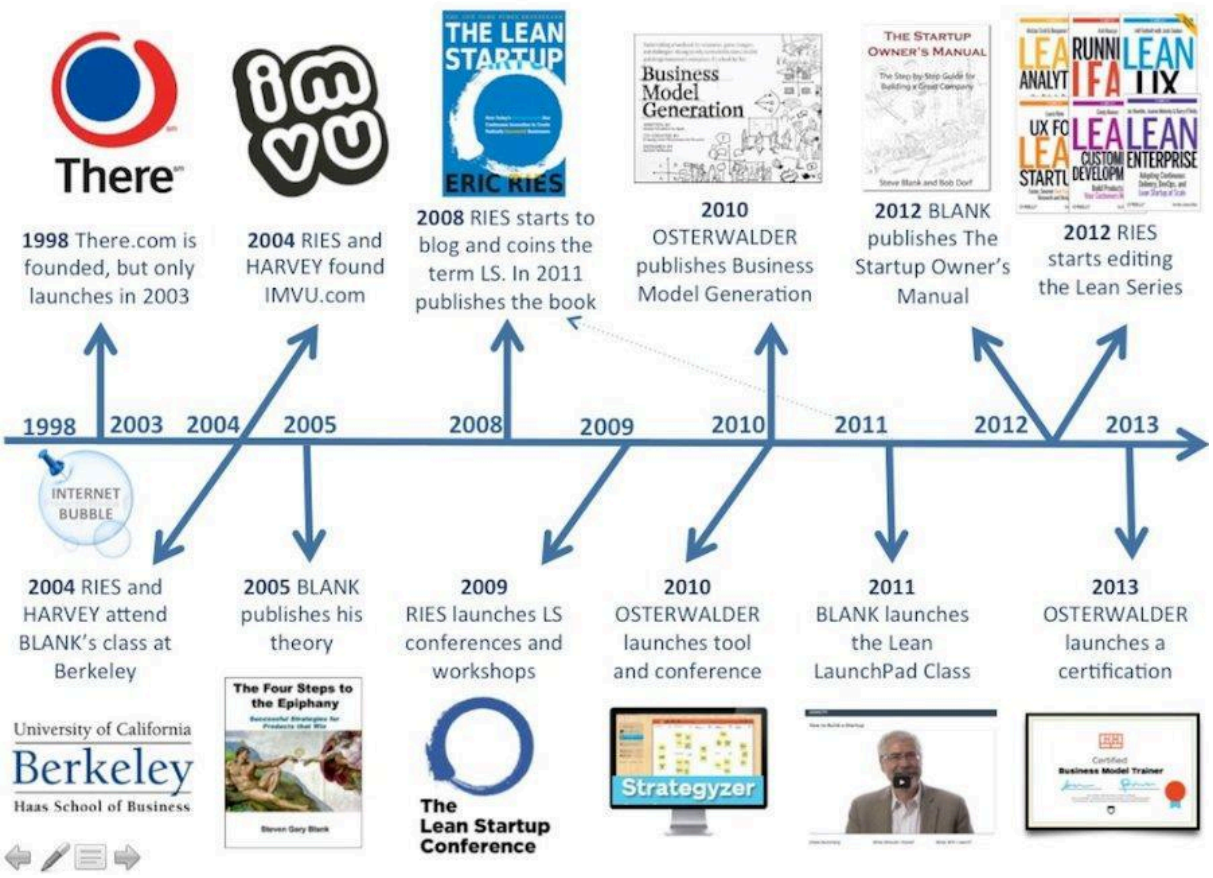


Fig. 2Lean start-up Methodology Development Timeline

There are many modern definitions of technology entrepreneurship. According to Tony Bailletti, technological entrepreneurship is the investment in a project that brings together specialists and heterogeneous assets, resulting in scientific and technological knowledge for the purposes of creating value for a company.

According to the NCIIA, technology entrepreneurship is the actions of the entrepreneur and team to bring the discovery on the path to market.

The Engineering Department of McGill University defines technological entrepreneurship as the innovative application of scientific and technical knowledge by one or more people who start and develop a business and take financial risks to achieve their vision or goal (McGill, 2014).

According to some authors, technological entrepreneurship is the process of creating new technological companies, which includes organizing, managing and dealing with the risk of the technological firm, bringing the invention to the market (Nicholas, 2003), (Florida, 2007). Technological entrepreneurship requires collaborative activities to interpret ambiguous data, to jointly understand and maintain technological activities, and a constant and coordinated effort to achieve technological change.

In Fig. 3shows activities and topics related to technological entrepreneurship described by Tony Bailletti.

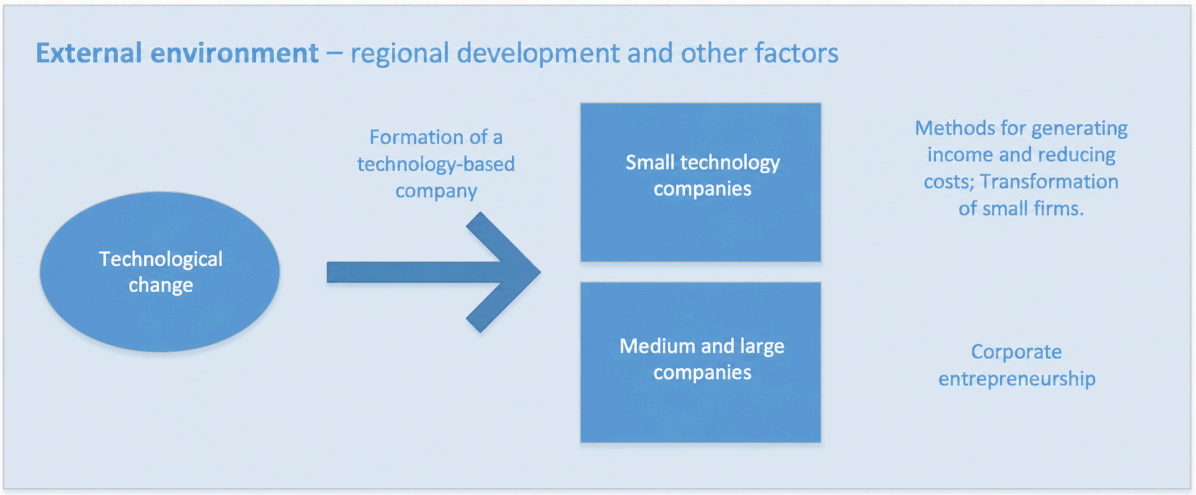


Fig. 3The relationship between technological change and entrepreneurship (created by the author as described by Tony Bailletti)

Based on the above definitions, for the purposes of this dissertation we assume that **technological entrepreneurship** is related **to scientific development or research activity, or innovative application** of existing technology or scientific and technical knowledge, resulting in **the creation of a product or service** that adds **value** to stakeholders and contributes **to business objectives.**

2.2.2 Definition of a start-up company

There are many definitions for the concept of a start-up company, most of which do not give concrete measures as to when a company ceases to be a start-up and becomes a mature company.

According to Eric Ries, a start-up company is a human institution designed to create a new product or service under conditions of extreme uncertainty (Ries, The Lean start-up, 2011). According to some authors, the founding of start-up companies is the result of the process of developing, producing and offering products or services that satisfy a certain market need for profit and growth (Gartner, 1988). According to Steve Blank, a start-up is a temporary organization created in search of a business model that can be executed and developed.

For the purposes of this dissertation, based on the listed definitions, we assume that a **start-up company** is one that is **at the beginning of its development and is SME**(micro, small or medium-sized) in size. For small and medium-sized companies, we accept those with 0to 250 employees (according to the definition applicable to the European Union).

2.2.3 Definition of a successful start-up company

The definition of start-up success is necessary to distinguish successful start-ups from unsuccessful start-ups. start-up success is a broad term that has been defined in multiple ways in the literature.

If we look in the Bulgarian interpretive dictionary (BAN, 2012), the meaning of the word success is defined as "1. Achieving a goal; progress, luck. Success in business. I am successful. I am not successful in love. Success in career. 2. Public recognition, approval, glorification. The film is a great success. The singer is touring with great Success. Degree of mastery, expressed in grades. 4. Success. In the first meaning of the word, success is the achievement of a certain goal, which also coincides with the definition of Brockhaus (Brockhaus, 1982): the fulfillment of certain goals (income, company growth, personal achievements, return on investment, and others). This shows that success **depends on whose goals we consider** - of the entrepreneur who starts a business; to the investor who invested money; to the bank that granted the loan; of the university that prepares entrepreneurs; of the government stimulating the economy; of the general public or one's other goals. In defining success, Keeley, for example, takes the side of investors and defines success as the internal rate of return realized by all shareholders (Keeley, 1989).

When choosing an appropriate definition of the success of start-up companies, we must consider their particularities. **Start-ups differ from established companies**by their younger age, smaller but often rapidly growing scale, and the uncertainty associated with them. For an established company, success could be measured purely financially, for example by profits or turnover from annual financial statements, by assets held and others, and these metrics would probably satisfy most stakeholders. With a start-up company, in the early stages of its development, there is usually a lot of investment in product development, and there is no profit and no return because there are no sales yet. The product or service is developed, the team accumulates experience and contacts, the market is studied, and step by step the company develops, but this does not give a positive impact on the income and expense report.

The development of a start-up company goes through different phases (Moore, 1991), from the point of view of product commercialization (Fig. 4). Gradually, different market segments are reached, which correspond to new types of consumers, whose requirements for the product are growing. In order to reach the mainstream market, it is necessary that the product be completed, accepted by the market and its shortcomings from the point of view of consumers be removed.

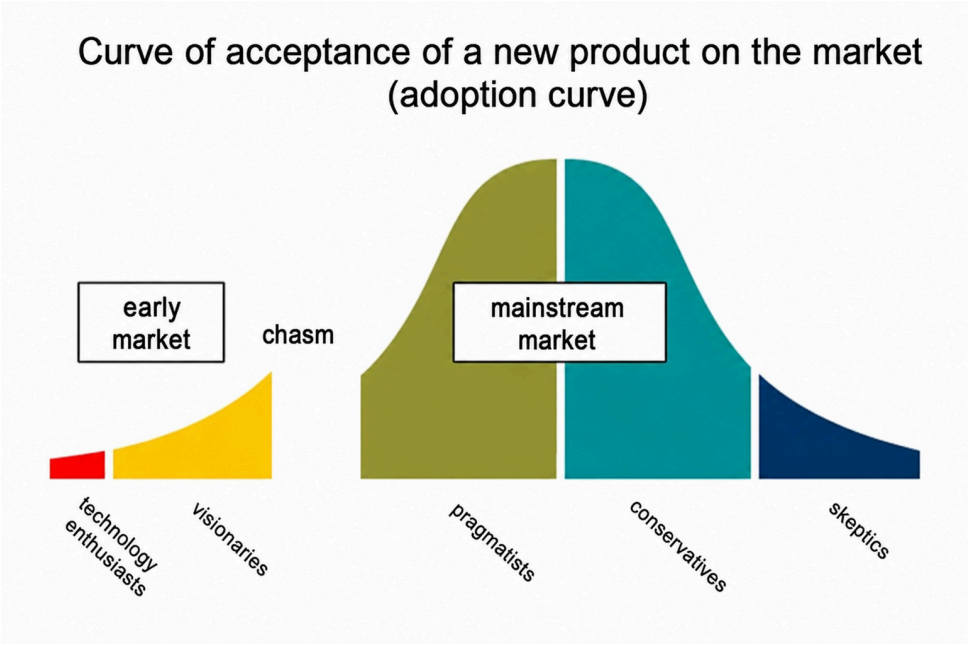


Fig. 4New product entry into the market - jumping the "gap" between the early and mainstream market (adapted from Tony Bailletti)

New product entry into the market is a long process associated with uncertainty and a lot of investment. This affects the cash flow (Grabenwarter, 2005) of start-up companies (Fig. 5), and to the greatest extent of technology companies. Because of the listed characteristics of start-ups, the success metrics that apply to mature companies are not directly applicable to start-ups.

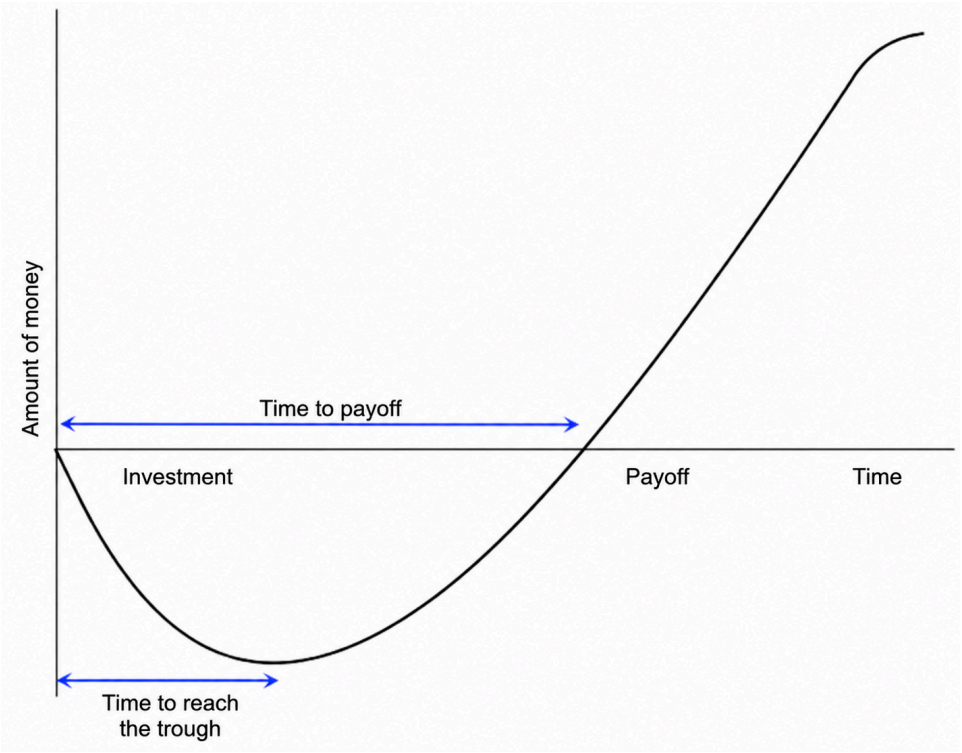


Fig. 5J-curve of cash flow of a technology start-up (adapted from Grabenwarter)

In order to determine the success criteria of start-ups, interviews **with business owners and entrepreneurswere** conducted. Respondents were asked the questions, "How do you determine which start-up is successful? How do you distinguish a successful start-up from a failed one?".

One of the interviewees of the author is Bob Dorf - one of the most popular experts on start-ups and corporate entrepreneurship in the world, a professor at Columbia Business University, a mentor, business consultant and entrepreneur. During the conversation, Bob Dorf responded as follows: "It is very difficult to put an exact line between which start-ups are successful and which are unsuccessful, and there is no universally recognized criterion. To me, a successful company is one that, if last quarter it had 10 customer inquiries and 2 sales, then this quarter it has 10 inquiries and 3 sales. It does not matterthestage of development of the company and the cash stream at the moment, as the company may still be at a very early stage. It is important to have development."

Interviews were also conducted with five founders and owners of Bulgarian start-up companies. Their answers can be summarized by saying that it is important for start-ups to develop and grow. While in mature companies, success can be considered purely financial result without increasing the size of the company, in start-ups development and growth are key to their further survival and existence.

Over 50% of start-up companies in Europe (as well as in Bulgaria) have between 0 and 4 employees (Eusostat, European Commission, 2009). Most start-ups fail to develop to a stage that ensures their survival during the first 5 years of operation and it follows logically that they are unsuccessful.

Based on the research and interviews conducted, **a successful start-up is considered to be one that has increased its size in the last five years**. A company that has ceased its activity

is considered unsuccessful. And a company that continues its activity, but has not increased its size, is considered neither successful nor unsuccessful.

2.3 Selection of the Research Methodology

For the purposes of the dissertation, various research methodologies were used. Their choice is based on the nature of the tasks being solved.

To solve the tasks related to research, research and analysis of the problem area - the current state of scientific works and achievements in the field of predicting the success of start-up companies, the success factors of start-up companies in the world and in Bulgaria, the method of **systematic literature research** was used (Tranfield, 2003). It is a type of literature review in which the existing knowledge regarding a specific question or problem area (in the form of so-called primary research) is identified, evaluated and synthesized in a way that ensures maximum objectivity, transparency and reproducibility.

The detailed model for predicting the success of start-up companies proposed in Chapter 3 is validated by applying the methods qualitatively and quantitatively research.

Qualitative research is conducting in-depth interviews with a small number of respondents in order to collect initial feedback on the model and make improvements.

Based on the detailed model for predicting success, **quantitative research**was conducted to determine the variables that have the greatest influence on the success of start-up companies, as well as the dependencies between individual variables. The quantitative research is based on the methodology CRISP-DM (Cross Industry Standard Process for Data Mining) (Chapman P., 2000), which describes a standard of the process of extracting knowledge from data (data mining) and was created by a consortium of companies as a project of the European Union in 1997. CRISP-DM is described as a hierarchical process model containing groups of tasks described at four levels of abstraction (from general to particular): phases, general tasks, specialized tasks and process instances, as shown in Fig. 6.

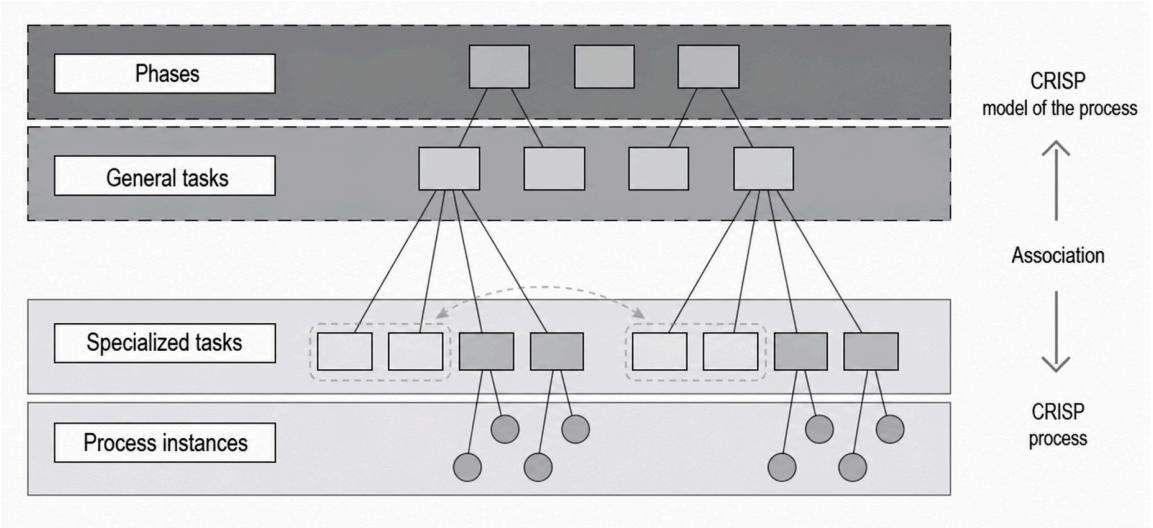


Fig. 6 Four-level decomposition of the CRISP-DM methodology (translated by the author from a Chapman figure)

The life cycle of a data mining project contains the project phases, their tasks and relationships. The life cycle consists of six phases as shown in Fig. 7. The sequence of phases is flexible, allowing movement back and forth between the different phases. The result of each phase determines which phase or task will be performed next. The arrows in the figure indicate the most important and common dependencies between phases.

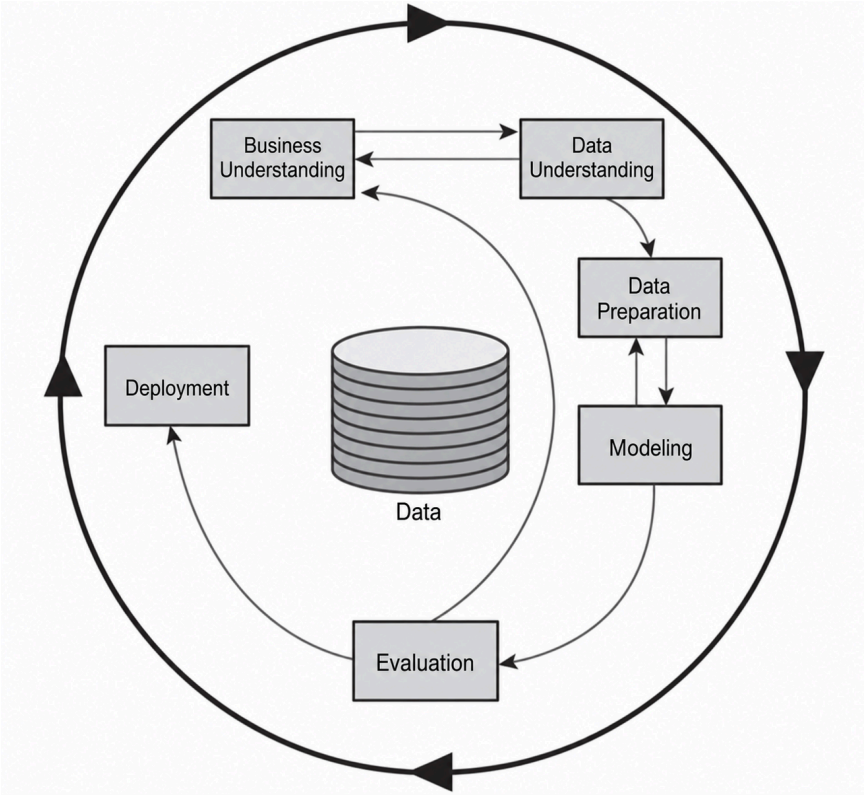


Fig. 7CRISP-DM data mining project life cycle (translated by the author from a Chapman figure)

The outer circle of the figure shows the cyclical nature of the data mining process, which does not end with the implementation of a solution. The conclusions obtained during the implementation of the methodology, and from the implemented solution, can unlock new, more focused business questions. Subsequent data extraction processes will be aided by the experience gained from previous ones.

Data modeling is performed using IBM SPSS Modeler and Wekasoftware products. Models have been created to predict the success of start-up companies from Bulgaria, by using the complete set of collected data. In addition, a model for predicting the success of technology start-ups has also been created by using only the technology company data, which is a subset of all data.

2.4 Proposed Solutions, Achievements, and Gaps

In this point, existing models for predicting the success of start-ups are reviewed and analyzed. Trends in creating models for predicting success have been identified, as well as established preferred models.

Software products for predicting the success of start-up companies have been studied and analyzed, paying attention to their strengths and weaknesses, as well as their suitability to solve the task of predicting the success of technological start-up companies from Bulgaria.

2.4.1 Comparison of existing theoretical models for predicting the success of start-up companies

There are many scientific developments that are concentrated on elucidating the characteristics,describing the success of start-up companies. Data on such developments have been available since 1974, when Wells (Wells, 1974) proposed a model of the success of a start-up company. The model is based on the variables of management commitment, product, marketing, skills (marketing, engineering, financial, production), marketing plan, references, other participants in the transactions, exit method.

A detailed classification of success prediction models is made by Ensley and Teal (Teal E. J., 1998), (Ensley, 1999), (Teal E. a., 2003) and it includes 21 models from 1976 to 1999. Similar classifications were made by Carland and Chrisman (Carland, A new venture creation model, 2000), (Chrisman, 1998). The classifications are united by the author in one, which is supplemented with many missing and more current models that are described, compared and analyzed (Abbas, 2008), (Astebro, 2000), (Biggadike, 1976), (Davis, 2005), (Gartner, 1988), (Shepherd D. , 1999), (Shirokova, 2010). The output contains 44 models for predicting success, which are presented and compared in Table 4.

Table 4Comparison of Models for Predicting start-up Success

Research	Key Factors
Wales, 1974	Management Commitment, Product, Marketing, Skills (Marketing, Engineering, Financial, Production), Marketing Plan, References, Other Dealing Participants, Exit Method
Bigdike, 1976	Entrepreneur and strategy
Poindexter, 1976	Management quality, expected rate of return, expected risk, Percentage of total capital investment, financial regulations for investor rights, firm growth rate, capitalization rate, investor control, application of tax reduction methods
Von Hippel, 1977	Entrepreneur and Capitalization
Vesper, 1980	Knowledge, Idea, Connections, Resources, Implementation
Hobson and Morrison, 1983	Entrepreneur and Strategy
Taibji and Bruno, 1984	Entrepreneur Recommendations, Profit Margin, Product Uniqueness, Product Patentability, Raw Material Availability, Market Access, Product Market Need, Market Size, Growth Potential market, freedom from regulations, protection from entry of competitors, resilience to economic cycles, protection against product obsolescence, protection against adverse development, exit opportunities, potential for mergers and acquisitions, hedge to current investment, tax advantage
Miller and Kemp, 1985	Strategy

Gartner, 1985	Person, Environment, Organization, Implementation
Sandberg, 1986	Entrepreneur, Strategy and Industry Structure
Moore, 1986	Person, Environment, Innovation, Implementation
McMillian and Day, 1987	Strategy
Romanelli, 1987	Strategy and the State of the Industry
McDougall, 1987	Strategy, Origins, Industry Structure and Interactions
Stewart and Abetti, 1987	Environment, Strategy, Structure and the Entrepreneur
Hall, 1989	Subjective evaluation
Bygrave, 1989	Person, Environment, innovation, triggering event, implementation
Gartner, 1989	Psychological predisposition, business start-up event, opportunity recognition, environmental and financial support, implementation
Manning, Burley and Norburn, 1989	Company characteristics, relationships, determinants, realization
Ruhr and Keely, 1990	Management, company strategy, competitive environment
Kankel, 1991	Strategy, Industry Structure and Interactions
Jensen, 1991	Entrepreneur and Industry
McCarthy, 1992	Strategy, Environment, Resources and Change
McDougall, Robinson and DeNici, 1992	Origin, Strategy, Industry Structure, and Strategy-Structure Interactions
Brüderl, Preisendorfer and Ziegler, 1992	Founder-related human capital, organizational characteristics, environmental conditions
Boland, 1993	Industry origin, strategy and structure
Blandon, 1993	Capitalization and type of financing
Niawali and Fogle, 1994	Entrepreneurial and business skills, socio-economic factors, government policies, financial resources, implementation
Stearns, Carter, Reynald and William, 1995	Location, Strategy, Industry Structure and Interactions
McGee, Dowling and Megison, 1995	Team Expertise, Strategy and Cooperative Arrangements
Robinson, 1995	Strategy, Industry Structure and Interactions
Lassier and Korman, 1995	Stage of market entry, planning, use of outside advisors, entrepreneur education
Stewart, 1996	Achievement motivation, risk taking, innovation preferences, psychological disposition
Teal, 1998	Entrepreneurial team, strategy and industry structure

Enzley, 1999	Entrepreneurial team, strategy, industry structure and strategy-structure interactions
Shepard, 1999	Stability, timing of market entry, time advantage over competition, competitive environment, educational capacity, competence - related to industry sector
Park and Lee, 2000	Strategy, resources, environment, interaction between the factors
Carland and Carland, 2000	Idea, entrepreneurial mindset, knowledge, creativity, experience, intuition
Thiel and Hofer, 2003	Strategy, industry structure, interaction between strategy and the structure
Astebrow, 2004	Expected profit, technological opportunity, development risk, suitable conditions
Davis and Zweig, 2005	Market, product or service, financial resources, managerial team
Abbas, 2008	Technological innovation, business concept, technological entrepreneur
Shirkova and Shatalov, 2010	Environmental factors, management style, strategic behavior
Manev, Manolova, Goshev, Harkins, 2012	Networking (social networks of the founder) and differentiating strategy

The considered models (Yankov, Overview of Success Prediction Models for New Ventures, 2012) differ according to their purpose and according to the conditions under which they were created and can be grouped into the following types:

- general - for evaluating start-up companies;
- to evaluate companies participating in a business plan competition;
- to evaluate companies operating in a certain local market;
- to evaluate technology start-ups;
- others.

It is characteristic of the models describing the successful creation of new companies from the distant years that they mainly consider individual success factors. Some of the factors considered in early models are knowledge, business idea, realization of the idea, business contacts, temperament of the entrepreneur, external environment. Some of these models treat the company as an independent entity and do not take into account external factors such as the business environment in which the firm will operate. Models for predicting success, created at a later stage, gradually begin to acquire structure, by categorizing the success factors into groups. Some of them (for example, Park and Thiel's models) (Park, 2000) define the influence of the external environment as the main factor in the process of creating new companies. Thus, to the main factors determining the success of a business, factors characterizing the environment in which the company operates are added: market, competition, industry structure, and others.

Brüderl, Prezendorfer and Ziegler (Brüderl, 1992) determine the chances of survival of newly created companies using a model in which three main categories of factors are used. These categories are: job and personal characteristics of the founder of the company, organizational characteristics and conditions of the environment in which the company develops its activity. The full list of factors from the model includes:

- *Personal and job description of the founder of the company*
    - *Basic - years of education, scientific degrees, years of work experience*
    - *Specific - accumulated experience in the industry in which the company will develop its activities, experience gained from independent activities, management experience*
- *Organizational characteristics*
    - *Innovation*
    - *Company size*
    - *Company management strategy*
- *Environmental characteristics*
    - *Location of the company*
    - *Industrial cluster*
    - *Market conditions*

Lussier and Korman in 1995 presented their analysis of the success of small start-up companies (Lussier, 1995). They identify four key success factors. These factors do not form a complete prediction model, but can be used as useful guidelines when starting a new business:

- *Product/service incompatibility*- companies that develop a product or service that is too new or too marketed have a greater chance of failure than companies that develop a product or service that is already on the market and has a growing search.
- *Planning*– companies that develop business plans have a greater chance of success than others.
- *Availability of consultants*– companies that do not use the help of professional business consultants have a greater chance of success.
- *Educational degree*– entrepreneurs starting their business without a university education have a lower chance of success than those who have graduated from university.

Sheppert presents another interesting model through 1999 (Shepherd D. , 1999) which evaluates the chances of success of a start-up company according to the following factors: stability, time of market entry, competitive environment, education and competencies for the industry in which the company will develop.

In 2008, Abbas proposed his model for predicting the success of a student competition for best business plan (Abbas, 2008). In his proposal, resources and the environment in which a company will develop are not identified as major factors, as student companies start in the same environment and fight for funding. The main factors of this model are technological innovation, the business idea and the entrepreneur.

Another model that deserves attention was presented by Davis and Zweig in 2005 (Davis, 2005). According to them, the factors that most influence the success of start-ups are: suitable product/service, financial resources and management team. This model is one of the few that include financial resources as a key factor for success.

In Bulgaria, there are not many developed models for predicting the success of start-up companies and almost no software to apply these models and make them available to interested parties. One of the studies in this direction (Manev, 2012), by Manev, Manolova, Goshev and Harkins concerns certain factors for the success of start-up companies in Bulgaria. It is based on the hypothesis that the networking capabilities of start-up companies (depending directly on the contacts and social networks of the founder) and the differentiation strategy are related to the success of start-up companies. The research assumes that the success of companies in the conditions of a transition economy depends on different factors, compared to companies developing their activities in countries with a developed economy. In order to prove the hypotheses, they apply a quantitative study of 334 Bulgarian companies. The research does not include a comprehensive set of factors that would affect the success of start-up companies.

2.4.2 Existing theoretical models for predicting the success of start-up companies based on Sandberg's model

A popular basic model used as the basis of many modern models for predicting the success of start-up companies is that of Sandberg (Sandberg, 1986). In his research,, Sandberg for the first time also adds characteristics that describe the entrepreneur, the business strategy of the given company and the structure of the industry in which it develops. The pattern can be described by the formula:

$$NVP = f(E, IS, BS) \quad (1)$$

Where NVP is the new firm's productivity, E is the entrepreneur, IS is the industry structure, and BS is the business strategy.

Further studies based on Sandberg include other factors such as the interaction between firm strategy and industry structure, available resources, and the human factor (entrepreneurial team) is often omitted. Table 5compares forecast models based on the main factors proposed by Sandberg: business strategy and industry environment. It shows the similarities and differences between the different models, with a plus sign "+" showing the factors included in the respective model, and additional factors are shown in the last column.

Table 5Comparison of success prediction models based on the model of Sandburg

Research	Key Factors

	Strategy	Industrial structure	Interaction	Other
Sandberg, 1986	+	+	-	Entrepreneur
Romanelli, 1987	+	+	-	
McDougal, 1987	+	+	+	Origin
Stewart and Abetti, 1987	+	+	-	Entrepreneur Environment
Kankel, 1991	+	+	+	
McDougall, Robinson and DeNici, 1992	+	+	+	Source
Boland, 1993	+	+	–	Origin
Stearns, Carter, Reynald and William, 1995	+	+	+	Location
Robinson, 1995	+	+	+	
Chrisman, 1998	+	+	–	Entrepreneur, resources, organizational structure
Teel, 1998	+	+	-	Entrepreneurial team
Enzly, 1999	+	+	+	Entrepreneurial team
Thiel and Hofer, 2003	+	+	+	
Shirkova and Shatalov, 2010	+	+	+	Global environmental factors, management style

McDougall (McDougall, 1987) develops a Sandberg-like model in which he includes the factors business strategy and industrial environment, but what is new about him is that he also adds characteristics that describe the connections between the chosen business strategy and the industrial environment in which a given company will develop its activity. Later developments also confirm that the relationship of a company's business strategy with the industrial environment is key and has an important role in determining the success of start-ups.

The importance of the "entrepreneurial team" factor is controversial and is often omitted from models. In an analysis of different models based on Sandberg's linear model of new firm productivity, Joan and James Carland observed that the results differed. They note that Sandberg fails to empirically link the characteristics of the entrepreneur to firm productivity. Most research on new firms focuses on their performance as a function of the entrepreneur, and therefore, despite his findings, Sandberg is reluctant to remove entrepreneur characteristics such as management competencies and industry experience from the model, as investors and venture capital funds value these characteristics.

Chrisman offers an extended model of new company productivity (Chrisman, 1998) based on Sandberg's model. In addition to the entrepreneur (E), industry structure (IS) and business strategy (BS), the model also includes the resources (R) and the organizational structure (OS). The model is represented by the formula:

NVP = f (E, IS, BS, R, OS) (2)

As the models for predicting success are further developed and validated with data, in some studies the results show that the human factor (usually characterizing the entrepreneurial team) has no relationship with the success of the company. Thiel and Hofer, for example, propose a model that omits the factors related to the entrepreneurial team and includes only the business strategy of a company, the industrial environment, and the relationship between the two factors.

Other important factors that stand out as primary according to the newer models are the entrepreneurial team, company location, environment and management style. Resources (such as assets) are often omitted because they are not important if funding is secured (eg in start-up competitions). In Shirkova and Shatalov's 2010 model (specially developed for the specifics of the Russian business climate), the industrial environment factor is expanded with factors that describe the overall environment more generally, and not just the specific industry.

From the forecast models considered so far, it can be concluded that **models that are based on basic factors such as strategy and industrial structure are a suitable option** to create a practical application of success prediction. These models can be supplemented with more factors characterizing successful companies and tested with statistical information. After validation, they can serve as the basis of a practical application for predicting the success of start-up companies.

2.4.3 Research and analysis of software products for predicting the success of start-up companies

A survey of software products for predicting the success of start-up companies shows that tools with a similar purpose are contained as part of crowdfunding portals for business ideas (crowdfunding).

One of the popular crowdfunding portals is “EquityNet” (EquityNet - The Leading Equity Crowdfunding Platform, 2014). According to the statistics published on the website, it has been used by more than 20,000 entrepreneurs and 20,000 investors, and the funding provided by investors to entrepreneurs on the portal amounts to more than 240 million dollars. Other more famous crowdfunding portals are "Sellaband" (Sell a Band - Where Fans Invest in Music, 2014), "IndieGoGo" (Indiegogo - Crowdfunding for What Matters to You, 2014), "GiveForward" (GiveForward - The Number 1 Way to Raise Money for a Loved One, 2014), "Kickstarter" (The World's Largest Funding Platform for Creative Projects, 2014) - with funding allocated for more than 1 billion dollars, "RocketHub" (RocketHub - Crowdfunding, 2914), "Fundly" (Fundly - Online Fundraising Websites to Raise Money for Anything, 2014), "GoFundMe" (GoFundMe - Number One for Crowdfunding and Fundraising Websites, 2014) with more than 350 million dollars of investments, "Appsplit" (AppSplit - Crowdsourcing for Apps, 2014).

Crowdfunding portals are quite different from the idea of success prediction software developed in this dissertation. The risk analysis tools they contain are based on a business plan presented by the entrepreneur. Entrepreneurs must present their business ideas publicly (to users of the portals). The downside of this approach is that entrepreneurs risk their business ideas being borrowed, or even stolen, before they ever receive any funding to implement them.

Three success prediction applications that rely on survey input, use statistical information, and estimate the chances of success and failure of a given survey company are selected and discussed in detail. These applications, as a working principle, are close to the software product developed for this dissertation.

2.4.3.1 StartupNation

The best developed tool for predicting the success of start-up companies was created by the company “StartupNation”. The tool is called "start-up Risk Calculator" (start-up Risk Calculator) (start-up Risk Calculator, 2014). This tool is affiliated with “EquityNet” because it uses their statistics to create its forecast model. Only a year after its inception, it became available as an EquityNet portfolio tool. The tool is a web application freely accessible on the Internet and its purpose is to determine the chances of success of start-up companies based on eight factors:

- size of investments,made so far in the given company,
- access to new funds of financing,
- financial management,
- depth of business planning by the company's management,
- annual growth of the company,
- management experience,
- experience of the company in the industrial sector in which it develops its activity,
- time interval of the assessment.

The application user is given the opportunity to answer eight questions related to the listed factors. After evaluating the responses to the survey, the application provides a percentage prediction of the probability of success of the company that completed the survey. The factors that influence the result are also presented as part of the result (Fig. 8). Due to the use of too few success factors, as well as the small amount of information the user has to enter to describe their business, the prediction generated by this tool does not seem very reliable. As a method of commercializing the tool, the results page has been developed as a marketing tool presenting offers for business advice, analysis and consulting.

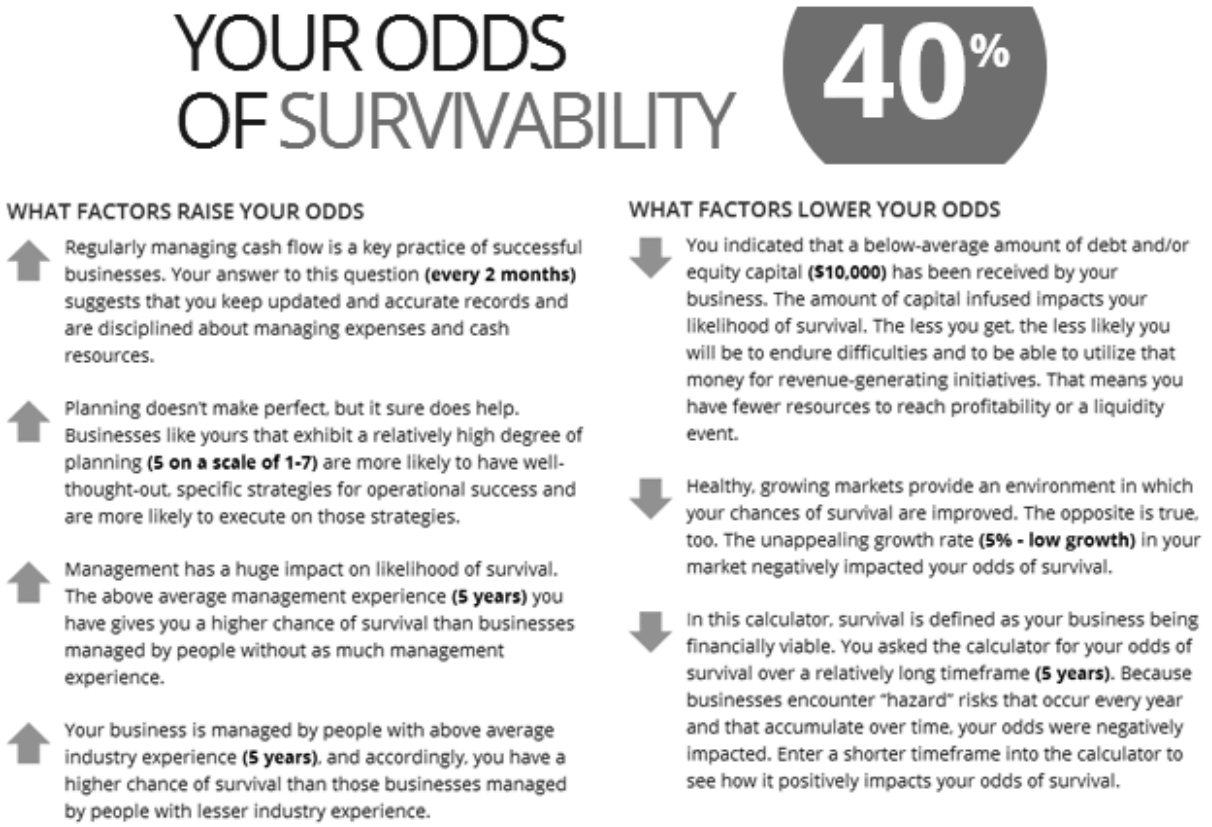


Fig. 8Results of the start-up Risk Calculator prediction tool (created by the author by implementing the tool)

StartupNation's tool can benefit entrepreneurs without any experience as it provides basic guidance on some of the important factors that can lead to success given business.

2.4.3.2 Compass for New Companies

Another app for predicting the success of start-ups is the start-up Compass (start-up Compass) (Compass - Benchmark Your Business Metrics From 30+ Data Sources, 2014). The app is a technology start-up assessment tool that allows you to fill out a survey. The user can update the filled data monthly. After completing the survey, results are displayed, evaluating the company according to its type and stage of development - in the form of various graphs showing the current results of the company based on the other companies using the tool (Fig. 9). The idea of the creators of the application is that it will serve as a personal assessment of a given business, and with its help, potential problems can be more easily noticed, priorities clarified, the process of forming the entrepreneurial team supported, and the progress of a given company can be more easily measured.

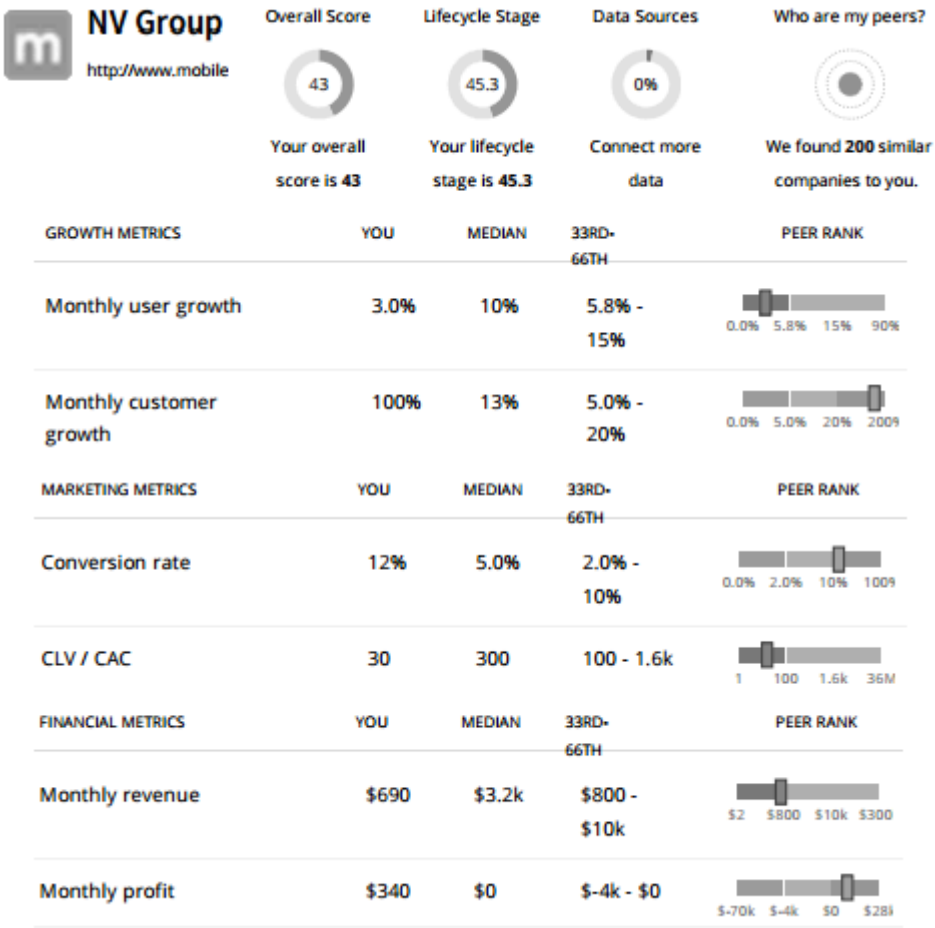


Fig. 9Snippet of start-up Compass success prediction results (created by the author by implementing the tool)

The tool uses statistics from over 600 technology start-ups, mostly from Silicon Valley, to determine key success factors and build its operating model. start-up Compass uses the following factors to make its analysis:

- customers, product / service,
- entrepreneurial team,
- business model and financing methods.

start-up Compass advises how to evaluate these factors at the right time. For example, companies that do not take the time to identify what the market for their product or service will be, do not analyze the market in the planning stage, but go directly to the creation of the product, are considered inconsistent and their chances of success are small. According to the tool, 80% of these types of companies fail, while 45% of companies that initially identify their potential customers and only then develop their product or service are successful. In the beginning, entrepreneurs can easily be fooled and concentrate all their efforts on creating their product without even having assessed the real need for its use. The creators of the application also note that the more companies use their tool, the more in-depth and accurate analyzes it will make, and the more closely related to a specific company development recommendations will be.

The start-up Compass tool was created to analyze technology companies only. It is free to use, but registration is required.

2.4.3.3 start-up Checker by Blueprint

Another web app for predicting start-up success is called start-up Tester (Blueprint New Business start-up Tester) (Blueprint New Business start-up Tester, 2014). It was created by the Australian company "Think Blueprint" and is a complete solution for business planning, including tools for creating business plans, assessing risk and highlighting the main activities when creating a new company. The tool itself is free to use, but a fee is charged to apply the full set of planning tools. The tool presented by Think Blueprint uses a short survey (about 10 questions) to obtain data about start-ups. Chances of Success results are emailed. The content of the email is shown in Fig. 10.

Title of Your New Business: NV Group

Your Preparation Index: 40% - 60%

Blueprint Preparation Index: 60%

Your Success Probability Score: 40% - 60%

Blueprint Success Probability Score: 63%

Fig. 10 Think Blueprint tool success prediction result (created by the author by applying the tool)

Application operation cannot be traced and no detailed information is given how the result was obtained. This application has the most basic functionality of all similar applications examined in the study.

2.4.3.4 Conclusions

The examined applications for predicting the success of start-up companies do not fulfill the task of predicting the success of technological start-up companies from Bulgaria to the maximum extent. They are a prerequisite and motivation for the development of the software for this dissertation. The main similarities are in the filling in of business data by the founder, owner or manager of the company to be analyzed and the use of statistical data for comparison and obtaining a prediction of success. In some of the applications, for easier perception, graphic images are added to the result.

Table 6 presents a summary comparison of the analyzed tools for predicting success, and their strengths and weaknesses, respectively.

Table 6Summary comparison of analyzed tools for predicting success

Tool	Strengths	Weaknesses
start-up Risk Calculator by StartupNation	• uses a huge database of statistics obtained from the company's other applications, which helps to increase accuracy	• the number of input parameters is small (eight variables) and this implies little precision and accuracy of the forecast

	● the environment in which the company develops its activity does not determine the accuracy, since the results in the database are collected from companies from different countries	
Compass for New Companies	● good application design and clear graphically presented results ● possibility of monthly update of the result and to measure the progress of the company ● offers for free software products that help businesses	● closely specialized in working with technology start-ups ● statistics used are from the US and the tool may not give accurate results for other countries.
start-up Checker from Blueprint	● has simplified (basic) functionality	● lacks information about the statistics used ● there is no data on how the result is calculated ● the number of input parameters is small (10 variables) and this implies little precision and accuracy of the prediction

From the review of the existing applications, it can be concluded that the software for predicting the success of start-up companies will not be the first such application. The most famous applications of this type are actually portals and include both users who have the role of entrepreneurs and investors who are ready to finance some of the entrepreneurs' ideas. When registering on these portals, an entrepreneur usually has to provide a large amount of information about his business or idea. In most cases, it is necessary to create a business plan and make it available on the relevant portal. Based on this business plan, the portal enables the assessment of the business idea by key factors and offers risk assessment tools. The data from these analyzes are visible both to the entrepreneur who registered his idea in the portal and to potential investors.

Investors, as users of such a portal, most often get the opportunity to review business plans and get acquainted with new business ideas of entrepreneurs registered on the portal and seeking funding. With the help of the analysis tools and based on the presented business plan, an investor, or several together, can decide to make an offer to finance a certain business, and in return they usually receive a share of the company. This method of financing is better known as crowdfunding. The application, the subject of this dissertation, could also be developed in the future as a web portal that meets the interests of entrepreneurs and investors.

The software that was developed for the dissertation outperforms existing ones by basing its prediction on significantly more input data (about 100 variables, compared to about 10 for competing tools), helping to make a more accurate prediction. The success prediction model and software were created specifically for the Bulgarian business environment. The statistical data that was used to create the forecast model was collected from Bulgarian start-ups. As a result, the success prediction application gives the most accurate results for Bulgarian companies.

2.5 Review of Data-Mining Software

This section reviews the author's survey of data mining software products and their characteristics. Various products have been compared and those that are suitable for application in the research to this dissertation have been selected in order to confirm the hypotheses through quantitative methods and synthesize models for predicting the success of start-up companies.

2.5.1 Overview of IBM SPSS Statistics

IBM SPSS Statistics is an integrated set of software products that address the entire analytical process of extracting knowledge from data – from planning, to data collection, reporting and implementation (IBM, 2014). The product, which was originally developed by SPSS Inc, is currently maintained by IBM and features more than ten fully integrated modules for conducting research and decision making. The software name "SPSS" originates from an acronym meaning "Statistical Package for the Social Sciences", but today the product is popular in various fields.

2.5.2 Overview of IBM SPSS Modeler

IBM® SPSS® Modeler is a set of software tools for extracting knowledge from data (data mining), that enable rapid development of predictive models by using business expertise and their implementation in business activities to improve the decision-making process. Designed around the accepted industry standard methodology CRISP-DM, IBM SPSS Modeler supports the overall process of extracting knowledge from data. It offers a variety of modeling methods borrowed from machine learning, artificial intelligence, and statistics. They enable the extraction of new information from available data and the development of predictive models. Each of the methods available in the product has certain advantages and is best suited for a specific type of task.

IBM® SPSS® Modeler is a commercial product and can be purchased as a standalone product or used as a client application in combination with SPSS Modeler Server. For the purposes of this study, the product IBM® SPSS® Modeler is provided for free use by IBS Bulgaria.

2.5.3 Overview of Weka

Weka is a set of machine learning algorithms applied to knowledge extraction tasks from data (Bouckaert, 2013), (Waikato, 2014). Algorithms can be applied directly to a dataset or called from Java code. Weka contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. The product is also suitable for developing new rules for machine learning. It contains command-line tools as well as a graphical user interface. The graphical user interface includes four different applications: Explorer, Experimenter, KnowledgeFlow, Simple CLI(Fig. 11), suitable for different tasks.

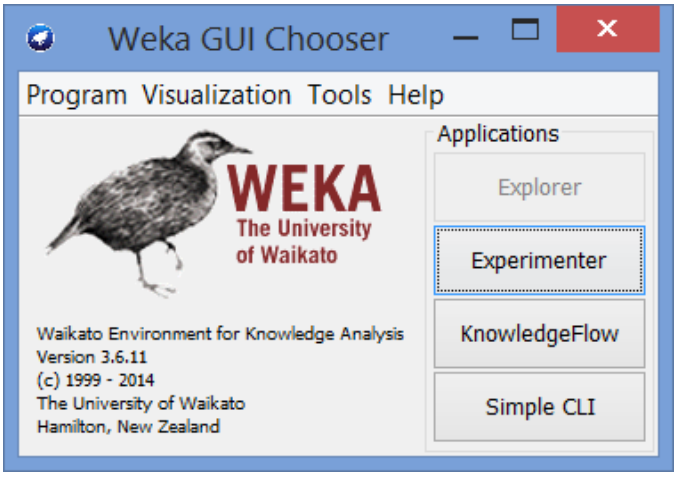


Fig. 11Selecting an application from the Weka GUI (screenshot created by the author)

Weka is a free product and is distributed as open source under the GNU General Public License.

2.5.4 Overview of HP Vertica

HP Vertica Analytics Platform is a big data analytics platform (HP Vertica, 2015). It enables large organizations to analyze massive volumes of structured, semi-structured and unstructured data, taking advantage of the platform's speed and analytics.

The main advantages of HP Vertica are faster execution of database queries (from 50 to 1000 times faster compared to traditional databases), from 10 to 30 times more stored information on each server of the infrastructure and the ability to apply specialized tools for business.

HP Vertica Analytics Platform supports the concept of extracting knowledge from data (data mining) through the application of"R" - an open source statistical software that includes an efficient programming language (The R Foundation, 2015). The HP Verticaplatform is suitable for running complex algorithms programmed in R, running on multiple data servers. Examples of such algorithms are product suggestions in online stores with millions of items, calculation of PageRank from search engines (Lawrence, 1998), calculation of prices of financial instruments, recognition of important people in social networks.

HP Vertica is a commercial product, the use of which is subject to the payment of a license.

2.5.5 Overview of Microsoft SQL Server

Microsoft SQL Server is a database platform for managing a large number of online transactions, storing data, e-commerce, as well as BI (business intelligence) (Microsoft, 2015).

One of the components of Microsoft SQL Server 2005 is called Analytical Services and is used for business data analysis, online analytical processing and knowledge extraction from data. Analytical Services contains the following capabilities and tools, by the joint application of which, complex data mining problems can be solved:

- a set of data mining algorithms following industry standards in the field;
- the Data Mining Designer tool, which allows creating, maintaining, viewing, and making predictions from data mining models;

- the DMX language, which can be applied to manage data mining models and create complex predictive queries.

Microsoft SQL Server 2005 is a commercial product, the use of which is subject to payment of license.

2.5.6 Findings and Conclusions

From the overview, it can be concluded that most products for extracting knowledge from data are commercial and for their application in the research it is necessary to purchase a license.

The product IBM SPSS Modeler will be used, because it specializes in solving tasks with knowledge extraction from data, is developed by an established company (IBM) and is easy to implement. For the purposes of this study, the product IBM® SPSS® Modeleris provided for free use by IBS Bulgaria. For data preparation and for some specific tasks, the IBM SPSS Statistics product, which is compatible with IBM SPSS Modeler, will be used in addition.

Since different data mining software products create classification models using different algorithms, it can be expected that the results obtained by applying of more than one product on the same data will be different. For this reason, one more product will be applied to classify the data and that is Weka –a popular open source software. An additional contribution of implementing one more product and creating more diverse models is that the ability to update and select the success prediction model will be developed in the dissertation prediction software, and users will be able to choose between alternative models.

2.6 Selection of Technologies and Tools for Information-System Development

In this point, they are discussed tools, programming languages and technologies, and different possible solutions for the development of the software system, the subject of the dissertation, are compared. The technologies and tools that best fit the objectives are selected and will be implemented.

2.6.1 Selection of application type

Before starting the development of the software I3SP, a decision must be made whether it will be implemented as a desktop application or as a client-server web application.

Desktop applicationshave the main disadvantage of being platform dependent and requiring installation. For each different operating system (Windows, Linux, OS X, Android),different applications must be created. A major advantage of desktop applications is that they have direct access to system resources, as a result of which they can run faster.

Client-server web applications are platform independent from the perspective of the end client. The user only needs a web browser to use them, and one is included in every modern operating system. As a disadvantage of web applications, it can be noted that the client needs a constant connection to the Internet and the speed of the application is dependent on the speed of data transfer between the server and the client browser. Constant connectivity is among the prerequisites for I3SP, due to the use of data from Google servers.

For the reasons described, it was decided to develop I3SPas a client-server web application.

2.6.2 Selection of an operating system

Once the decision has been made I3SP to be developed as a client-server web application, an operating system must be selected on which to install the server part. Table 7presents the main characteristics of the two most popular types of operating systems (Surveys, 2014).

Table 7Comparison of web server operating systems

	Linux and Unix-based	**Microsoft Windows**
Market Share	67.6% Web Server Market Share	32.5% Web Server Market Share
Price	Most distributions are distributed for free under the GPL license. Server services (virtual server and web hosting) running on Linux are available at lower prices.	Distributed for a fee. It is necessary to purchase a license for use. Server services (virtual server and web hosting) running on Windows are offered at higher prices.
Software	Linux is available in a large number of distributions: Ubuntu, OpenSUSE, Fedora, Debian, CentOS, Mandriva, Gentoo and its source code is open. Most web server applications are free (eg Apache, MySQL, PHP, Perl). Linux and most applications for it are often more complicated to use and set up than those for Windows.	Some applications are paid (eg Microsoft SQL Server), although free ones also exist. When creating paid apps, more effort and resources are often put into development and testing, and some of them also offer support.
Security and stable operation	Proven security, options for various configurations, flexibility and stability of operation, no frequent need to reboot, complete independence from the graphical interface.	Good security and stable operation, greater capabilities of the graphical interface, easier initial configuration, fewer options for settings and adaptation.

For the purposes of the application, Linux will be used as the server operating system, and the main factors of choice are the possibility of legal free use and more affordable web hosting services.

2.6.3 Choosing a Web Server Application

A client-server web application needs web server software to handle requests coming over the HTTPprotocol from the client browser and return the appropriate content.

One of the most common and well-known web servers is **Apache**- most often installed under Linux, but with versions for other operating systems as well. It is open source, freely distributed, has a reputation as a secure and well-scalable web server, particularly suitable for small and medium-scale projects.

With the server versions of the Windows operating system, the web server **IIS**is available for free, which also offers sufficiently stable operation and scalability, as well as a large number of settings.

For the purposes of the I3SPapplication, Apache web server will be used because it is multi-platform and free to use.

2.6.4 Selection of server-side development tools

Web server-side programming languages enable dynamic generating web pages. The following server-side programming languages are most commonly used:

- **PHP** – a popular general-purpose scripting language specifically tailored for rapid web application development (Achour, 2014). Its main advantages are that it is free, works fast, is intuitive to use (with moderate project complexity) and allows relatively fast development of dynamic web server applications. There are many PHP modules that add different functionalities. PHP is distributed under an open license and runs on a variety of platforms.
- **ASP. NET-** a web server application development platform created by Microsoft (Microsoft, 2014). ASP. NET allows programming in various object-oriented languages, such as C#, Visual Basic, C++. The platform uses a virtual machine called CLR (Common LanguageRuntime), which, with the help of CLI (Common Language Infrastructure), executes the code written in the different languages. ASP. NET runs on Windows and does not require payment of an additional license.
- **JSP (Java Server Pages)**– technology for the development of web server applications using the Java programming language (Oracle Corp., 2014). Application code is compiled using Java Virtual Machine technology to allow source code to be translated to native code, enabling cross-platform operation. Use of Javais not subject to license fees.
- **PERL**- general purpose scripting language. It is distributed under the GNU General Public License, which is a free software license. Works under multiple platforms.

For software development purposes, I3SPselected PHP programming language because it is free to use, runs under Linux and is suitable for relatively fast web application development.

2.6.5 Customer Selection of Development Software

The language that web browsers interpret to visualize information is HTML (translated **as** "hypertext markup language"). HTML consists of words called tags enclosed in square brackets that describe the semantics and in some cases the appearance of the document. HTML documents can include embedded code in a scripting language such as JavaScript. HTML code is interpreted by web browsers and, in combination with the CSS language, defines how the information between HTML tags is displayed to the user. xHTML is a variant of **the** HTML language that has the characteristics of the XML language and is accepted as a good practice for building HTML documents. Since xHTML is the most commonly used language for front-end development of web applications and is supported by all web browsers, it will be used for the development of the success prediction software.

JavaScriptis a scripting programming language used to create snippets of code that are embedded in an HTML document. JavaScript code is interpreted by the client browser and cannot be used directly to write data to the server, but with its help, information can be sent to the server and processed there by server scripts. The importance of JavaScript has grown significantly in recent years with the widespread adoption of AJAX, JSON and jQuery.

jQuery is a JavaScript library published in early 2006 by John Resick. jQuery simplifies access to any element on a web page through a syntax that is independent of the web browser

used, thus allowing easy building of dynamic functionality into pages. jQuery is free and open source software licensed under the MIT License or the GNU GPL.

CSS (Cascading Style Sheets) is a programming language used to describe the visual presentation of a document written in a formatting language (HTML, xHTML, XML). Cascading styles allow better control over page elements and give additional options for their visualization.

Due to the set connection requirements with Google drive, I3SP needs asynchronous data transfer between the client browser and the server. Two **technologies for asynchronous data transfer between client and server** will be considered:

- **AJAX(Asynchronous** JavaScript and XML) is an asynchronous data transfer technology that works with server languages PHP, ASP. NET, PERL, JSP, and others and is supported by almost all web browsers available today. It is based on JavaScript. AJAX is a web development technique for creating interactive web applications by asynchronously exchanging small chunks of data without reloading the entire web page. There is no license fee required to develop AJAX applications.
- **Adobe Flex**is a set of technologies providing an opportunity to build interactive web applications running on the Adobe Flash platform. Adobe Flex is a comprehensive framework for the development of RIA (Rich Internet Application, in translation rich Internet applications functionality).

For the development purposes of I3SP, the AJAX technology was chosen and asynchronous requests were often used. Also used is the JSON-P technique, which allows requests to be directed to different Internet domains.

2.6.6 Selection of databases

The application database will be selected based on the expected workload, the types of data stored, the availability of client software for administration and query writing, and whether the chosen programming language has good database support. Some of the most popular databases will be covered:

- **MySQL** - a popular relational database that is distributed under the GNU General Public License. Most versions can be used for free. MySQL has all the important elements expected of a modern database (stored procedures, triggers, views, cursors, transactions, full-text index, and more). Most often, applications written in PHP use a MySQL database. It can be installed under Windows, Linux and OS X. In addition to the database, free client software for administration and writing queries is distributed - MySQL Workbench and phpMyAdmin.
- **Microsoft SQL Server-** relational database management system. MSSQL is a well-developed database offering excellent capabilities for storing and accessing large volumes of data. Applications written in ASP. NET and the . NET framework most commonly use this database. MSSQL Server runs only under the Windows operating system and is distributed under the Microsoft EULA (closed source) license. Usually, this database is used for very large applications (enterprise) - due to the high prices of licenses for the database and for the operating system.
- **PostgreSQL** – an open source relational database, which does not yield to the capabilities of paid products. The query language used by this database is PL/SQL. PostgreSQL can be installed on different platforms.

For the development purpose of the I3SPapplication, MySQL database will be used because MySQL is well supported by PHP, has very good performance and is free to use.

2.6.7 Selection of Development Environment

For software development environment I3SPNetBeans IDE and PSPad will be used. **NetBeans IDE**is free, open source software that helps develop web applications by offering specialized text editors that recognize PHP, HTML, CSS, and JavaScript code. It supports code coloring, which greatly improves readability. The development environment also offers quick descriptions of methods and classes (purpose, expected parameters, etc.). **PSPad** is a lightweight text editor for software development that also offers code coloring and descriptions of PHP functions and their parameters. PSPad also offers the added convenience of working directly via FTP protocol, which is useful when developing web applications in a remote web hosting server.

2.7 Findings and Conclusions

From the overview of the problem area, it can be concluded that technological entrepreneurship is a current topic, and the problem of predicting the success of start-up companies has gained more and more popularity in the last 25 years and is becoming more and more relevant. Due to the fact that the development of entrepreneurship is set as a top priority for the European Union for the new program period Horizon 2020, and is of strong interest to investors, the development trends in this area and the growing number of publications are expected to be maintained.

The definitions of technological entrepreneurship, start-up company and the success of a start-up company in the literature are examined. The success of start-ups has been considered as a multi-faceted problem with different trends and options for its definition. Interviews were conducted with entrepreneurs, founders and owners of companies, on the basis of which a definition of a successful start-up company was chosen for the purposes of this dissertation.

Various success prediction models created from 1974 to the present are examined. The development of the models was analyzed and the success factors of the companies proposed in them were compared. A model created by Sandberg is examined in more detail, which serves as the basis for many subsequent models, the factors of which have been analyzed in various studies. Sandberg's model can serve as a basis for building new models for predicting the success of start-up companies from Bulgaria.

A methodology was chosen for the individual stages of the research.

Different software products for predicting the success of start-up companies were studied and analyzed. Three products have been identified that are closest in type to the software that will be developed for the dissertation. The three products are compared based on their advantages and disadvantages. Crowdfunding web portals that connect entrepreneurs with investors and in some cases offer functionality to predict the success of companies are reviewed. The main advantages and differences of the software to be developed compared to the existing ones have been identified.

Various technological solutions, tools and programming languages have been compared, and those that will be applied in the development of the software product to the dissertation - a system for predicting the success of start-up companies from Bulgaria have been selected.

CHAPTER 3: DEVELOPMENT OF A SUCCESS-PREDICTION MODEL

This chapter develops and proposes hypotheses for:

- a model of the company creation process,
- an abstract model for predicting the success of a start-up company, containing the categories and subcategories of success factors,
- a detailed model of company success, containing proposal for the individual success factors.

3.1 Model of the Company-Formation Process

Joan and James Carland (Carland, A new venture creation model, 2000) argue that the success of new firms can be predicted if we first model the company creation process. They analyze 8 models of firm creation processes and propose an improved model (Fig. 12), in which the entrepreneur develops a vision for the firm based on his understanding, entrepreneurial attitude, knowledge, creativity, experience and intuition. This early vision is further developed, as a result of which it may be rejected or developed into a new business.

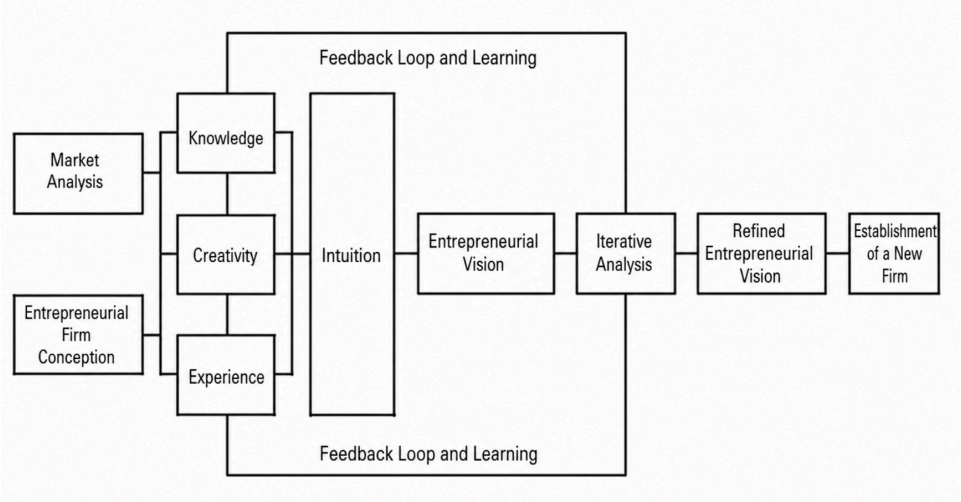


Fig. 12New Firm Creation Process Model (translated by the author after Joan and James Carland)

The author has developed and supplemented the new firm creation process model by adding context including company strategy and industry structure (from Sandberg's model) and replacing the improved entrepreneurial vision with the more formal definition - business plan (Fig. 13). The whole process happens in and depends on the context. The initial part of the process – the formation of the entrepreneurial vision happens in the mind frame of the technology entrepreneur. After analysis, changes and rethinking, the vision can be developed into a (written or informal) business plan. Then the new firm is formed within the relevant industry.

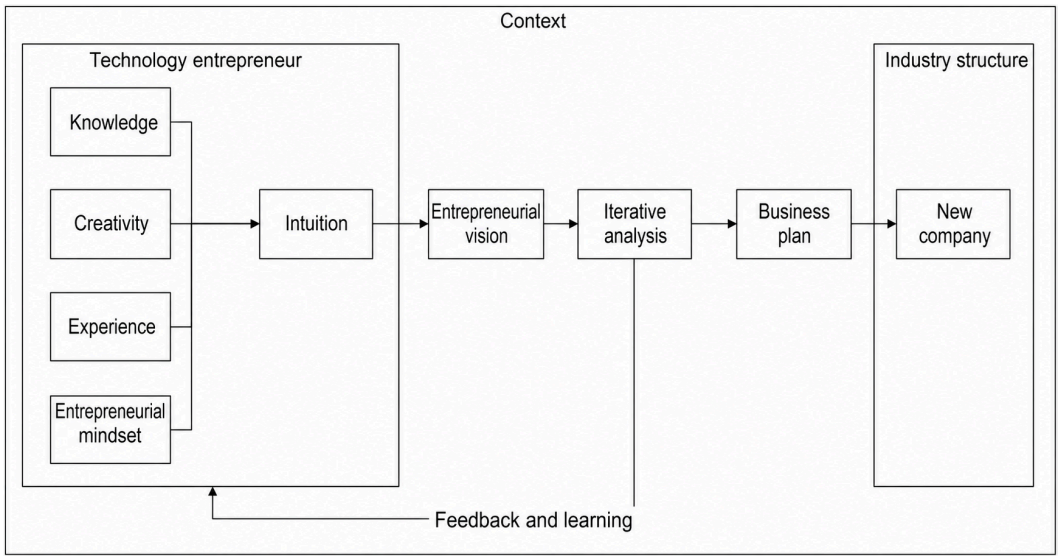


Fig. 13Added model of the process of creating a new company (created by the author)

Thus formulated, the process model serves to make sense of the stages, actions and prerequisites that lead to the creation of a new company.

3.2 Abstract Model for Predicting the Success of Start-up Companies

Based on the reviewed previous attempts to create models for predicting the success of start-up companies, it is concluded that the success prediction models based on the Sandberg model are a good template for designing new models. A model for predicting start-up success should include the following factors: firm strategy, industry structure, and in some cases the interaction between strategy and structure. Incorporating other factors into the model, such as the entrepreneurial team, would be beneficial as they may prove important in validating the model. Each factor in the model must be measurable and clearly defined.

By analyzing the requirements for models for predicting success, the author proposes an extended model for predicting the success of new companies based on Sandberg. The model is represented by the formula:

$$NVP = f(E, IS, BS, R) \quad (3)$$

Where NVP stands for the productivity of the new firm, E is the entrepreneur, IS is the industry structure, BS is the business strategy, and R is the resources.

Compared to the model by Chrisman, Bauerschmidt and Hofer, the proposed model lacks the organizational structure, which is defined as the processes and systems developed by the firm to implement its strategy and achieve its goals. The developed processes and systems are considered part of the company's intangible assets and, therefore, part of the resources.

The team is classified as human resources and could fall under the category of resources, but in the context of entrepreneurship and start-ups, the entrepreneurial team is a key success factor that plays a key role in the creation and development of the company, and for this reason it is separated into a separate category.

Each of the main categories in the model for predicting the success of companies consists of subcategories as shown in Fig. 14 (Yankov, A Model for Predicting the Success of New Ventures, 2013).

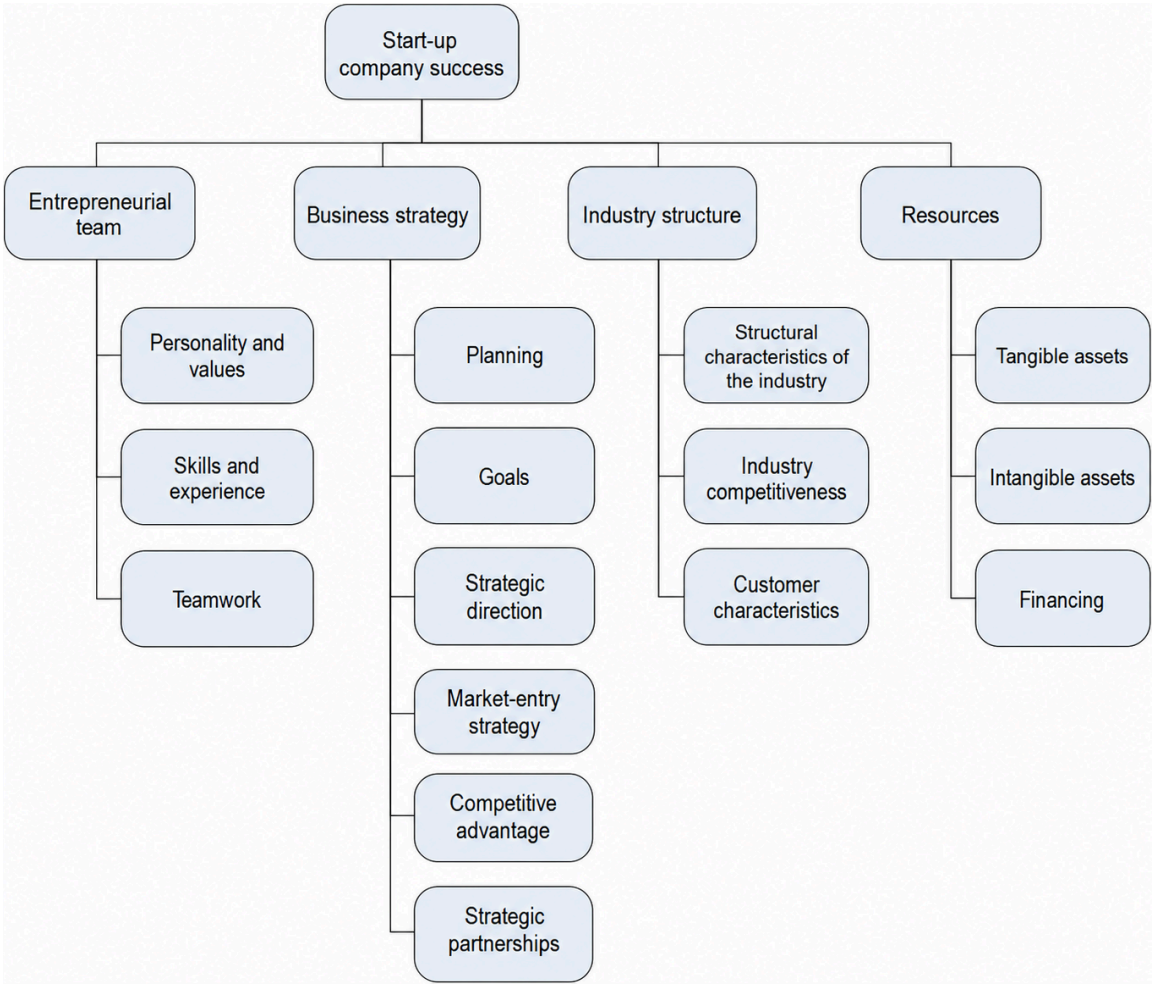


Fig. 14A model for predicting the success of start-up companies (created by the author)

The entrepreneurial team category includes the personality and values of the founder of the company (e.g. confidence, locus of control, risk tolerance), the skills and experience of the management team (e.g. technical skills, management skills, marketing skills, experience in a similar position) and the work in team (e.g., the team's ability to work together and generate value).

The business strategy category includes planning (e.g., depth and breadth of planning), goals, strategic direction (e.g., aggressiveness of company strategy), stage of market entry (e.g., innovator, early vs. late), competitive advantage, and strategic partnerships (e.g., partnerships with competitors, with customers, with the government).

The industry structure category includes industry characteristics (e.g., barriers, size, and predictability), industry competition, and customer characteristics (e.g., concentration, heterogeneity).

The resource category includes tangible assets (e.g., cash, inventory), intangible assets (e.g., established business reputation, patents, social networks), and the resources used. source of financing (eg own funds, venture capital or bank loan).

To ensure completeness and comprehensiveness for the model for the Bulgarian entrepreneurial and business environment, a qualitative study was conducted with entrepreneurs and business owners (described in Chapter 4) and their recommendations and suggestions were included in the proposed model. The described categories and subcategories represent a hypothesis that will be verified with the help of quantitative research and the application of factor analysis and modeling techniques.

3.3 Detailed Model for Predicting the Success of Start-up Companies in Bulgaria

The detailed model for predicting the success of start-ups in Bulgaria is developed based on the proposed abstract model for predicting success. Each of the four main categories and their subcategories of the abstract model are represented by concrete variables. The variables are based both on the success prediction models applied to other countries by other authors, which are discussed in the literature review, and on adaptations and innovations specific to Bulgaria.

One of the innovations in the model is the inclusion of intangible assets to the company's resources, which after the conducted quantitative research actually turned out to be a key factor for the success of the companies. Other innovations in the model are the addition of variables determining the personality qualities and values of the entrepreneur, the presence of a crisis in the industrial sector (a factor that has been relevant in recent years in Bulgaria), the source of the company's financing (a factor related to start-up and young companies).

Some of the variables in the detailed model, such as age groups of the entrepreneurs, completed education and industrial sectors of the companies, are adapted according to the NSI data, which makes possible the comparability of the survey data with the average levels for Bulgaria.

The presented detailed model has been previously revised and improved compared to its initial version, by means of qualitative research - interviews with entrepreneurs and business owners to whom the model was presented. Adaptations and improvements have been made based on the responses and recommendations received. Details and results regarding the qualitative research are presented in Chapter 4.

The detailed model for predicting success,created by the author, includes the following categories, subcategories and variables (questions determining the success of start-ups):

Entrepreneurial team

- Personality and values
  - I can do my job without need from cooperation.
  - I am self-confident.
  - I am enterprising, proactive.
  - What happens to me in life depends on me.
  - I am ambitious.

        - I tend to take over risks.
        - I am not worried about making decisions without having the necessary information.
    - Skills and experience
        - I have entrepreneurial skills (skills in starting a business).
        - I have skills in management.
        - I have marketing skills.
        - I have technical skills (for technologies related to my business).
        - I have human resources (HR) skills.
        - I have investment skills.
        - I have previous experience in starting a business.
        - I have experience in a similar position (to the one currently held).
        - I have experience in the same industry.
        - I have management experience.
        - What education do you have graduates?
        - In what field did you graduate?
        - What is your age group?
        - Are your parents entrepreneurs / business owners?
    - Teamwork
        - The company has a complete team.
        - The team has the necessary knowledge.
        - The team has the necessary skills.
        - The team has the necessary positive attitude.
- **Business strategy**
    - Planning
        - I plan, covering a wide area.

        - I plan in depth.
        - I make a formal (written) strategic plan.
        - I plan frequently.
        - I plan by functional areas / departments.
        - When planning, I consider multiple alternatives.
        - I often analyze the competition.
    - Goals
        - The goals I set for my business are ambitious.
        - The target market share of the company is:
        - The target annual profit of the company is (in BGN)
    - Strategic direction
        - I manage to stick to the initially chosen strategy.
        - The strategy I have chosen is aggressive.
        - The targeted strategy is detailed.
        - I have a clear strategy for my business.
    - Entry strategy on the market
        - How to enter market
        - My business enters the market in the following stage.
        - The product I offer has been accepted by the market to the following extent.
    - Competitive advantage
        - My business has a clear competitive advantage.
        - My business has the following main competitive advantage advantage.
        - I believe the following factors are key to success for my business.
    - Strategic partnerships
        - The firm engages in partnerships with competitors.
        - The firm engages in partnerships with customers.

        - The firm engages in partnerships with the government.
        - The firm engages in partnerships with suppliers.
        - The firm engages in partnerships with other stakeholders.
- **Structure of the industry**
    - Characteristics of the industry in which the firm operates
        - There are barriers (barriers) to new firms entering the industry.
        - There are barriers (barriers) to exiting the business.
        - Most companies in the industry have a good profit.
        - The industry in which the firm operates generates a large portion of the gross domestic product (GDP).
        - There are many bankruptcies in the industry.
        - There is stable demand in the industry.
        - The industry sector in which the company operates is large compared to other sectors.
        - The industry is in a period of growth.
        - There is a crisis in the industry.
        - The industry is predictable.
        - The company operates in the following industry sector.
        - The industry in which the company operates is in the stage of.
    - Competitiveness of the industry
        - The industry is highly competitive.
        - Competitors in the industry are aggressive.
        - The company is exposed to competitive attacks.
        - Small firms are also present in the industry.
    - Characteristics of customers
        - There is a concentration of customers in the industry.
        - Customers are of different types.

- **Resources**
  - Tangible assets
    - Company owns equipment / machinery.
    - Company owns inventory.
    - Company owns real estate.
    - Company owns cash.
  - Intangible assets
    - The company owns patents / trademarks.
    - The company has a recognizable brand.
    - The company has software systems for managing business processes (ERP, CRM).
    - The company has an established business reputation.
    - The company has developed its own products.
    - The company has licenses.
    - The company has established business contacts (social network)
    - The company has access to distribution channels.
    - The company has access to markets in labor.
    - The company has access to capital markets.
    - The company has access to materials and suppliers.
  - Financing
    - The main source used for initial financing of the company is.

3.4 Findings and Conclusions

Based on the literature review of the current state of the problem area, the conducted survey and the qualitative research, including interviews with Bulgarian entrepreneurs, the conclusion follows that the development of a new and up-to-date model for predicting the success of start-up companies should be **based on previous research** of established authors and more specifically on the model of **Sandberg**, as well as the specific conditions of the **environment**described by the Bulgarian entrepreneurs.

This chapter presents a proposal of a **model of the process of creating a new company.** An abstract model for **predicting the success of start-up companies from Bulgaria is** proposed,

including the main categories and subcategories of success factors. The abstract model is elaborated with individual success factors, forming a detailed **model for predicting success.**

The proposed models represent **hypotheses**which are validated by conducting qualitative and quantitative research detailed in Chapter 4.

CHAPTER 4: RESEARCH, APPLICATIONS AND EXPERIMENTS

In this chapter, the conducted research, applications and experiments are described. The hypotheses presented in Chapter 3 have been tested. Through factor analysis, the grouping of factors in the proposed models for predicting success was validated. Through statistical and data mining methods, various success prediction models have been synthesized and analyzed and compared. The most successful models that can be used to create a software product to predict the success of start-ups are identified.

4.1 Qualitative study of the success factors of start-up companies

A qualitative study was conducted to collect opinions, discover trends and better understand the motivation of Bulgarian entrepreneurs. The technique used is conducting in-depth interviews with a small number of non-representative cases, by asking questions against a prepared interview guide.

The interviewed entrepreneurs are presented in advance with the abstract and detailed model for predicting the success of start-up companies. The aim is for their responses and comments related to the model to help develop the model before proceeding with validation with quantitative methods and thus to obtain better results in predicting the success of start-ups using the model.

The interview was conducted based on the following interview guidelines:

1. What is the main field of activity of your business?
2. What is your position in the company?
3. How long have you been in business?
4. What do you think about the model for predicting the success of start-ups?
5. Do you think there are gaps in the model? What?
6. Do you think there are redundant criteria in the model? Who are they?
7. Do you think software for predicting the success of start-ups and young companies would be useful for your business? With what?
8. What do you think about the individual criteria from the model: are they understandable, are they measurable, could you evaluate them, could you add something?
9. How do you determine which start-up is successful? How do you distinguish a successful start-up company from an unsuccessful one?

The respondents are five Bulgarian entrepreneurs and owners of young companies in the fields of information technology, medicine and agriculture. The companies were established between 1 and 5 years ago and all of them are active. All companies use their founders' own funds for seed funding and are founded without the help of business incubators, venture capital funds, business angels or the government.

The duration of each interview is from 0:30 to 2:30 hours, depending on the answers and the demonstrated interest of the interviewees. The results of the interviews are presented in a summarized form in Table 8.

Table 8Summary results of the qualitative survey of opinions in relation to the proposed model for predicting success

Question #	Qualitative research results
1	The companies' fields of activity are information technology, medicine and agriculture.
2	All interviewees are business owners.
3	The companies were founded 1 to 5 years ago and all of them are active.
4	All interviewees described the model for predicting the success of new companies as logical and correct.
5	Respondents did not find gaps in the model, but suggested numerous improvements and additions to the individual criteria.
6	Respondents did not find redundant data in the model.
7	All interviewees believed that software to predict the success of start-ups and young companies would be useful. Only some of them are willing to pay for such software, but all would use it provided it is free. The respondents would use the software to help their start-up companies, initial research, to evaluate the company, to locate and improve weak points. Some of the respondents would trust the software, but others need evidence: information about the algorithm and logic of work, details and explanations about the model.
8	Respondents described most criteria as logical, clear and measurable. They are able to understand and answer most questions of the presented draft survey without difficulty. However, respondents felt uncertain about many questions related to the team, the entrepreneur's personality, and the industry. None of them know their temperament type according to the Myers-Briggs test. All interviewees asked for a definition of management or marketing terms that they failed to understand or misunderstood. They suggested numerous improvements and additions to the individual criteria.
9	Respondents defined successful start-ups as those that have growth and development. According to them, in the initial stages of the company's development, when a product or service is being developed and a team is formed, financial indicators are not as important as the progress achieved.

All respondents describe the model as logical and do not find significant gaps. I. M. - a web design company owner said "The model is logical and does not have any significant gaps. It has improved by adding additional categories to the traditional model." V. V., owner of an internet trading company, believes that "The model covers the main areas". Respondents identified all categories as important to the firm's success and found no redundant data. L. M., owner of an organic farming company, believes that "There is nothing to remove". Based on these responses, a decision was made to retain the proposed categories in the abstract model.

Respondents were shown the individual criteria representing the subcategories in the model diagram, as well as a draft of the survey. They were able to understand and answer most

questions without difficulty. However, they felt uncertain about a number of issues related to the team, the entrepreneur's personality, and the industry. I. M. shared that for start-ups the team is not important to the success of the firm, only the core management team is important and suggested providing a definition of an entrepreneurial team. The interviewees did not have information about their temperament type (according to the Myers-Briggs test and Carl Jung's classification) (Jung, 1971), (Briggs, 1995). Some of them did not understand or misinterpreted marketing and management terms such as "product portfolio". They requested that definitionsof **all** terms be provided so that they can understand the questions properly. Respondents identified human resource (entrepreneurial team), customers, unique product and competitors (and also competitive analysis) as the most important factors for the success of the young company. They suggested **numerous improvements and additions** to the draft survey, which have been taken into account in the detailed model and in the quantitative study.

Based on the entrepreneurs' suggestions for the criteria in the detailed model for predicting success, numerous changes have been made to it. All the terms that made it difficult for entrepreneurs have been translated into Bulgarian and explained in the survey so that they can answer the questions correctly. The question on temperament according to the Myers-Briggs classification was removed and the original idea of conducting a temperament test as part of the survey was rejected. Instead, other questions defining the psychological profile of the entrepreneur have been proposed. Many of the questions have been revised to sound more specific and be clearly measurable.

All respondents demonstrate **an interest in the study** and share that software **to predict the success of start-ups and young companies would be useful to them.** However, some of them remain reserved about the practical implementation, usability and logic. I. M., founder of a web design company believes that "The results will not be taken seriously until a logical explanation and a list of weaknesses are provided. I need the weaknesses to make improvements in the business." P. P., owner of a specialized medical equipment distribution company claims the following: "The software would be useful. It would focus the entrepreneur's attention on details, such as the team and the strategy." PM, owner of a software development company shared, "It's good to have such a product on the market, and I'll use it if it's free, but if it's paid, I'd rather do my own research."

Interviewed business owners expressed a clear interest in the research and a need for software to predict the success of young companies and to suggest possible business improvements. This is proof that **Bulgarian entrepreneurs need a similar tool**to help them develop their companies. Business owners have a positive attitude and are pro-active because they see potential benefits in research. They are happy to share their experience and wish to help improve the business climate in Bulgaria.

On the basis of these answers, a number of improvements are foreseen in the design and development of the software for predicting the success of start-up companies. It is possible to graphically display the model on which the forecast was made. The success prediction model obtained through the application of quantitative methods is easy to understand and represents a classification tree, and thus entrepreneurs can analyze for themselves the algorithm on which the prediction for their business is built. On the classification tree, the specific factors that led to the corresponding forecast result are marked and can be considered as opportunities for improvement if the forecast result is not satisfactory. The software is intended to give a percentage prediction of the company's probability of success, not just a final decision.

The conclusion of the conducted qualitative research is that the proposed model for predicting the success of new companies is appropriate according to the interviewed Bulgarian

entrepreneurs and is accepted by them as logical and complete. They identified the proposed category of "resources" as an integral and important part of the model. The entrepreneurs suggested improvements to the model, which were implemented both in it and in the quantitative study. The success prediction software is tailored to the recommendations and requirements received.

4.2 Quantitative Study of the Success Factors of Start-up Companies

The study is based on the data mining methodology CRISP-DM. The survey data were analyzed using the software products IBM SPSS Modeler and Weka, which generated success prediction models (eg, a classification tree) that gave a practical dimension to the theoretical prediction model. Through IBM SPSS Statistics, a factor analysis was prepared, through which the groups of factors from the proposed hypothesis for a model for predicting the success of start-up companies were validated.

In order to validate the model for predicting the success of start-up companies,a quantitative study was conducted using an online survey, through which data were collected data. The survey was designed by selecting for each of the factors from the abstract success prediction model a set of relevant questions that best characterize that factor. The answers to these questions make up the quantitative research data. The processing and analysis of the survey data will serve to validate the model for predicting success and to extract knowledge from data – key success factors and models based on the survey data.

Given the size of the data set from the surveys in other countries, as well as the fact that the intended survey is detailed and takes a significant amount of time to complete (about 20 minutes), **the initial objectives** were as follows:

- To conduct a survey using a survey containing about 100 the question,
- Conducting the survey within 6 months,
- Completing the survey by at least 100 representatives of the target group - entrepreneurs, business owners, managers of start-up small and medium-sized companies.

The intended survey analyzes in detail multiple factors of start-up success. Questions are marked as optional in order to get as many answers as possible and allow respondents to skip difficult or unwanted questions. The time to complete the survey by one respondent is about 15-20 minutes.

The target group (survey respondents) are owners, managers and managers of small and medium-sized companies that are start-ups or young. According to the definitions proposed in Chapter 2, start-up and young companies are those which are **at the beginning of their development and are SME**(micro, small or medium-sized companies) in size. For small and medium-sized companies, we accept those with 0to 250 employees (according to the European Union definition). The number of companies falling into the target group would be important for the research, but unfortunately NSI does not provide such a sample for Bulgaria.

Approximately **the number of companies falling into the target group**will be determined from the available data from NSI. According to enterprise demography (NSI, National Statistical Institute - Republic of Bulgaria, 2013), published by NSI, in 2012 there were 323,745 active enterprises in Bulgaria. Unfortunately, at the moment (year 2014) no more up-to-date data is available. Of all operating enterprises, only 54% (or about 174,500) have at least one employed person. This fact, the possibility that some enterprises are not active, as well as the presence of a gray economy, raise doubts about the real number of active enterprises in Bulgaria. 91% of enterprises have up to 10 employees, and only 0.1% have

more than 250 employees. Unfortunately, the number of enterprises falling into the target group is not reported. In 2012, 13% of the total number of enterprises were created, of which 81% survived a year later. From these data, it is obviously not possible to determine exactly the number of enterprises falling into the target group, but only to estimate approximately. According to the NSI reports, I calculated that **the number of enterprises falling into the target group at the end of 2012 was between 106,000 and 193,000**.

According to a study done by the author, in similar surveys to validate models for predicting the success of start-up companies, the number of respondents varies between 33 and 334, depending on the specifics of the survey and the size of the target group. The following is a list of similar studies conducted in different countries, arranged in chronological order:

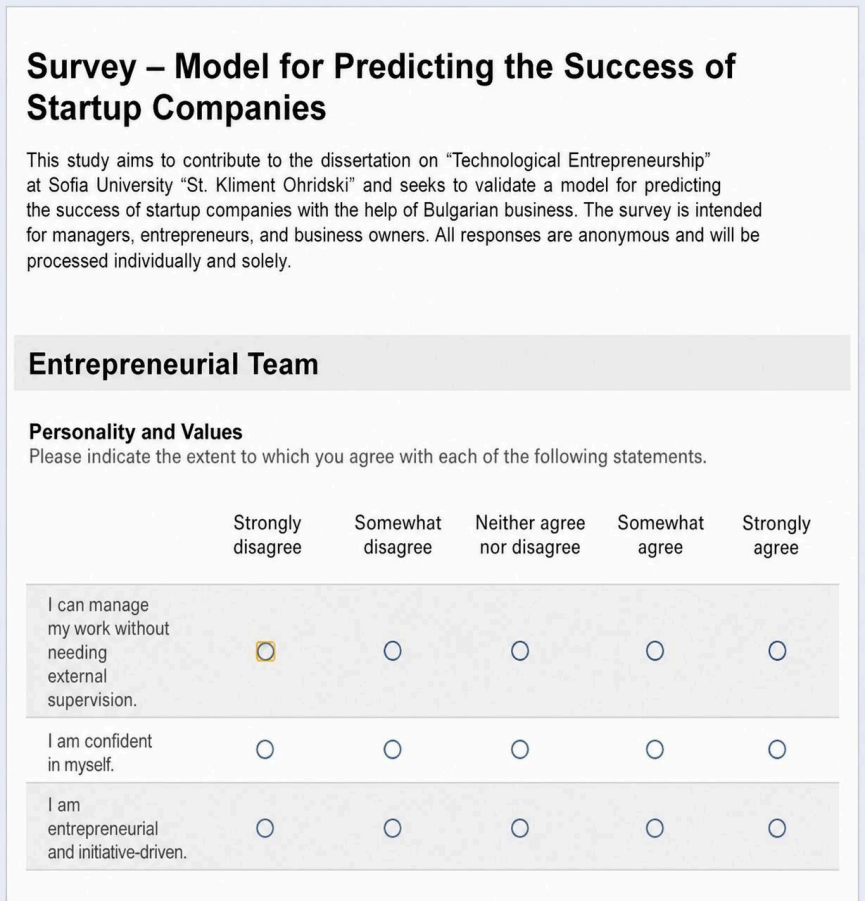
- 46 US venture capitalists, (Tyebjee, 1984),
- 97 US venture capital firms, (Poindexter, 1975),
- 155 manufacturing enterprises (Chandler G. H., 1994),
- 225 companies from the USA, (Carland, The Theoretical Bases and Dimensionality of the Carland Entrepreneurship Index, 1996),
- 63 firms from China, (Luo, 1999),
- 137 firms from Korea, (Park, 2000),
- 56 companies from China, (Peng M. L., 2000),
- 334 US Small Manufacturing Enterprises, (Ebben, 2005),
- 33 firms from USA (Abbas, 2008),
- 162 firms from Russia (Shirokova, 2010),
- 334 companies from Bulgaria - study of social capital factors and competitive strategies, (Manev, 2012).

By means of a survey, for a period of 12 months (in 2013 - 2014), data was collected **for 142 companies from Bulgaria,** which is within the set goals.

Multiple sources and channels were used to collect data for the survey conducted by the author:

- personal business contacts,
- online social networks (Facebook, LinkedIn, Google+),
- business and entrepreneurial communities (students, graduates of technology entrepreneurship, participants in competitions for educational companies and others),
- email marketing to active Bulgarian companies,
- personal website http://byankov.com/,
- advertising campaign on websites through the platform Google AdWords.

The survey conducted by the author is internet based to allow maximum accessibility for respondents. It was implemented through the tool Google Forms (Google Inc., 2014), which allows creating surveys, collecting and storing data from them and exporting the data in file format . CSV. The survey consists of 5 internet pages with about 100 questions to be filled in of the types of choosing one of many options, choosing several of many options, text field. An excerpt from the first page of the survey is shown in Fig. 15.



Survey – Model for Predicting the Success of Startup Companies

This study aims to contribute to the dissertation on "Technological Entrepreneurship" at Sofia University "St. Kliment Ohridski" and seeks to validate a model for predicting the success of startup companies with the help of Bulgarian business. The survey is intended for managers, entrepreneurs, and business owners. All responses are anonymous and will be processed individually and solely.

Entrepreneurial Team

Personality and Values

Please indicate the extent to which you agree with each of the following statements.

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I can manage my work without needing external supervision.	○	○	○	○	○
I am confident in myself.	○	○	○	○	○
I am entrepreneurial and initiative-driven.	○	○	○	○	○

Fig. 15Part of the start-up success prediction model validation survey (created by the author)

A standard 5-level Likert scale (Likert, 1932) was chosen to answer most of the survey questions. The questions are formulated as statements, and the possible responses are the degrees of agreement: "Strongly disagree", "Rather disagree", "Neither agree nor disagree", "Rather agree" and "Strongly agree". The Likert scale should be symmetrical and the 5-point option is preferred because the 3-point option is imprecise (does not express degrees of agreement) and the 7-point option is more complex for respondents and may cause difficulty for them. In addition, the 5-level Likert scale is the most popular.

4.2.1 Analysis of the Quantitative-Study Data

In order to determine the types and profiles of the companies that completed the survey, the collected data will be reviewed and analyzed.

4.2.1.1 Size of companies

The majority of companies (136) are micro, small and medium enterprises (SME), and only 3 are large companies (Fig. 16). This is an expected distribution of companies by size, considering the examined data from NSI. As a profile, start-up micro companies predominate. The total number is less than 142 because some respondents did not specify company size.

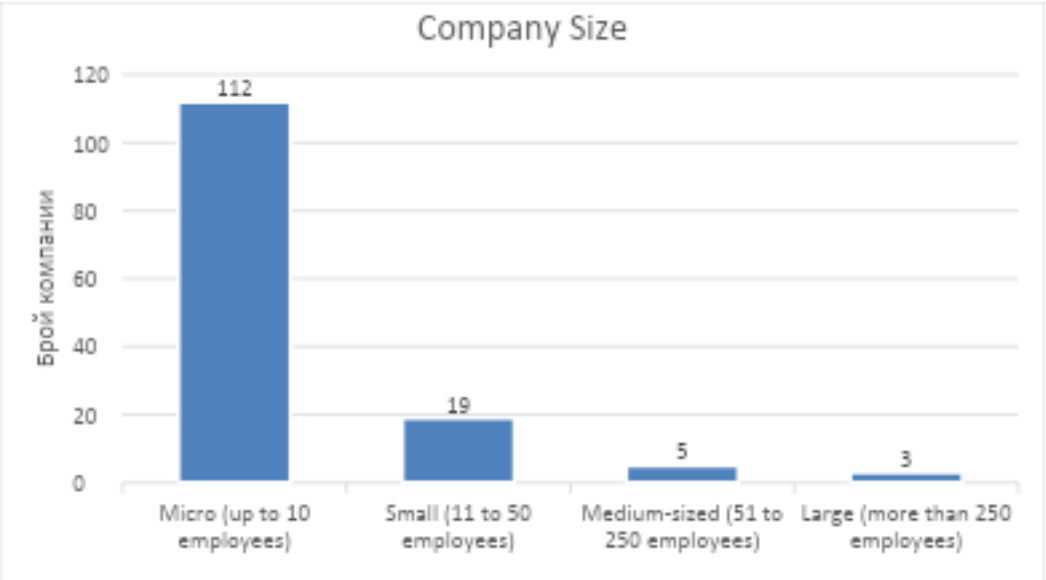


Fig. 16Company size (created by author)

4.2.1.2 Age of the companies

Of the companies participating in the survey, those between 1 and 5 years old prevail (62 in total), but there are also companies over 5 years old (41), as well as companies that are just starting their activities (33). The total number is less than 142, because part of the respondents did not indicate the age of the company (Fig. 17).

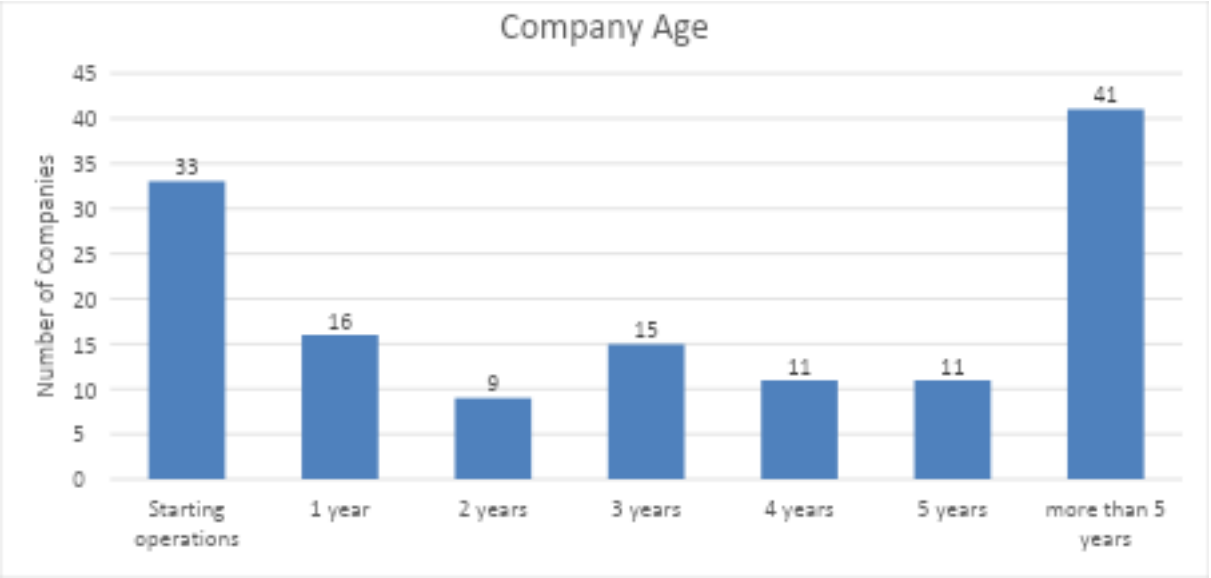


Fig. 17 Age of companies (created by author)

4.2.1.3 Industrial sector of the companies

The industrial sectors of the companies were checked according to the NSI classification (National Statistical Institute - Republic of Bulgaria, 2014) so that the data could be compared

if necessary. From the represented companies, most of them work in the spheres "Creation and distribution of information and creative products; telecommunications" (25 companies), "Trade; repair of cars and motorcycles" (11 companies), "Education" (10 companies). The exact distribution of industrial sectors is shown in Fig. 18.

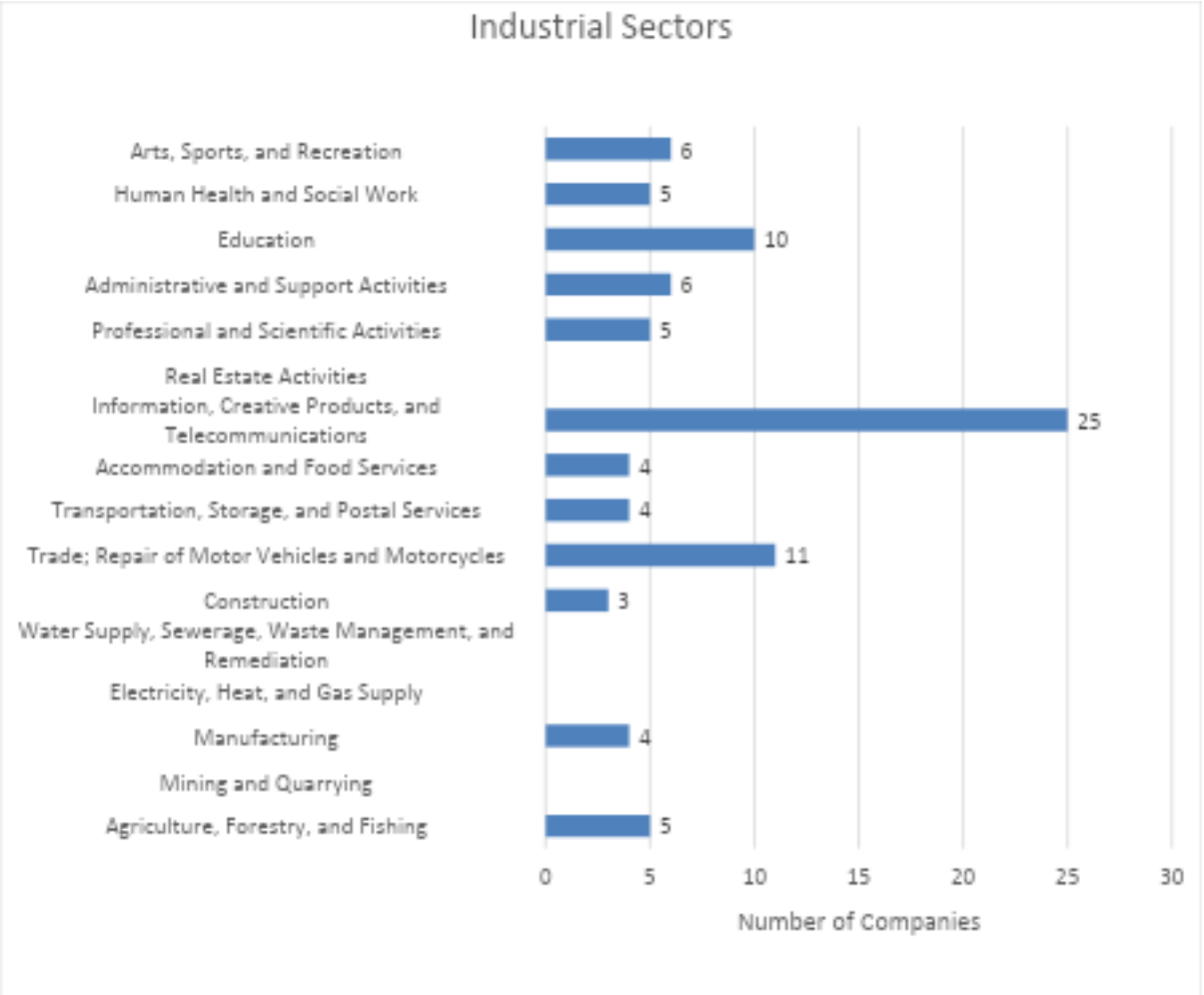


Fig. 18Industry sector of companies (created by author)

4.2.1.4 Gender of respondents

The data includes survey completions from entrepreneurs of both sexes, with males (90) predominating over females (48) (Fig. 19).

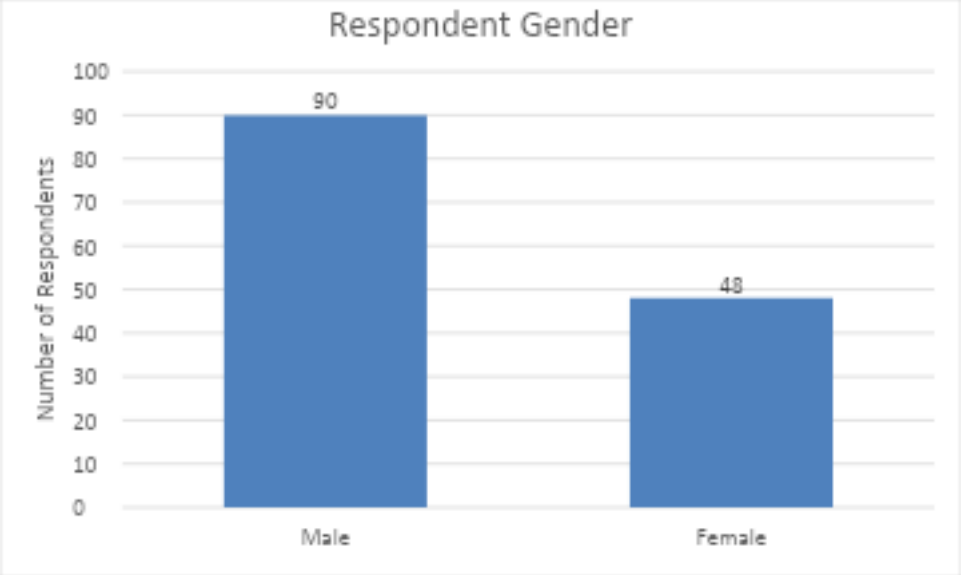


Fig. 19Gender of the respondents (created by the author)

4.2.1.5 Age of the respondents

The data set is dominated by entrepreneurs aged 30-39, 20-29 and 40-49 years, but entrepreneurs from all age groups are represented (Fig. 20). The division into age groups is organized in the same way as the NSI samples to allow comparisons.

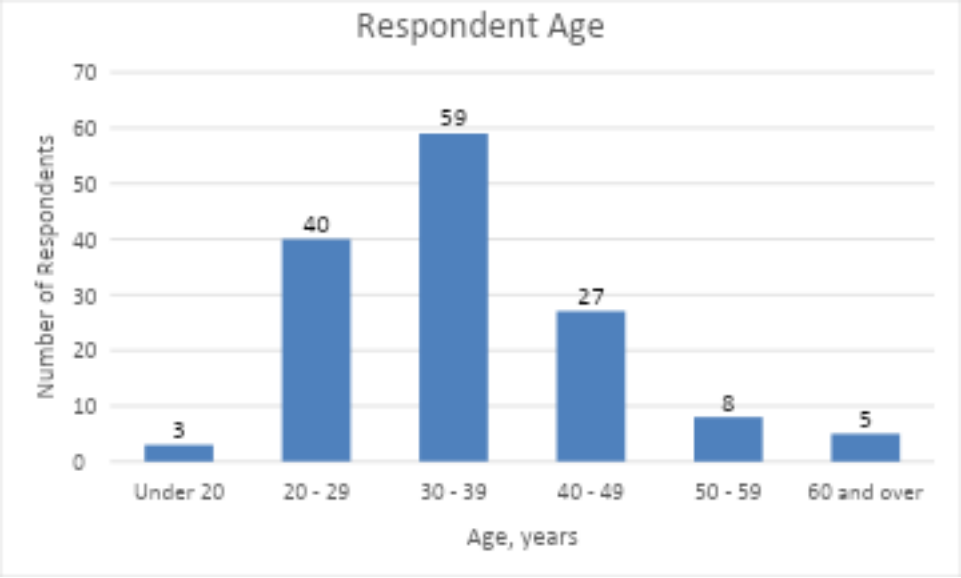


Fig. 20Age of the respondents (created by the author)

4.2.2 Preparation of the Quantitative-Study Data

The data collected through the survey on Bulgarian start-up companies includes a set of valid entries from 142 companies. When they are received from completed surveys, they are saved in an electronic file stored in the Google drive service. The data as received is not suitable for

analysis and needs preliminary preparation. To be comprehensible to data mining and statistical analysis software, they must be prepared in a specific way. Fields in the database must be assigned a specific type (nominal, ordinal, continuous) and the data in them must be converted to numeric values.

From Google drive it is possible to export (export) the data in file format CSV, which contains tabular data. CSV format is text and contains one row of data for each row of the table (corresponding to one survey completion). Within each row, the data in individual columns is separated by a special character. The first line in the file contains the column headings. The survey data, viewed in Google drive, looks as shown in Fig. 21.

	A	B	C	D	E	F	G
1	Timestamp	Personality and values [I can handle challenges at work when I need to take action.]	Personality and values [I believe in myself.]	Personality and values [I am entrepreneurial, innovative.]	Personality and values [I would rather depend on myself in life.]	Personality and values [I am ambitious.]	Personality and values [I analyze and assess risks.]
2	6/23/2013 9:05:51	Strongly agree	Strongly agree	Completely agree	Completely agree	Completely agree	Strongly agree
3	6/27/2013 6:16:47	Strongly agree	Completely agree	Completely agree	Completely agree	Completely agree	Completely agree
4	6/27/2013 23:10:36	Strongly agree	Neither agree nor disagree	Strongly agree	Completely agree	Strongly agree	Strongly agree
5	6/28/2013 11:06:40	Completely agree	Completely agree	Strongly agree	Strongly agree	Strongly agree	Strongly agree

Form Responses

Fig. 21Sample of survey data viewed in Google Drive (author created)

4.2.2.1 Preparation of Survey Data in Excel

The exported survey data file is in CSV format and can be read by the Microsoft Excel software product (Microsoft Excel - Spreadsheet Software, 2014), which was used for the bulk of the data preparation. Through formulas in Excel, the data has been processed so that the individual values have a numerical form, with each number having a specific meaning. The majority of the questions are on a 5-point scale of agreement from "Strongly Agree" to "Strongly Disagree", which is represented numerically as follows:

1 => Strongly Disagree

2 => Rather Disagree

3 => Neither agree nor disagree

4 => Rather agree

5 => Strongly agree

Since in the survey the answers are presented in the form of text, the conversion to numerical data is done using the following Excel formula (Form. 4):

=SUBSTITUTE(*(4)*
SUBSTITUTE(
SUBSTITUTE(
SUBSTITUTE(
SUBSTITUTE(
'Form Responses'!B2;
"Strongly agree";
"5");
"Rather agree";
"4");
"Neither Agree nor Disagree";
"3");
"Rather Disagree";
"2");
"Strongly Disagree";
"1")

The survey also contains other types of questions that are converted by other formulas – a total of 49 different formulas that are applied to about 100 variables in the set from data.

The data also includes a small number of questions whose answers contain free text (open questions). This is necessary in order to explore the exact opinion of entrepreneurs on important issues without setting prior restrictions. An example of such a question is "I believe the following factors are key to the success of my business:", where entrepreneurs have the freedom to list the factors that are important to them. The preparation of this data is more difficult as all the responses have to be read and analyzed and then grouped appropriately. Since a respondent may indicate none, one or more than one factor, each group of factors is presented as a separate variable in the processed data. The presence of this factor for the respective respondent is marked with one value of the variable - in the case of "1", and the absence of this factor - with "2". A value of zero is avoided as an option because some data mining software and their algorithms might interpret it as missing data. In order to at least partially automate the processing of the data, as well as the processing of future survey completions, each group of factors is represented by keywords and phrases, the presence of which in the data is checked by a formula. For example, Formula 5 is used for the team factor:

```
=IF(                                                    (5)
AND(
 ISERROR(SEARCH("team";'Form Responses'!BI2));
 ISERROR(SEARCH("people";'Form Responses'!BI2));
 ISERROR(SEARCH("employee";'Form Responses'!BI2));
 ISERROR(SEARCH("human";'Form Responses'!BI2));
ISERROR(SEARCH("staff";'Form Responses'!BI2));
 ISERROR(SEARCH("professional";'Form Responses'!BI2))
);
"2";
"1"
)
```

The survey also contains two questions that determine whether a start-up is successful or not. The wording of these questions is based on the definition of a successful start-up company adopted in this dissertation. These are the questions of whether the company is still in business at the present time and whether it has registered growth in the last 5 years. Their answers divide the firms into three categories: "successful", "neither successful nor unsuccessful" and "failed". All other questions and answers in the survey describe the characteristics of a given company, through independent variables.

After processing the data in Excel the rows of the table corresponding to filling in the survey contain only numbers corresponding to answers to the questions (Fig. 22).

	A	B	C	D	E	F	G	H
1	Personality and values [I can manage my work without the need for others' assistance.]	Personality and values [I am confident in myself.]	Personality and values [I have entrepreneurial spirit, initiative.]	Personality and values [The things that happen in my life depend on me myself.]	Personality and values [I am ambitious.]	Personality and values [I tend to take risks.]	Personality and values [I do not worry about taking decisions without consulting others.]	Personality and experience [Magic skills (skills in starting a business).]
2	4	4	5	5	5	4	4	4
3	4	5	5	5	5	5	5	5
4	4	3	4	5	4	4	4	4
5	5	5	4	4	4	4	3	3
6	5	4	4	5	4	4	5	5
7	4	4	5	5	4	4	4	4
8	4	5	5	4	5	4	1	5
9	5	5	5	5	5	5	4	5
10				5		4	2	4

Fig. 22Sample of survey data, after processing in Microsoft Excel (author created)

4.2.2.2 Preparation of survey data in IBM SPSS Statistics

In order for survey data to be properly read and interpreted by IBM SPSS Statistics and IBM SPSS Modeler, it needs to be saved in the IBM SPSS Statistics file format (IBM SPSS Statistics, 2014), which is of type ".sav". For this purpose, the data was copied from Excel and inserted into IBM SPSS Statistics, which recognizes them as values - numbers (Fig. 23).

	Leadership				Opportunities			Organization				Compensation		Workplace		Environment		Values			Overall satisfaction		Targeted position company name	S	C
	Str.	Vis.	Mot.	Com.	Train.	Dev.	Career	Str.	Proc.	Inno.	Team	Pay	Ben.	Safe.	Res.	Phys.	Soc.	Eth.	Val.	Diversity	Job	Life			
1	4	4	5	5	4	4	4	4	4	3	4	5	5	3	1	4	4	5	4	3	4	2	150000	4	3
2	4	5	5	5	4	5	5	4	4	3	5	5	2	2	4	5	5	5	5	5	5	2	50000	5	4
3	4	3	4	4	4	5	4	4	4	3	4	6	2	2	4	4	5	4	3	4	5	1	12800	4	3
4	5	5	4	4	3	3	5	3	3	2	5	3	1	3	3	5	4	4	4	5	5	4	50000	4	3
5	5	4	5	4	4	5	3	3	5	2	5	6	2	2	4	4	4	4	2	4	4	2	10000	3	2
6	4	4	5	4	4	3	4	3	3	4	5	6	1	3	2	2	3	4	2	4	0	1	50000	2	3
7	5	5	5	5	4	5	4	3	4	5	4	1	5	2	4	4	5	4	5	4	5	...	60000	5	4
8	0	0	5	0	4	4	4	4	4	1	1	5	2	3	4	5	3	4	4	4	0	0	0	4	3
9	4	5	5	5	2	5	5	4	2	5	5	3	2	2	4	5	5	5	4	4	5	7	200000	4	3
10	4	4	5	4	3	3	4	3	4	4	3	1	3	3	4	5	4	3	3	4	3	3	0	3	2
11	3	4	4	4	4	4	4	3	4	5	3	1	3	5	5	5	3	4	2	4	5	7	0	5	4

Fig. 23Sample survey data entered into IBM SPSS Statistics (author created)

In addition to transferring the data to IBM SPSS Statistics, it is also necessary to set their type and format (in the form of metadata to each variable). Each of the numerical values is also assigned an appropriate label that is easier for the user to read. Each variable is assigned a role: "input" - variables used as input data for the analysis or "target" - target variables whose values we analyze. The view of the variable columns is shown in Fig. 24.

	Name	Type	Width	Deci...	Label	Values	Missing	Columns	Align	Measure	Role
91	Non-material active assets	Numeric	1	0	Non-material active assets	{1, Fully intangibl...	0	1	R...	Ordinal	Input
92	Own equity	Numeric	1	0	Own equity	{1, Yes}...	None	1	R...	Nominal	Input
93	Average household size	Numeric	1	0	Average household size	{1, Yes}...	None	1	R...	Nominal	Input
94	Family type	Numeric	1	0	Family type / source of funds	{1, Yes}...	None	1	R...	Nominal	Input
95	Loan / credit - principal	Numeric	1	0	Loan / credit - Principal	{1, Yes}...	None	1	R...	Nominal	Input
96	Business incubator - principal	Numeric	1	0	Business incubator - Principal	{1, Yes}...	None	1	R...	Nominal	Input
97	Individual investment - principal	Numeric	1	0	Individual investment - Principal	{1, Yes}...	None	1	R...	Nominal	Input
98	Investment fund - principal	Numeric	1	0	Investment fund - Principal	{1, Yes}...	None	1	R...	Nominal	Input
99	Venture capital fund - principal	Numeric	1	0	Venture capital fund - Principal	{1, Yes}...	None	1	R...	Nominal	Input
100	Partnership - principal	Numeric	1	0	Partnership - Principal	{1, Yes}...	None	1	R...	Nominal	Input
101	Individual investment - new	Numeric	1	0	Individual investment - New	{1, Yes}...	None	1	R...	Nominal	Input
102	Position in the company	Numeric	1	0	Position in the company	{1, Employee}...	None	5	R...	Ordinal	Input
103	What is your position?	Numeric	1	0	What is your position?	{1, Managemen...	None	1	R...	Nominal	Input
104	What stage is your company at?	Numeric	1	0	What stage is your company at?	{1, Start-up}...	None	1	R...	Ordinal	Input
105	What is the current state of the company?	Numeric	1	0	What is the current state of the company?	{1, Micro (up to ...	None	5	R...	Ordinal	Input
106	Company is expanding	Numeric	1	0	The company is expanding	{1, Yes}...	None	5	R...	Nominal	Target
107	Company is growing	Numeric	1	0	The company is growing	{1, Yes}...	None	5	R...	Nominal	Target
108	Company is successful	Numeric	1	0	The company is successful	{1, No}...	0	9	R...	Ordinal	Target

Fig. 24Sample of survey variables in IBM SPSS Statistics (created by the author)

4.2.2.3 Preparing survey data in Weka

In order to read the data from the software product Weka (Bouckaert, 2013), they need to be converted to ".arff" file format. This format contains the data in numeric form separated by commas (".csv"), which is a directly supported format in Microsoft Excel. Additionally, the ".arff" file contains metadata describing the variables - their names, types, and valid values. In Fig. 25shows the survey data loaded into the software product "Weka".

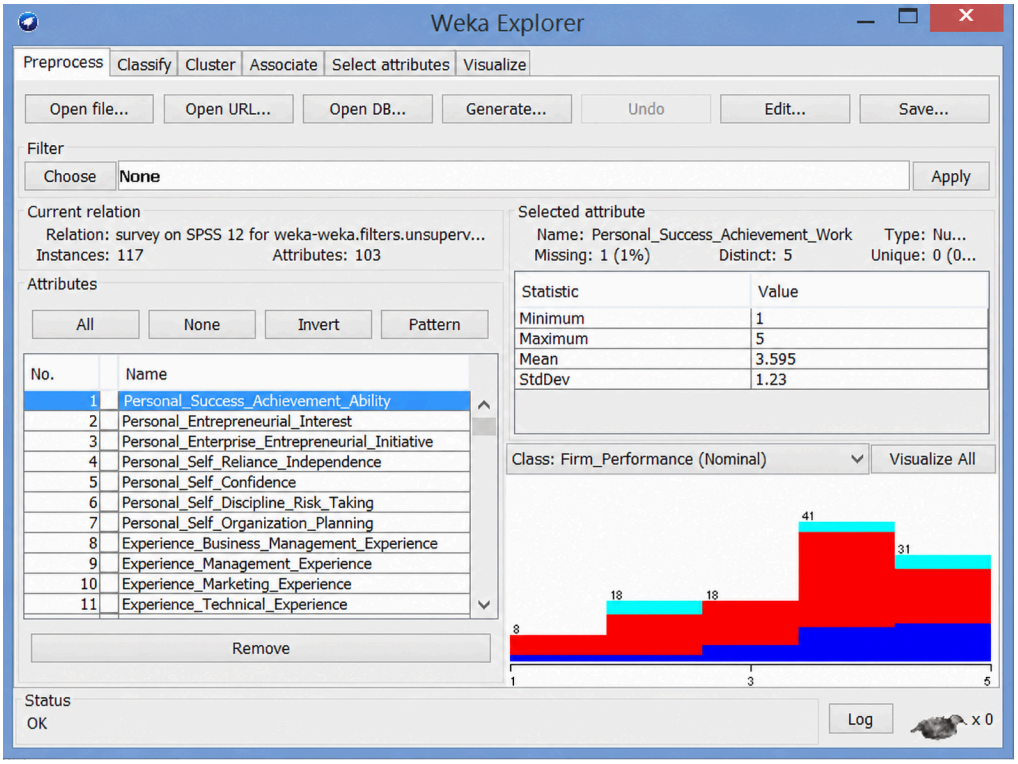


Fig. 25Sample data loaded into the Weka software product (created by the author)

4.2.3 Grouping success factors

Factor analysis (Kharalampiev, 2012) is a statistical method created by Charles Spearman and used to highlight the factors that are common to a set of variables between which correlations exist. Factor analysis will be applied to the variables that represent the success factors of start-ups to obtain clusters of factors. The software product that will be used for the analysis is IBM SPSS Statistics.

As a first step of the analysis, some variables were excluded because the method only works with ordinal indicators. Missing values are small in number but occur in almost all indicators, so the method of replacing missing values with mean values was chosen in order to perform the factor analysis. Sampling adequacy was checked for all factors as well as for each factor separately. Sampling adequacy was unacceptable due to the presence of a large number of factors, some of which had poor sampling adequacy. These factors were excluded in the next iteration of the analysis. All factors excluded from the analysis are listed in Appendix 2.

In the next iteration of the analysis, sampling adequacy according to the Kaiser-Meier-Olkin test (Kaiser, 1974) is acceptable (Table 9).

Table 9Kaiser-Meyer-Olkin and Barlett test

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.697
Bartlett's Test of Sphericity	Approx. Chi-Square	3154.260
	Df	1431
	Sig.	.000

After setting 11 groups and applying rotation to the obtained correlation matrix, the distribution of the factors by groups (Table 10). The factors are located in the left column of the table, and the groups are labeled as components (Components) and numbered from 1 to 11. The numbers that associate the factors with the components indicate the degree to which the factor relates to the corresponding group, with larger numbers indicating a stronger correlation. Correlations below 0.5 are not shown in the table.

The eleven components of the correlation matrix were generated by IBM SPSS Statistics and represent groups that are not titled because the software does not automatically indicate their business sense, but only groups the factors presented. Based on the factors included in the components, their significance will be analyzed.

Table 10Correlation matrix, after applying rotation

Rotated Component Matrix[a]											
	Component										
	1	2	3	4	5	6	7	8	9	10	11
Personality and values [I can manage my work without the need for assistance.]											
Personality and values [I am self-confident si.]		0.754									
Personality and values [I am enterprising, initiative.]		0.761									
Personality and values [I am ambitious.]		0.683									
Personality and values [I tend to take risks.]		0.643									
Skills and experience [I have entrepreneurial skills (skills in starting a business).]											
Skills and Experience [I have management skills.]				0.775							
Skills and Experience [I have marketing skills.]				0.784							
Skills and Experience [I have HR skills.]				0.777							
Skills and Experience [I have HR skills. investing.]				0.546							
Skills and experience [I have previous experience in starting a business.]					0.645						

Skills and experience [I have experience in a similar position (to the one currently held).]					0.782						
Skills and experience [I have experience in the same industry.]					0.748						
Skills and experience [I have management experience.]					0.695						
Teamwork [The company has a complete team.]			0.743								
Teamwork [The team has the necessary knowledge.]			0.883								
Teamwork [The team has the necessary skills.]			0.886								
Teamwork [The team has the necessary positive attitude.]			0.775								
Planning [I plan covering a wide area.]	0.635										
Planning [I plan thoroughly.]	0.768										
Planning [I make a formal (written) strategic plan.]	0.653										
Planning [I plan often.]	0.633										
Planning [I plan by functional areas / departments.]	0.659										
Planning [I consider multiple alternatives when planning.]	0.659										
Goals [The goals I set for my business are ambitious.]											
The target market share of the company is:											
Strategic direction [I manage to stick to the initially chosen strategy.]											
Strategic direction [The strategy I have chosen is aggressive.]											
Strategic direction [The target strategy is detailed.]	0.704										
Strategic direction [I have a clear strategy for my business.]	0.653										
My business enters the market at the following stage:										-0.515	
Competitive advantage [My business has a clear competitive advantage.]											
Strategic partnerships [The company engages in partnerships with competitors.]											
Strategic Partnerships [The firm engages in partnerships with customers.]						0.689					
Strategic Partnerships [The firm engages in partnerships with suppliers.]						0.728					
Strategic Partnerships [The firm engages in partnerships						0.699					

with other stakeholders. countries.]											
Characteristics of the industry in which the firm operates [Most companies in the industry have a good profit.]									0.684		
Characteristics of the industry in which the firm operates [The industry in which the firm operates generates a large portion of gross domestic product (GDP).]									0.609		
Characteristics of the industry in which the firm operates [The industry has stable demand.]									0.634		
Characteristics of the industry in which the firm operates [The industry is in a period of growth.]									0.704		
Industry Competitiveness [Industry competitors are aggressive.]											0.821
Industry Competitiveness [The company is exposed to competitive attacks.]											0.7
Tangible assets [The company owns inventory.]								0.77			
Tangible assets [The company owns real estate.]								0.685			
Tangible assets [The company owns cash.]								0.504			
Intangible assets [The company has a recognizable brand.]							0.736				
Intangible assets [The company has software systems for managing business processes (ERP, CRM).]							0.647				
Intangible assets [The company has an established business reputation.]							0.619				
Intangible assets [The company owns licenses.]							0.636				
Intangible assets [The company has established business contacts (social network).]											
Intangible assets [The company has access to distribution channels.]											
Intangible assets [The company has access to labor markets.]										0.536	
Intangible assets [The company has access to capital markets.]											
Intangible assets [The company has access to materials and suppliers.]								0.539			
Extraction Method: Principal Component Analysis.											
Rotation Method: Varimax with Kaiser Normalization.											

a. Rotation converged in 10 iterations.

The obtained components of the factor analysis largely correspond to the groups of factors of the proposed abstract model for predicting success (Chapter 3). The first column of Table 11lists the categories and the second column lists the subcategories of the theoretical model for predicting success. The third column lists all groups obtained by factor analysis that have common factors with the corresponding subgroup of the theoretical model. As can be seen from the table, there is a clear correspondence and overlap for almost all of the subcategories from the theoretical model with groups from the factor analysis. The distinction of the individual subcategories and that they correspond to different groups of the analysis is clearly visible. When examining the main categories of the theoretical model, it can be seen that several groups from the factor analysis correspond to each of them.

Table 11Comparison of the categories from the theoretical model for predicting success and the resulting categories from the factor analysis

Category of factors from the theoretical model	**Subcategory factors from the theoretical model**	**Corresponding groups of factors from the factor analysis**
Entrepreneurial team	Personality and values	Component 2
	Skills and experience	Component 4 Component 5
	Teamwork	Component 3
Business strategy	Planning	Component 1
	Goals	-
	Strategic direction	Component 1
	Entry strategy on the market	Component 10
	Competitive advantage	-
	Strategic partnerships	Component 6
Structure of the industry	Characteristics of the industry in which the firm operates	Component 9
	Competitiveness of the industry	Component 11
	Characteristics of customers	-
Resources	Tangible assets	Component 8
	Intangible assets	Component 7 Component 8 Component 10
	Financing	-

The subcategories of the theoretical model for which there is no correspondence with groups from the factor analysis are:

- Goals
- Competitive advantage
- Characteristics of customers
- Financing

The reasons for the lack of compliance for the listed subcategories can be sought in several hypotheses:

- the fact that part of the variables that may contain important information were excluded from the analysis,
- the possibility of creating more groups that possibly correspond to the subcategories,
- the factors of the subcategories may actually have a correlation with some of the already created groups, but it is weaker and therefore not indicated in the correlation matrix,
- incorrect hypothesis of grouping the factors of the theoretical model into subcategories for which groups of the analysis.

Comparison of the eleven components of the correlation matrix generated by IBM SPSS Statistics, with their corresponding variables allows to determine their significance (Table 12). The components at the top of the table (with smaller numbers) represent the categories that are more clearly distinguished. It can be concluded that there is a distinction between expert skills (Component 4) and entrepreneurial and business management skills (Component 5). In general, the meaning of the components of the correlation matrix matches the hypothesis of an abstract model for predicting success(Chapter 3).

Table 12Meaning of the components of the correlation matrix

Components of the correlation matrix	**Meaning**
Component 1	Planning
Component 2	Personality and values
Component 3	Teamwork
Component 4	Skills and experience – expert
Component 5	Skills and experience – entrepreneurship and management business
Component 6	Strategic partnerships
Component 7	Intangible assets
Component 8	Material assets and access to materials
Component 9	Characteristics of the industry in which the firm operates

Component 10	Entry strategy on the market
Component 11	Competitiveness of the industry

Based on the prepared factor analysis and comparison of the obtained results with the theoretical model for predicting success from Chapter 3, it can be concluded that the proposed model is adequate and largely corresponds to the results obtained from the analysis. The complete analysis output is presented in Appendix 2.

4.2.4 Data Mining and Success-Model Development with IBM SPSS Modeler

This section examines the process of extracting knowledge from data and creating models to predict the success of start-ups with the software product IBM SPSS Modeler (IBM Corporation, 2012). The created models are analyzed and compared in terms of their working principles and efficiency. The software algorithms by which the most effective models were obtained are examined.

4.2.4.1 Synthesizing models for predicting success

Using the software product IBM SPSS Modeler success prediction models have been created that classify companies - predict which category they fall into: "successful", "neither successful nor unsuccessful" or "failed". The software product offers a tool that automatically generates multiple models that solve the given task, compares them and suggests the best of them.

In IBM SPSS Modeler the sequences of actions performed by the software can be represented graphically (Fig. 26). The sequence at the top of the figure shows model acquisition by the automatic classifier, and the sequence at the bottom shows model acquisition by the C5.0 algorithm with manual adjustments and applying cross-validation.

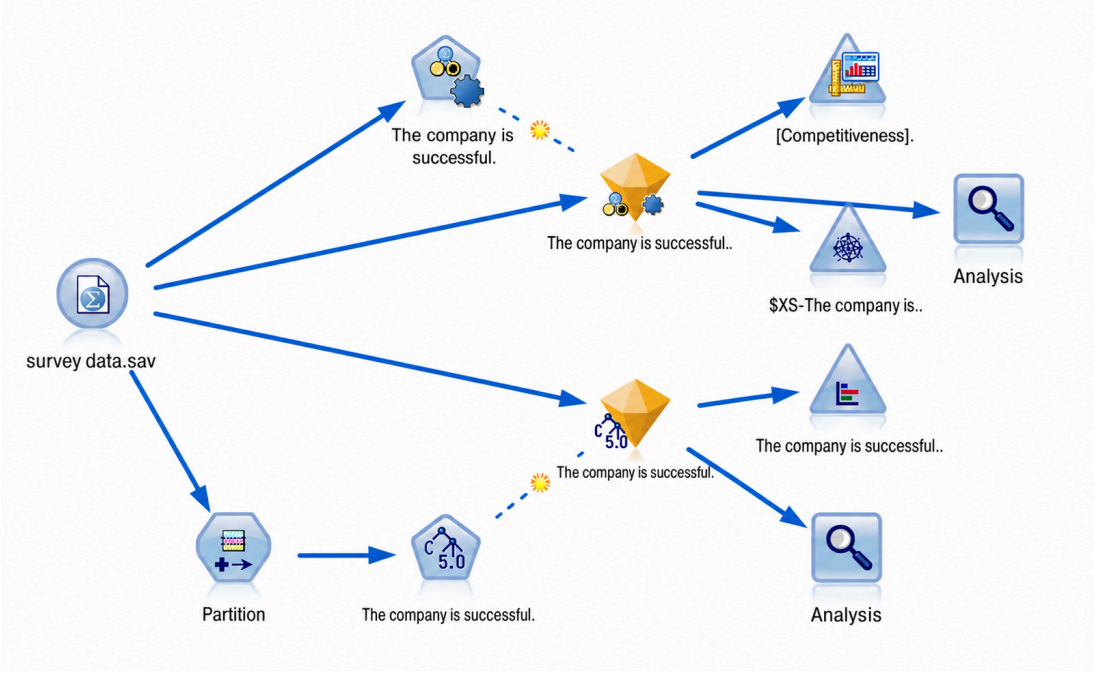


Fig. 26Sequence of actions in IMB SPSS Modeler to obtain classification models - graphical representation (created by the author)

Survey data are stored in an IBM SPSS Statistics file (.sav file format) and loaded into IMB SPSS Modeler via the Statistics File item located on the left side of the figure and labeled survey data spss. The Auto Classifier unit (pentagon-shaped at the top of the figure) synthesizes, evaluates and compares classification models for the selected target using different algorithms. The resulting models for predicting success can be viewed in the diamond-shaped unit (model nugget) located at the top of the figure.

The automatic classification unit is set up by selecting a target and input data (Fig. 27). When synthesizing for models to predict success, I choose the company success variable as the target, indicating how successful the company is. The input data are the remaining independent variables whose types and data are prepared in Excel and IBM SPSS Statics.

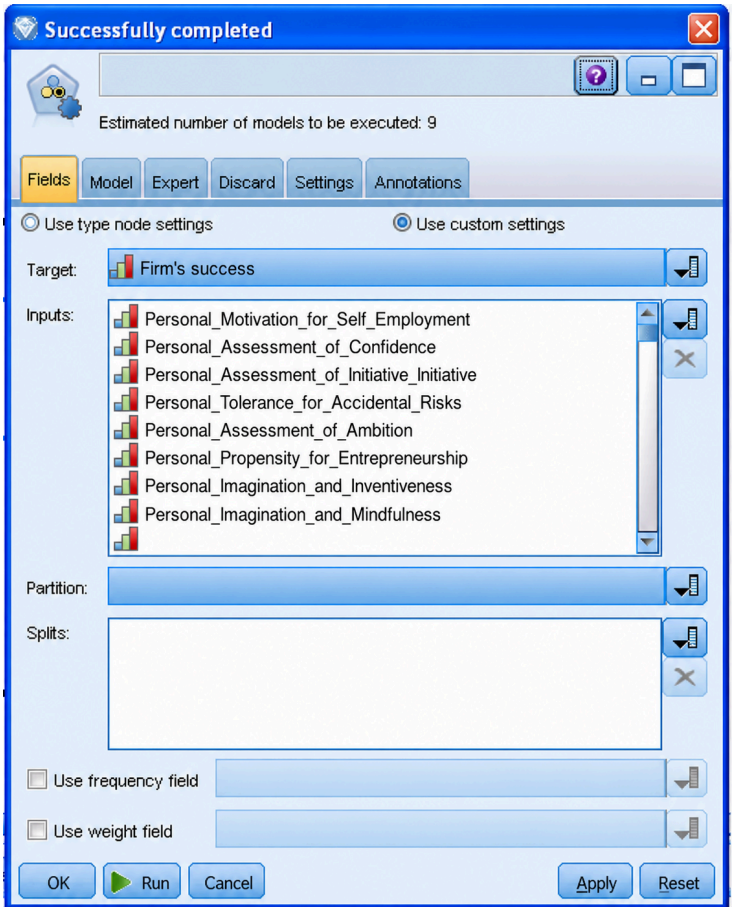


Fig. 27Selection of fields for the automatic classifier - input and objectives (created by the author)

The automatic classifier synthesizes and compares the resulting models and saves the best three success prediction models (those with the highest accuracy) for further analysis (Fig. 28).

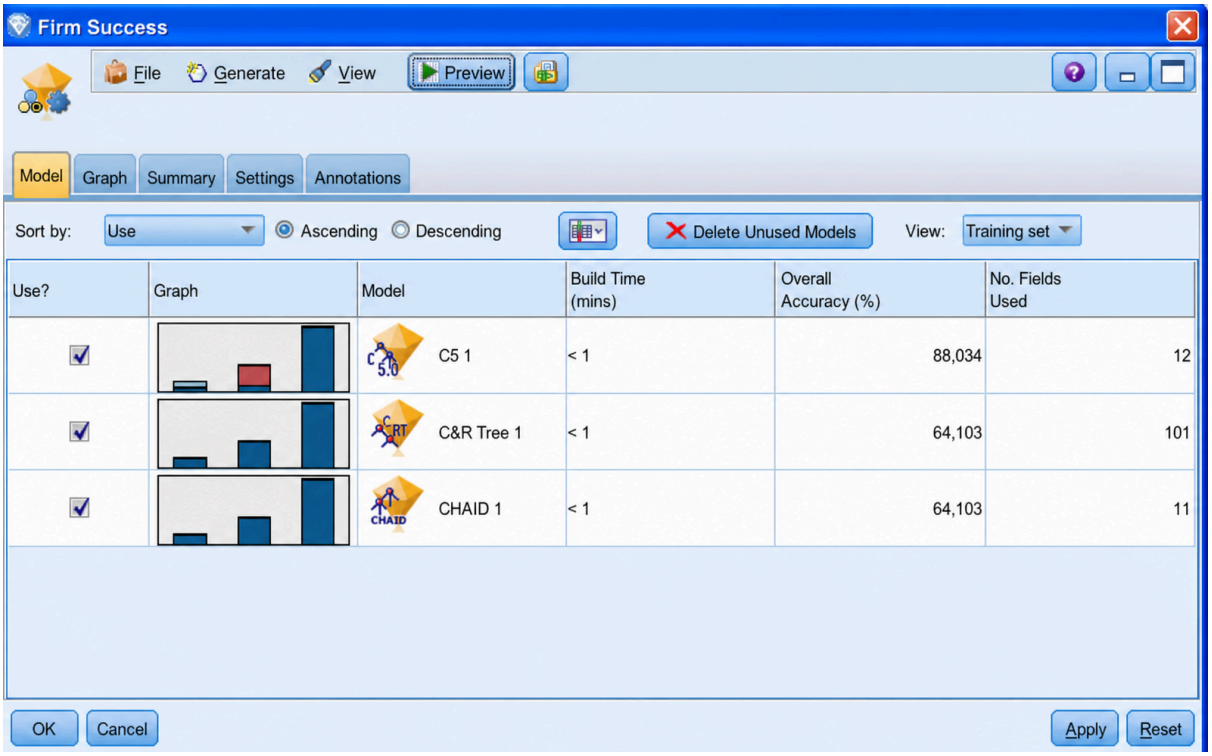


Fig. 28Best received models for predicting success in automatic classification (author created)

The obtained models are compared according to their accuracy (overall accuracy), in which case no cross-validation is applied and the specified accuracy is rather maximal than realistic. The most accurate is the model obtained by the C5.1 algorithm – with an accuracy of 88.03%. The remaining models were obtained by the C&R Tree 1 (classification and regression tree) and CHAID 1 algorithms, both of which had an accuracy of 64.10%. Classification trees start with the most important factors predicting success and divide cases into groups (represented by nodes) depending on the responses. The process continues until reaching an end node (leaf) that shows the predicted value of the goal - the success of the company.

The model obtained by algorithm C5.1 uses 12 variables from the input data to predict the company's success. The model obtained by the C&R Tree algorithm 1 used 101 variables and the one obtained by CHAID 1 used 11 variables. With equal accuracy, models using fewer variables are preferred.

When initially synthesizing models using IMB SPSS Modeler and the automatic classifier, the cross-validation method was not enabled. This method is relevant to model accuracy prediction, and without its application, model accuracy predictions are inflated.

Based on the obtained models, we choose the algorithm C5.1because it obtained the model with the highest accuracy. The algorithm will be used as a starting point to synthesize a high-fidelity model by applying cross-validation. In this way, a model will be created whose accuracy will be evaluated more realistically.

4.2.4.2 Development of a Cross-Validated C5.0 Model

When creating models to predict success, the cross-validation method can be applied. In cross-validation, the data is divided into two sets: a training set (training set), used to create the model and a testing set (testing set), model checker. The advantage of this method is that

the assessment of the accuracy of the resulting model is significantly more realistic. A disadvantage of the method is that when we have a limited amount of data, we have to set aside part of it for model validation. As a consequence, a smaller data set is used to synthesize the model, resulting in a less accurate model.

I apply the algorithm C5.0, which is an implementation of the algorithm C5.1 chosen in the previous point. C5.0 supports multiple manual settings and data cross-validation. After setting the algorithm to use cross-validation and applying it to the data, a model with an accuracy of 83.76% was obtained (Fig. 29).

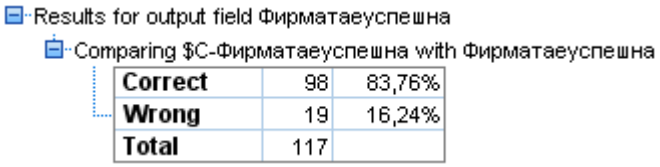


Fig. 29Accuracy assessment of a model obtained by the C5.0 algorithm and applying cross-validation (created by the author)

The distribution of the data correctly and incorrectly classified by the model is shown in Fig. 30. The top bar of the figure shows the failed companies, with the correctly classified ones marked in light blue. The middle bar shows companies that are neither successful nor unsuccessful, with those correctly classified in this group marked in red. The bottom bar shows successful companies, with correctly classified ones shown in dark blue. The bottom bar is of the largest size because the successful companies are the largest number.

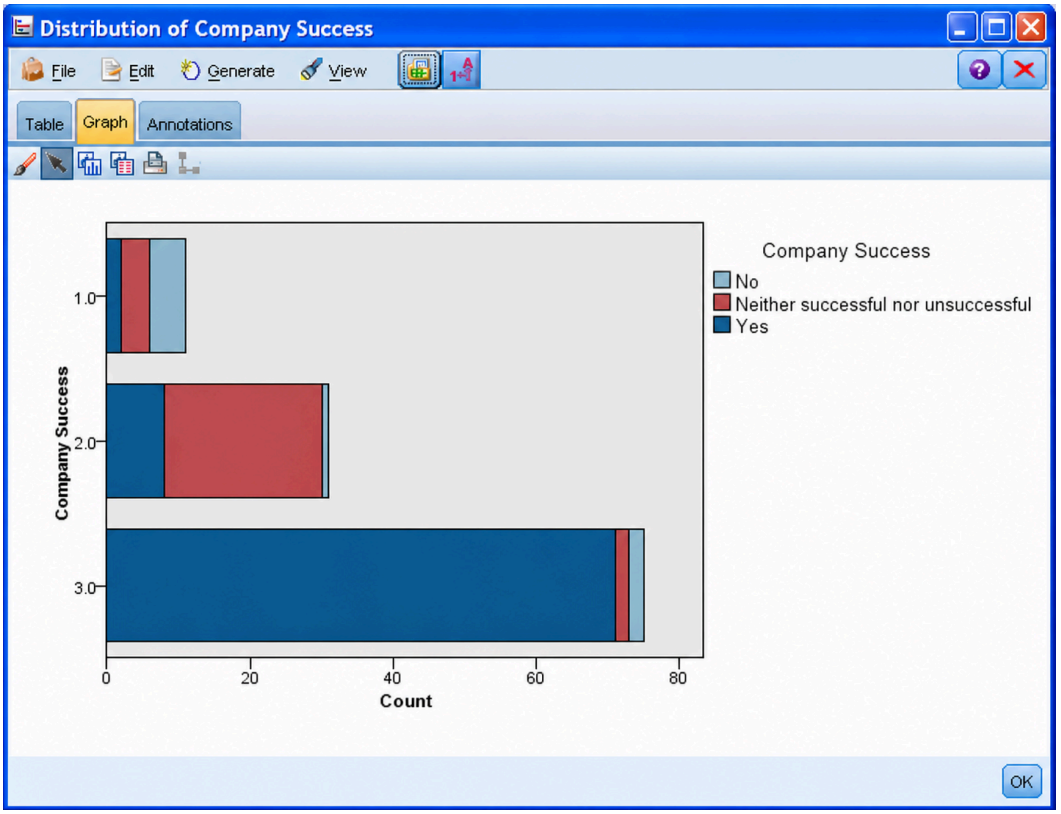


Fig. 30 The distribution of companies correctly and incorrectly classified by the model (created by the author)

The model obtained by cross-validation is a classification tree with 9 variables. To familiarize ourselves with the model in detail, we will look at its text representation, which comprehensively describes the tree. In the **text representation,each** node of the tree is presented on a separate line. Hierarchy in the tree is shown with spaces (tabs) at the beginning of the lines. The more distances there are in front of a node, the more nodes there are above it in the tree. On each line, the abbreviated name of the variable is written first, followed by (in curly brackets) the data values that lead to the current node. After the arrow, the predicted value for the target (start-up success) next from the current node is shown. At the end of the line, in parentheses are shown the number of cases from the data that fall into the node and the probability of success – in absolute units (from 0 to 1). The text representation generated by the data mining software is as follows:

Competitive AdvantageMyBusinessp in [1 2] [Mode: 2] => 2.0 (3; 0.667)

Competitive AdvantageMyBusinessp in [3 4 5] [Mode: 3] (94)

FSuccess_ExternalEnvironment = 1 [Mode: 2] (9)

Experience_Is_Ask_Like_Position in [1 2 3 4] [Mode: 2] => 2.0 (6; 0.833)

Experience_Is_Ask_LikePosition in [5] [Mode: 3] => 3.0 (3; 0.667)

FSuccess_ExternalEnvironment = 2 [Mode: 3] (85)

NMActive_builtbusinessreput in [1] [Mode: 1] => 1.0 (8; 0.625)

NMActive_builtbusinessreput in [2 3 4 5] [Mode: 3] (77)

Mode of market entry = 0 [Mode: 2] => 2.0 (3; 0.667)

Mode of market entry = 1 [Mode: 3] (30)

Profit_Industry in [1 2 3 4] [Mode: 3] (27)

Partner_Others in [1 2 3] [Mode: 2] (8)

Client_InIndustryHasConcentration in [1 2 3 4] [Mode: 2] => 2.0 (5; 1.0)

Client_Inindustryhasconcentration in [5] [Mode: 3] => 3.0 (3; 1.0)

Partner_others in [4 5] [Mode: 3] => 3.0 (19; 0.947)

Industry_MoreProfit in [5] [Mode: 2] => 2.0 (3; 0.667)

Mode of market entry = 2 [Mode: 3] (25)

NMActive_RecognitionBrand in [1] [Mode: 2] => 2.0 (2; 1.0)

NMActive_Recognition in [2 3 4 5] [Mode: 3] => 3.0 (23; 0.957)

Mode of market entry = 3 [Mode: 3] => 3.0 (0)

Mode of market entry = 4 [Mode: 3] => 3.0 (19); 0.947)

The model contains the following 9 variables that describe the successful start-up, with each variable in parentheses followed by the corresponding node number in the tree whose branches depend on the variable:

- Competitive advantage: My business has a clear competitive advantage. (node 0)
- I consider the following factors to be key to success for my business: External environment. (node 2)
- Skills and experience: I have experience in a similar position (to the one currently held). (node 3)
- Intangible assets The company has an established business reputation. (node 6)
- Mode of market entry (node 8)
- Intangible assets: The company has a recognizable brand. (node 17)
- Characteristics of the industry in which the firm operates: Most companies in the industry have a good profit. (node 10)
- Strategic partnerships: The firm engages in partnerships with other stakeholders. (node 11)
- Characteristics of customers: There is a concentration of customers in the industry. (node 12)

A complete graphical representation of the classification tree (tree map) is shown in Fig. 31, with its number written in each of the nodes. The topmost node (0) is primary and branches into two nodes (1 and 2) depending on the value for the company's primary success

factor – the presence of a competitive advantage. Then node 2 branches depending on the next factor and so on until reaching a final node.

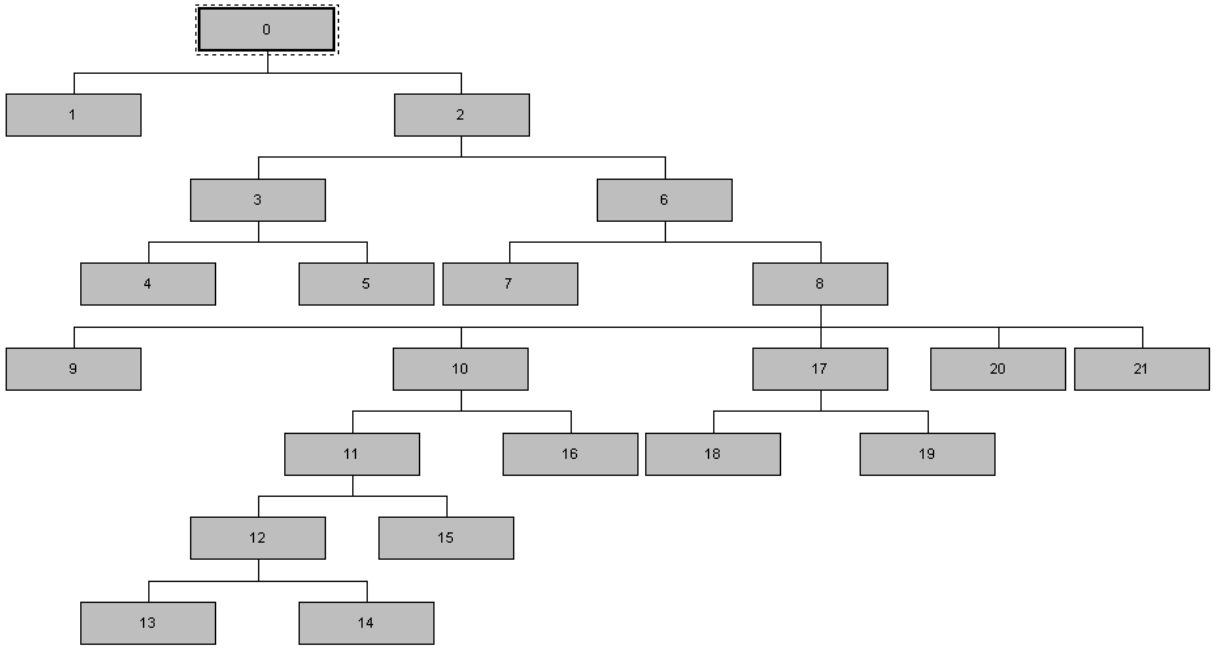


Fig. 31Overall graphical representation (map) of the classification tree (created by the author)

Some of the nodes of the classification tree, which contain the most important success factors, as well as their branches, will be considered. By looking at the structure of the classification tree, we can establish the importance of the success factors, since the most important factors are located at the top nodes of the tree and determine the initial partitioning of the data.

The first level of the tree (node 0) has two branches depending on the variable "presence of a clear competitive advantage" (Fig. 32). This variable is most important for predicting the success of the companies in the analyzed data set. start-ups that have a clear competitive **advantage(node** 2) are more successful than those that do not (node 1). The dissertation, which is also found in other studies (Ruskov, 2012) that the presence of a competitive advantage is a key factor for the success of start-up companies, is confirmed.

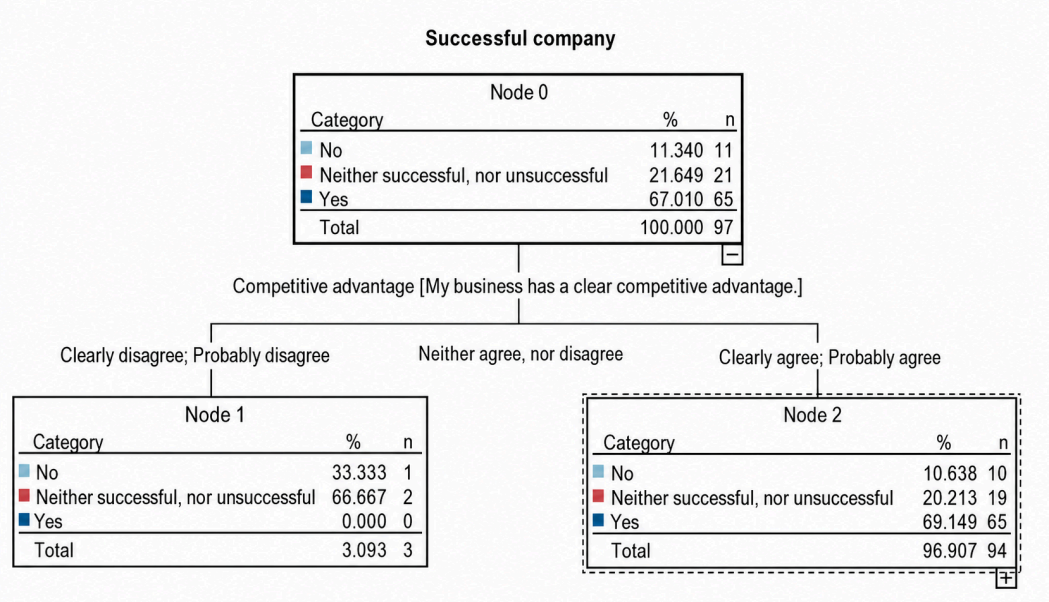


Fig. 32Classification tree - first level: Existence of a clear competitive advantage (created by the author)

Node 2 has two branches depending on the next success factor of the company "**the environment is a key success factor**" (Fig. 33). Those companies in the analyzed data set that consider the environment as a key success factor (node 3) are less successful than the others (node 6). Node 3 has a small number of dataset elements, and for this reason I will not analyze its descendants in depth.

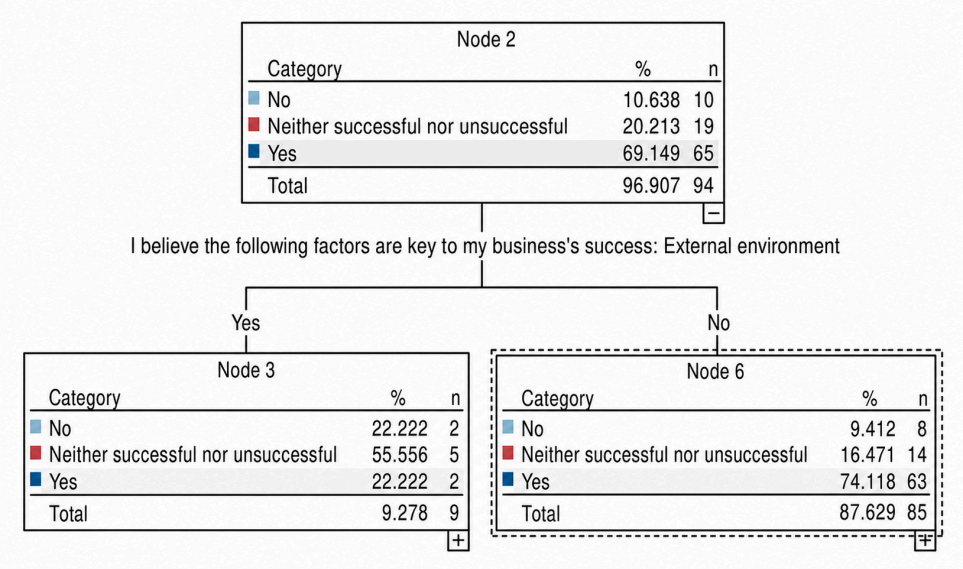


Fig. 33Classification tree - branch 2: the external environment is a key factor for the company's success (created by the author)

Node 6 has two branches depending on the next factor for the company's success "**presence of an established business reputation**" (Fig. 34). Those companies from the analyzed data

set that do not have an established business reputation (node 7) are less successful than the rest (node 8).

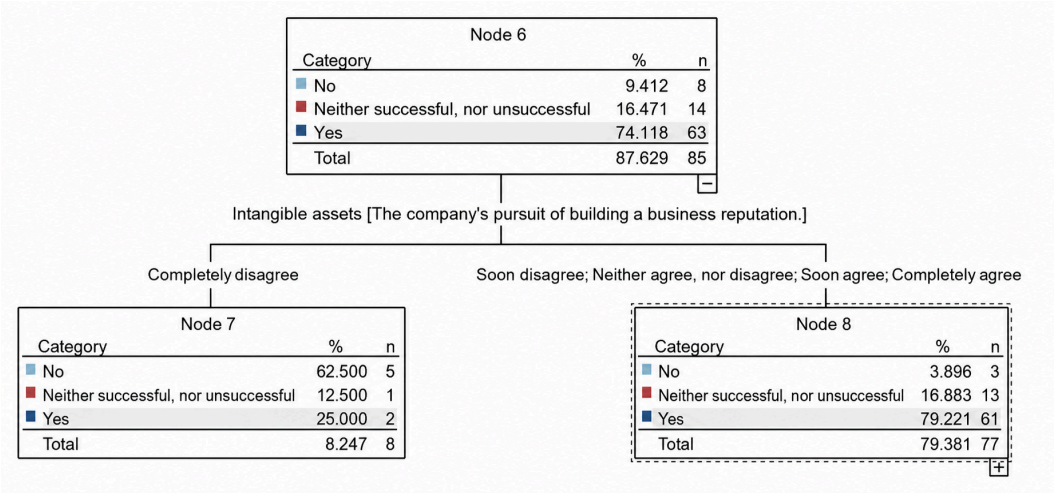


Fig. 34Classification tree - branch 2.2: intangible assets - established business reputation of the company (created by the author)

Node 8 has multiple branches depending on the next success factor of the company "way **of entering the market of the start-up company"** (Fig. 35). Changing an existing product or service is the most successful way of entering the market for the companies in the analyzed data set. Companies that develop a new product or service, as well as those that enter the market through parallel competition, are also very often successful.

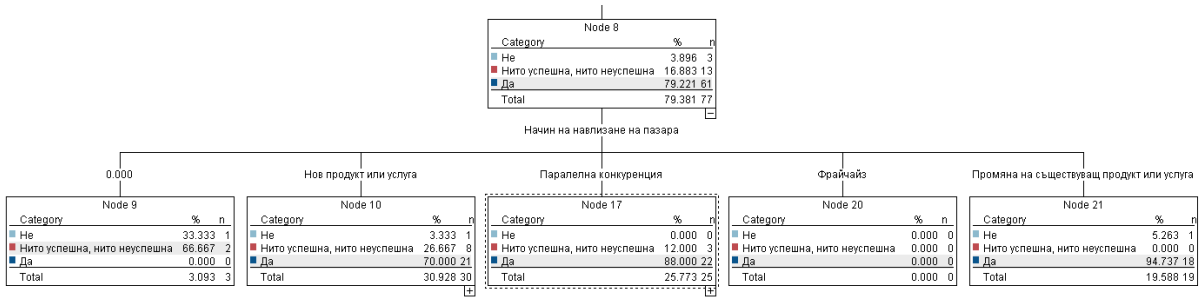


Fig. 35Classification tree - branch 2.2.2: how the start-up company enters the market (created by the author)

Further analysis of the tree structure shows the location by nodes, the interrelationships and the relative importance of the remaining factors for the success of the company. The complete output of the analysis is presented in Appendix 2.

4.2.4.3 Comparison of algorithms for obtaining models

The model generation algorithms proposed by IBM SPSS Modeler (C5.1, C&R Tree and CHAID) use the rule induction technique. Software algorithms synthesize a classification tree, representing a set of rules that describe certain segments in the data in relation to the goal,

which in our case is the success of the company. They traverse the input data set and sequentially divide it into subsets based on relationships between the data and the target.

The models obtained by the algorithms clearly represent the logic for each rule and can therefore be used to understand the decision-making process that leads to the corresponding result (company success). Another advantage of rule induction, compared to other methods such as neural networks, is that the process automatically eliminates fields that are not important for decision making.

To explain **how the rule induction algorithm**works, I will use the decision to buy a house as an example. The most important factor may be the price, as it reflects the ability to afford the home. The next factor could be the type of property we are looking for - a house or an apartment. The next factor could be location and so on. By presenting the factors in the form of questions and by hierarchically combining them, obtaining a classification tree (Fig. 36).

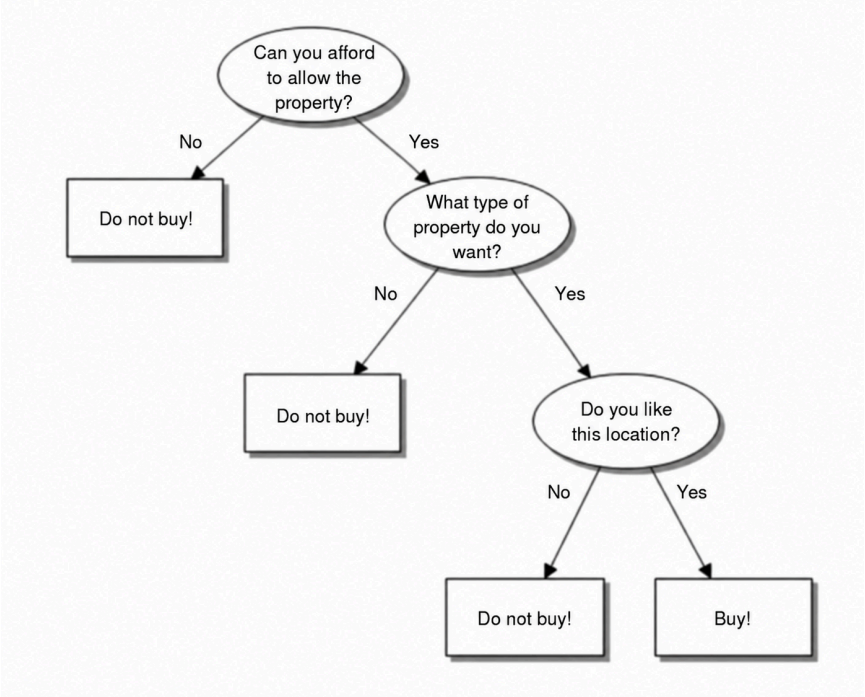


Fig. 36Graphic representation of a classification tree (translated by the author)

The C5.1, C&R Tree, and CHAID algorithms differ in the way they generate a result (classification tree).

C5.1 targets only **category-type** fields (flag, nominal, and ordinal), while C&R Tree and CHAID support category-type and continuous targets. In our case, the target is of type category. When algorithms recursively traverse the data, they create different types of data splits (splits) into subgroups (nodes) based on data variables (predictors). C&R Tree supports only binary splits (into two subgroups), while CHAID and C5.1 support splits into more than two subgroups.

Algorithms differ in the criterion **used to split the data.** In C5.1 (respectively also C5.0) a meter from the information theory is used - the coefficient of information extraction (information gain ratio). When the C&R Tree algorithm predicts a field of category type, a

measure of variance (the Gini coefficient by default) is used. And the CHAID algorithm uses a chi-square criterion.

The considered algorithms **allow missing values in** the data fields, although they apply different methods for their interpretation. C5.1 uses an allocation method that allocates a portion of the record with missing data to each of the analyzed branches of the tree. C&R Tree uses field substitution to proceed with tree generation. CHAID creates a separate category for the missing values and uses them when building the tree.

There are a small number of fields with missing values in the analyzed survey data. It is not recommended to completely remove records that have a small number of missing fields. In addition, omitting the answer to a particular question may be relevant. For this reason, missing values are marked as such in the files and their interpretation is entrusted to the relevant software algorithms.

The trees that different algorithms generate are not necessarily identical because they apply different **criteria to select the variables** that predict the target value. The CHAID and QUEST algorithms use more standard statistical methods, while C5.0 and C&R Tree use non-statistical metrics.

The C5.0 algorithm has yielded the best model for predicting the success of start-ups, and for this reason, C5.0 will be **discussed in detail. The review** also concerns the version of the algorithm when applying automatic classification, which is designated as C5.1. It should be noted that the C5.0 algorithm is closed source and its exact behavior on the Windows version implemented in IBM SPSS Modeler is not publicly available. However, its basic operating principles are formally documented (IBM, 2010). The C5.0 algorithm was developed by the company RuleQuest (RuleQuest, 2014), which offers its version for the Linux operating system under the Gnu GPL open license. The main part of the code for the **Linux version** of the algorithm is included in Appendix 2, and the complete code is available for download on the manufacturer's website. C5.0 is known to be based on the open source classification algorithm C4.5 (Learning, 1993), the definition of which is fully available.

Algorithm C5.0 works on the principle of selection based on **gain ratio (Gain Ratio Selection)**, which decides how to divide the data, and which is based on information theory. Such a division of the data in the branches of the classification tree is chosen so that the extraction of information is maximized. The information extracted by dividing the data based on the field categories X (target predictor variable) is measured by Formula 6.

$$GAIN(X) = INFO(DATA) - INFO_X(DATA) \qquad (6)$$

GAIN(X) denotes the extracted information. INFO(DATA) represents the averaged information needed to identify the class (outcome category) of an entry in the data set. And INFOX(DATA) represents the expected required information after once the data has been divided according to each result for the current field.

According to the information theory, **the information extraction criterion is defined**as follows: “The information conveyed by a message depends on its probability and can be measured in bits as the minus logarithm to the base 2 of that probability.' If, for example, there are 8 equally likely messages, the information conveyed by any of them is – log2 (1/8), or 3 bits.

Although the information extraction criterion gives good results, it has the disadvantage of favoring splits that have a large number of cases. Thus, a categorical input variable that has many values takes precedence over another with fewer values.

The C5.0 algorithm solves this problem by using the **information extraction ratio criterion.** Expectations in the information extraction criterion can be adjusted by a type of normalization in which the extracted information is adjusted for multivalued input data. For this reason, C5.0 uses an information extraction ratio, which is a proportion of the information generated by dividing the data in the parent node into each of the field categories X, which is useful in classification (Formula 7).

$$GAIN\ RATIO(X) = GAIN(X) / SPLIT\ INFO_X(DATA) \quad (7)$$

Where GAIN RATIO(X) is the gain ratio. GAIN(X) indicates the extracted information. SPLIT INFOX(DATA) $_{\text{represents}}$ the potential information generated by splitting the data into n cases, given that information extraction measures information relevant to classification.

The C5.0 algorithm chooses to split the data based on the fields where the information extraction ratio is maximum. This rule is limited by an additional rule: the extracted information must be a lot, or at least as much as the average extracted information from all cases considered. In this way, instability of **the algorithm criteria is** avoided in cases where the extracted information is small.

When **evaluating a record by the C5.0 algorithm** (classification and confidence calculation), the input record is evaluated by the class and confidence of the rule that generated it.

If a set of rules **is directly generated from a tree node**, then the confidence for the rule is calculated as follows (Formula 8):

$$< rule\ confidence > \ = \ \frac{<correct\ terminal{-}node\ records> + 1}{<terminal{-}node\ records> + 2} \quad (8)$$

If a set of rules is**generated from a classification tree generated by the node**, then the confidence for the rule is calculated as follows (Formula 9):

$$< rule\ confidence > \ = \ \frac{<correct\ terminal{-}node\ records> + 1}{<terminal{-}node\ records> + <target\ classes>} \quad (9)$$

When doing a **rule-set evaluation**, for each record all the rules that apply to it are considered. For each rule, prediction (classification) and confidence values are generated and associated. For each value of the prediction results, a confidence sum is calculated, and the one with the highest value of this sum is selected as the final prediction result. The confidence of the final prediction result is calculated as the sum of the confidence of this value divided by the number of applied rules for the corresponding record.

4.2.4.4 Comparison of the obtained success prediction models

Before comparing the obtained models, it is necessary to note that the automatic classification tool in IBM SPSS Modelerwas implementedwithout using cross-validation of the data. This means that the same data is used to synthesize the model and to check its accuracy. As a result, the stated accuracy estimates are inflated. The models obtained by the automatic classifier were used to identify the most suitable algorithm - C5.1. After that, a model was generated by applying cross-validation through the algorithm C5.0using manual settings, whose accuracy was evaluated.

If we consider only the factor of highest predictive accuracy, the most accurate would be the model, generated by C5.1. But bearing in mind that cross-validation should be applied in order for the accuracy to be calculated more objectively, the best model is similar to the first one, but generated with the algorithm C5.0 and applying cross-validation.

4.2.5 Data Mining and Success-Model Development with Weka

This section discusses the process of extracting knowledge from data and creating models to predict the success of start-ups with the software product Weka (Hall, 2009). The created models are analyzed and compared in terms of their working principles and efficiency. The software algorithms by which the most effective models were obtained were examined.

4.2.5.1 Synthesizing success prediction models

With the help of the software product Weka success prediction models were created, which perform a classification of companies - predicting which category they fall into: "successful", "neither successful nor unsuccessful" or "failed".

Weka offers multiple algorithms for creating classification models (Fig. 37) that can be selected and run from the graphical user interface.

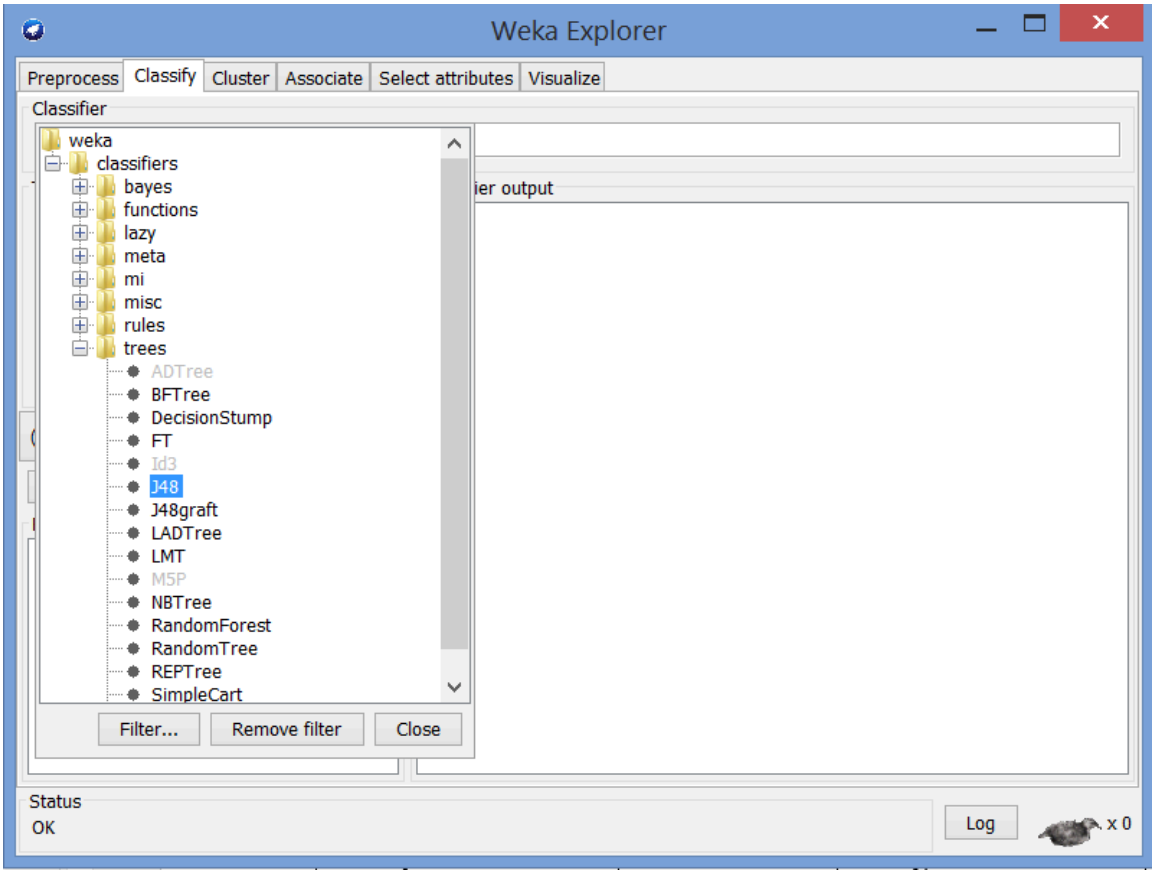


Fig. 37Algorithms for classification in Weka (created by the author)

After choosing an algorithm, a **objective for the model** is selected - the variable whose value is predicted, in this case - the success of the company. In addition, the model generation algorithm can be set - whether to apply **cross-validation** of the model and other settings specific to the particular algorithm.

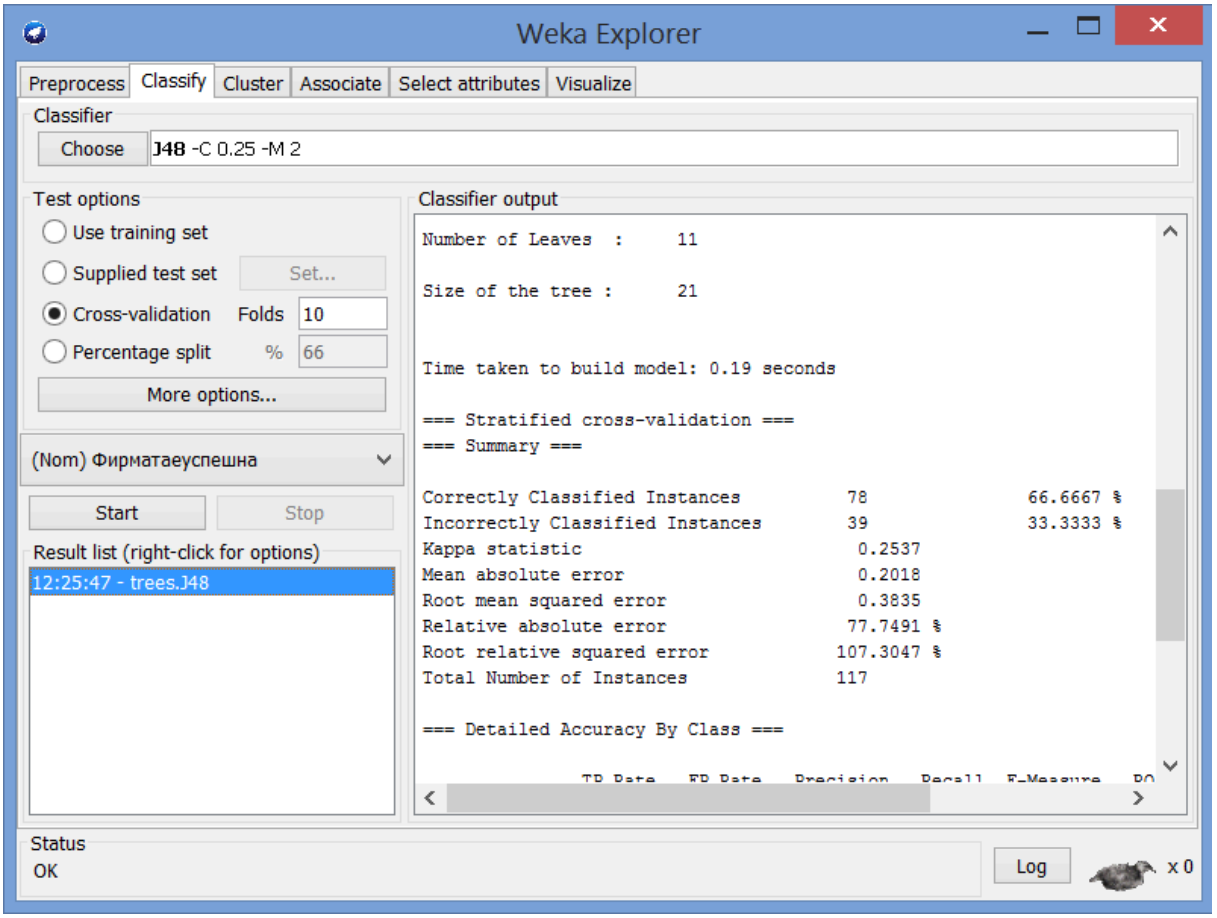


Fig. 38Setting up the J48 algorithm in Weka and creating a classification model (created by the author)

After setting the algorithm and pressing the "Start" button, a text history of the data obtained by the classification algorithm is obtained in the "Classifier output" field (Fig. 38). This timeline contains information about the model and its synthesis. Part of the obtained data also contains the prediction of model accuracy. These steps are repeated until the desired patterns are generated, which are then analyzed and compared.

Different classification algorithms were tested, all performed with and without applying cross-validation. The results are shown in Table 13where the models are ranked in descending order according to accuracy when applying cross-validation.

Table 13Comparison of classification models obtained with Weka

Classification algorithm	Type of model	Accuracy (without cross validation)	Accuracy (with cross-validation)
J48	Tree	83.76%	66.67%
J48graft	Tree	83.76%	66.67%
DecisionTable	Rules	68.38%	64.10%
LMT	Tree	66.67%	64.10%

BayesNet	Bayes	65.81%	64.10%
BFTree	Tree	64.10%	64.10%
REPTree	Tree	69.23%	63.25%
SimpleCart	Tree	64.10%	63.25%
FT	Tree	100.00%	60.68%
DecisionStump	Tree	64.95%	60.68%
RandomForest	Tree	98.29%	59.83%
NBTree	Tree	91.45%	54.70%
NaiveBayes	Bayes	80.34%	52.14%
LADTree	Tree	83.76%	48.72%
RandomTree	Tree	99.15%	47.86%

The table presents models that have a heterogeneous principle of operation (the terminology used in Weka is preserved): classification tree (Tree), rule induction (Rules), Bayesian classifier (Bayes).

The full analysis output for the model J48 is presented in Appendix 2.

4.2.5.2 Comparison of Classification Algorithms

Weka's proposed algorithms that generate the highest accuracy models using cross-validation are J48 and J48graft. The models generated are **classification trees**, representing a set of rules that describe certain segments in the data in relation to the goal, which in our case is the success of the company.

J48 is an open source implementation of the C4.5 algorithm implemented in the Java programming language. The algorithm is implemented as a class "public class J48 extends AbstractClassifier implements OptionHandler, Drawable, Matchable, Sourcable, WeightedInstancesHandler, Summarizable, AdditionalMeasureProducer, TechnicalInformationHandler, PartitionGenerator".

Due to its similarity to C4.5, J48 resembles the C5.0 algorithm from IBM SPSS Modeler. J48 builds a classification tree from a dataset by applying the concept of **information entropy**. The training data is a set of already classified cases (samples) S = s1, s2, ... Each sample siconsists of a p-dimensional vector (x1,i, x2,i,..., xp,i), where xj represents the attributes of the sample as well as the class in which si falls.

At each node of the tree, the algorithm selects the attribute from the data that most efficiently divides the set of samples into subsamples. The separation criterion is the **normalized extracted information**(entropy difference), with the attribute with the highest normalized extracted information being selected for decision making (and data splitting). The algorithm is then repeated recursively on the resulting subsamples.

The algorithm J48 (and C4.5) has the following basic principles:

- If all samples from datasets belong to the same class, the algorithm simply creates a single terminal node (leaf) for the classification tree.

- If none of the attributes provide extracted information, the algorithm creates a decision node high in the tree using the expected class value.
- If it encounters an instance of a class not encountered before, the algorithm creates a decision node high in the tree using the expected class value.

4.2.5.3 Comparison of the resulting success prediction models

Of the obtained classification models, the model obtained with the J48algorithm has the highest accuracy when applying cross-validation: 66.67%. This accuracy is obtained by using different data for training and testing the model, and is reasonably realistic. Model accuracy is generally relatively low, as well as lower than the accuracy achieved using the software product IBM SPSS Modeler.

I will examine in greater depth the model obtained by J48applying cross-validation in order to identify the factors that led to the company's success and their relative importance. The chronology generated by the classification algorithm shows the structure of the classification tree in text form:

Competitive AdvantageMyBusinessp <= 2: 2.0 (6.0/1.0)

Competitive AdvantageMyBusinessp > 2

| FSuccess_Cost <= 1

| | Change_ManagementExperience <= 3: 1.0 (2.0)

| | Change_Management_Experience > 3: 3.0 (2.0)

| FSuccess_Cost > 1

| | FSuccess_ExternalEnvironment <= 1

| | | Experience_I have an experience similar to the position <= 4: 2.0 (5.0)

| | | Experience_I have an experience similar to the position > 4: 3.0 (3.0/1.0)

| | FSUCCESS_ExternalEnvironment > 1

| | | NMactive_builtbusinessreput <= 1

| | | | The product we offer accepted <= 2: 3.0 (3.03/1.03)

| | | | The product we offer accepted > 2: 1.0 (4.04/0.04)

| | | NMactive_builtbusinessreput > 1

| | | | NMAactive_recognizebrand <= 2

| | | | | Working with one basic client Basic <= 1: 3.0 (2.0)

| | | | | Working with one basic client Basic > 1

| | | | | | Personal_Don't bother me I'll take it <= 3: 2.0 (9.15/1.0)

| | | | | | Personal_Don't bother I'll take it > 3: 3.0 (4.0/1.0)

| | | | NMActive_recognitionbrand > 2: 3.0 (76.78/13.78)

In the text representation of the tree, the factors are described with abbreviated names. The most important factor for the company's success is the presence of a clear competitive

advantage. This result is also consistent with the results obtained by the algorithm C5.0 in IBM SPSS Modeler. The remaining factors in the model generated by Weka and J48 are:

- Competitive advantage: My business has a clear competitive advantage.
- I consider the following factors to be key to the success of my business: Price
- I believe the following factors are key to success for my business: External environment
- Skills and experience: I have management skills.
- Intangible assets: The company has an established business reputation.
- Skills and experience: I have experience in a similar position (relative to the one currently held).
- Intangible assets: The company has a recognizable brand.
- The product I offer has been accepted by the market to the following extent.
- The main source used for initial funding of the company is: Individual investor (business angel)
- Personality and values: I am not worried about making decisions without having the necessary information.

Part of the listed success factors of the company are similar to those obtained with the help of IBM SPSS Modeler and the C5.0 algorithm: competitive advantage, the external environment as a key factor for success, established business reputation, recognizable brand. The remaining factors are specific to the model.

Weka does not offer a detailed visualization of the classification tree, unlike IBM SPSS Modeler. Only a schematic visualization (Fig. 39) is possible, which shows the nodes of the tree, the connections between them and abbreviated names of the variables.

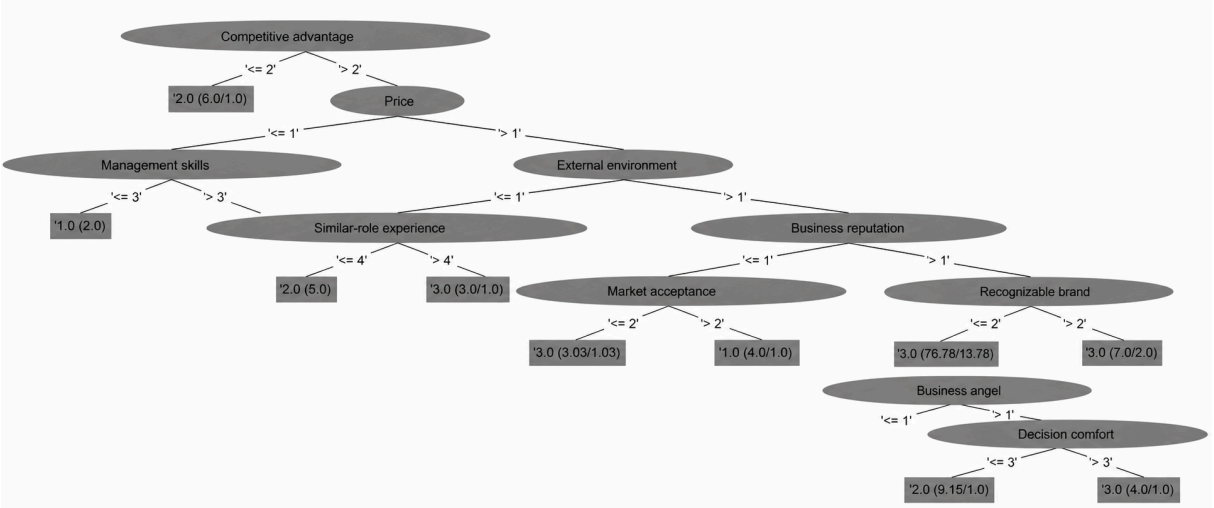


Fig. 39Schematic visualization of the classification tree generated with the J48 algorithm in Weka (created by the author)

4.2.6 Development of Success-Prediction Models for Technology Companies

Based on the definition that technology **companies** develop a new or change an existing product (or service) proposed in Chapter 2, 72 technological start-up companies from Bulgaria were selected from the data set.

The selected data for technology start-ups was analyzed with IBM SPSS Modeler data mining software using the CRISP-DM methodology. The experimental set-up is the same as for data mining for all start-ups, only the input data set is different. The tools applied are again an automatic classifier and a C5.0 classification algorithm with cross-validation applied. For this reason, the experimental setting and the principle of operation of the algorithms will not be repeated.

The automatic classifier ranks the resulting models and keeps the best three success prediction models (those with the highest accuracy) for further analysis (Fig. 40).

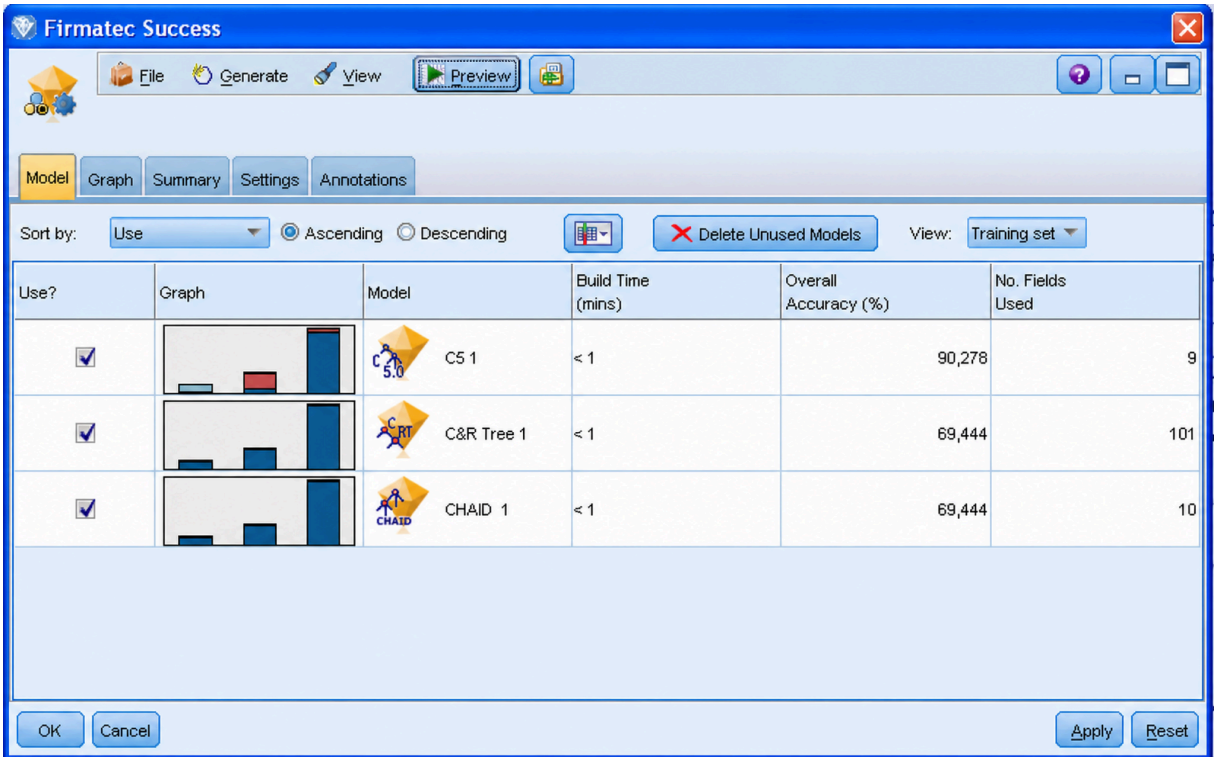


Fig. 40Best received models for predicting success of technology companies in automatic classification (created by the author)

The obtained models are compared according to their accuracy (overall accuracy), in which case no cross-validation is applied and the specified accuracy is rather maximal than realistic. The most accurate is the model obtained by the C5.1 algorithm – with an accuracy of 90.28%. The remaining models were obtained by the C&R Tree 1 (classification and regression tree) and CHAID 1 algorithms, both of which had an accuracy of 69.44%.

To obtain a model with more realistic predictive accuracy, I apply the algorithm C5.0. The C5.0 algorithm was chosen because it is an implementation of the C5.1 algorithm, which obtained the best model using the automatic classifier. C5.0 supports multiple manual settings and data cross-validation. After tuning the cross-validation algorithm and applying it to the data, a model with an accuracy of 87.50% is obtained (Fig. 41).

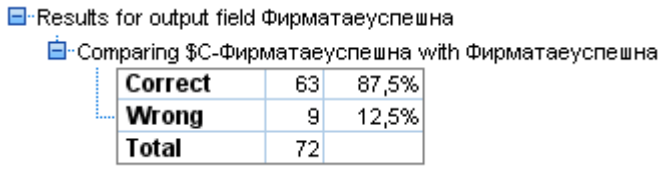


Fig. 41Evaluation of the accuracy of a model for predicting the success of technology companies obtained by the C5.0 algorithm and applying cross-validation (created by the author)

The model obtained by cross-validation is a classification tree with 8 variables. To get acquainted in details the model, we will look at its textual representation, which comprehensively describes the tree:

Industry_MoreProfit in [1 2 3 4] [Mode: 3] (55)

FSuccess_Price = 1 [Mode: 1] (4)

Experience_Investment Management in [1 2 3] [Mode: 1] => 1.0 (2; 1.0)

Experience_Imamumeniyapomanagement in [4 5] [Mode: 3] => 3.0 (2; 1.0)

FU_Price_Success = 2 [Mode: 3] (51)

FU_Marketing_Success = 1 [Mode: 2] => 2.0 (4; 0.75)

FMarketing_Success = 2 [Mode: 3] (47)

Strategy_Strategythatwerecollected in [1 2 3] [Mode: 3] (32)

Competition_The industry is strong in [1 2] [Mode: 2] => 2.0 (2; 1.0)

Competition_IntheIndustryisstrongerthanku in [3 4 5] [Mode: 3] (30)

NMActive_builtbusinessreput in [1 2 3] [Mode: 2] (9)

Strategic partnerships The company in [1] [Mode: 3] => 3.0 (3; 1.0)

Strategic partnershipsThe company in [2 3 4 5] [Mode: 2] => 2.0 (6; 1.0)

NMActive_built business reputation in [4 5] [Mode: 3] => 3.0 (21; 0.905)

Strat_Strategythatwerecollected in [4 5] [Mode: 3] => 3.0 (15; 1.0)

Industry_MoreProfit in [5] [Mode: 1] => 1.0 (5; 0.6)

The model contains the following 8 variables that describe the successful technology start-up, with each variable in parentheses followed by the corresponding node number in the tree whose branches depend on the variable:

- Most companies in the industry have a good profit. (node 0)

- I consider the following factors to be key to success for my business: Price (node 1)
- I consider the following factors to be key to success for my business: Marketing and Advertising (node 5)
- Skills and experience: I have management skills. (node 2)
- The strategy I have chosen is aggressive. (node 7)
- There is strong competition in the industry. (node 8)
- Intangible assets: The company has an established business reputation. (node 10)
- Strategic partnerships: The company participates in partnerships with competitors. (node 11)

From the listed success factors of technology start-ups, it can be seen that they are very different from the success factors of start-up companies from Bulgaria such as whole. Many of the factors are related to the external environment. A **complete graphical representation** of the success prediction classification tree (tree map) is shownin Fig. 42.

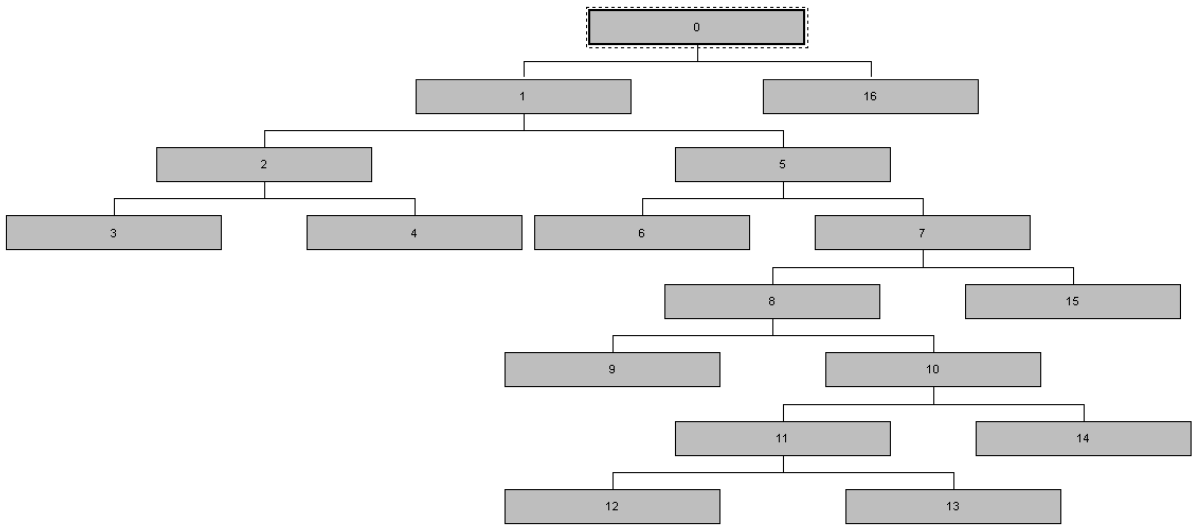


Fig. 42Overall graphical representation (map) of the classification tree for technology companies (created by the author)

Some of the nodes of the classification tree that contain the most important success factors will be examined, as well as their branches. By looking at the structure of the classification tree, we can establish the importance of the success factors, since the most important factors are located at the top nodes of the tree and determine the initial partitioning of the data.

The first level of the tree (node 0) has two branches depending on the variable "**most companies in the industry have a good profit**" (Fig. 43). This variable is most important for predicting the success of the companies in the analyzed data set. Start-ups who strongly agree that most companies in the industry have a good profit (node 16) are less successful than others (node 1). At first glance, this conclusion seems illogical, and most likely there is a hidden reason, which could be the connection of the profits of companies in an industry with the competitive environment and the presence of large firms in the market.

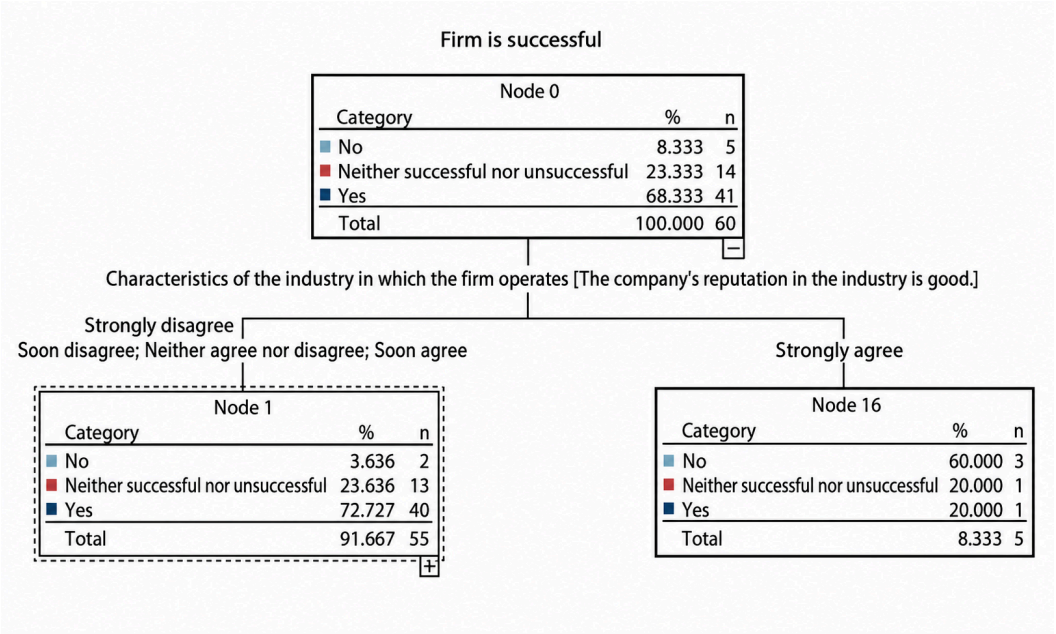


Fig. 43Classification tree - first level: Most companies in the industry have a good profit (created by the author)

Node 1 has two branches depending on the next success factor of the company "**price as a key success factor**" (Fig. 44). Those companies in the analyzed data set that consider price as a key success factor (node 2) are less successful than the rest (node 5). Node 2 has a small number of dataset elements, and for this reason I will not analyze its descendants in depth.

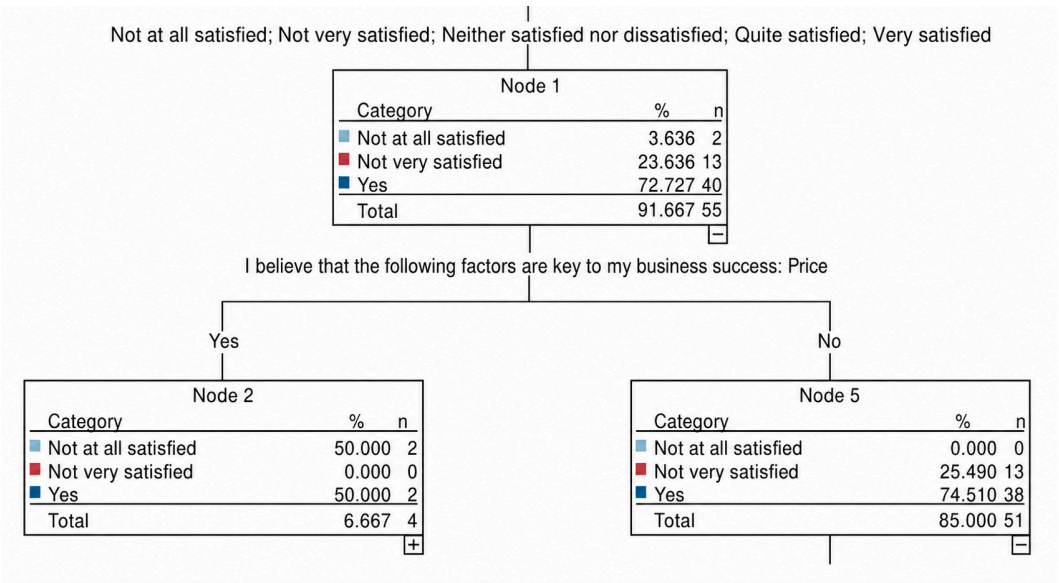


Fig. 44Classification tree - branch 1: key success factors: price (created by the author)

Node 5 has two branches depending on the next success factor of the company "**marketing and advertising as key success factors**" (Fig. 45). Those companies in the analyzed data set that consider marketing and advertising as a key success factor (node 6) are less successful than the rest (node 7).

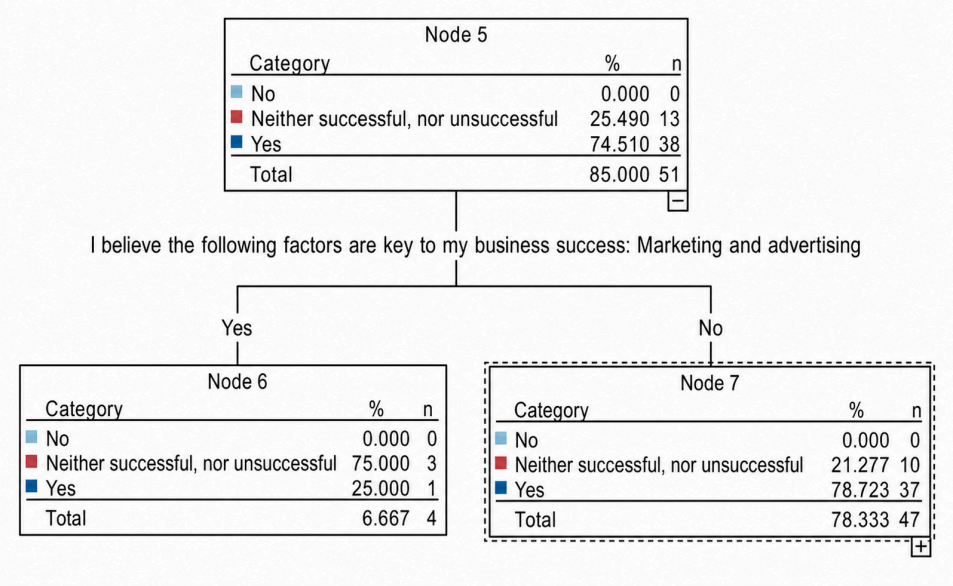


Fig. 45Classification tree - branch 1.2: key success factors: marketing and advertising (created by the author)

Node 7 has two branches depending on the next success factor of the company "the **chosen strategy is aggressive"** (Fig. 46). Those companies in the analyzed data set that chose an aggressive strategy (node 15) are more successful than the rest (node 8).

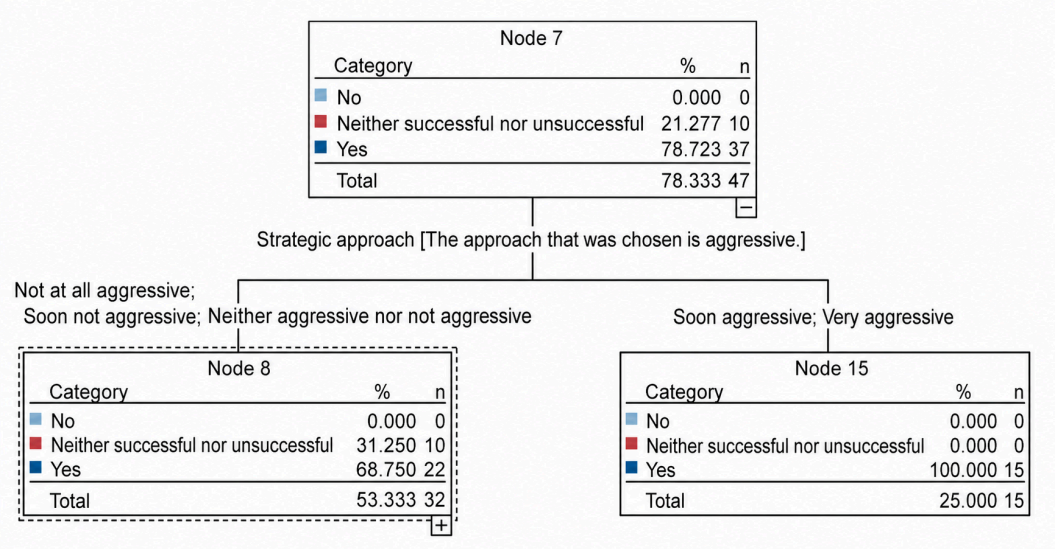


Fig. 46Classification Tree - Branch 1.2.2: The chosen strategy is aggressive. (created by the author)

Further analysis of the tree structure shows the node placement, interrelationships, and relative importance of the remaining company success factors. The complete output data of the analysis is presented in Appendix 2.

4.2.7 Comparison of the Resulting Models

Based on the analysis, the two **most accurate models for predicting the success of start-up companies (when** applying the cross-validation method) are:

- classification tree generated using the product IBM SPSS Modeler and the algorithm C5.0 with accuracy 83.76%,
- classification tree generated using the product Weka and J48 with accuracy 66.76%.

A model was also synthesized that predicts the success of technology **start-ups(when** applying cross-validation.):

- classification tree generated using the product IBM SPSS Modeler and the algorithm C5.0 with accuracy 87.50%.

A comparison of the success factors of the three listed models is shown in Table 14.

Table 14A comparison of models for predicting the success of start-up companies

Classification tree for start-up companies obtained by C5.0	**Classification tree for start-ups obtained through J48**	**Classification tree for technology start-ups obtained through C5.0**
• Competitive advantage: My business has a clear competitive advantage. • I believe the following factors are key to success for my business: External environment • Skills and experience: I have experience in a similar position (relative to the one currently held). • Intangible assets: The company has an established business reputation. • How to enter market • Intangible assets: The company has a recognizable brand.	• Competitive advantage: My business has a clear competitive advantage. • I consider the following factors to be key to the success of my business: Price • I believe the following factors are key to success for my business: External environment • Skills and experience: I have management skills. • Intangible assets: The company has an established business reputation. • Skills and experience: I have experience in a similar position (relative to the one currently held).	• Most companies in the industry have a good profit. • I consider the following factors to be key to the success of my business: Price • I believe the following factors are key to success for my business: Marketing and Advertising • Skills and experience: I have management skills. • The strategy I have chosen is aggressive. • The industry is highly competitive.

• Characteristics of the industry in which the company operates: Most companies in the industry have a good profit. • Strategic partnerships: The company engages in partnerships with other stakeholders. • Characteristics of customers: There is concentration of customers in the industry.	• Intangible assets: The company has a recognizable brand. • The product I offer has been accepted by the market to the following extent • The main source used for initial funding of the company is: Individual investor (business angel) • Personality and values: I am not worried about making decisions without having the necessary information.	• Intangible assets: The company has an established business reputation. • Strategic partnerships: The company engages in partnerships with competitors.

The table shows that the two best models for predicting start-up success are similar in the factors they include. The model for predicting the success of technology start-ups is specific, and although it has some common factors with the other models, most differ.

4.3 Results, Findings, and Conclusions

This chapter describes the conducted qualitative research with **the aim of** preliminary verification and improvement of the proposed in Chapter 3 abstract and detailed model for predicting the success of start-up companies from Bulgaria. As a result of the collected data from the qualitative research and the improvements and additions made to the models, they meet **the requirements** and preferences of the Bulgarian **entrepreneurs** who took part in the research. This helps to achieve completeness **and comprehensiveness** of the proposed models.

The grouping of factors from the abstract model for predicting success proposed in Chapter 3 has been validated **by** the statistical method **factor analysis, which** showed a lot grouping close to the author's hypothesis. From the results of the factor analysis, it can be concluded that the proposed abstract model **correctly groups the proposed success factors.**

By applying the CRISP-DM methodology and products for machine learning and knowledge extraction from data, models were synthesized **to predict the success** of start-ups from Bulgaria as well as technology start-ups. The resulting models were analyzed in terms of the algorithms used to generate them, the factors contained, as well as the accuracy of the models. The best success prediction models, which are of the classification tree type, were compared and selected, and which can serve to build an information system for predicting the success of start-up companies. The best models predict **the success of** start-ups and technology start-ups with **high accuracy,** from which it can be concluded that they **successfully meet their goals.**

CHAPTER 5: IMPLEMENTATION OF SUCCESS-PREDICTION SOFTWARE

This chapter will look at the implementation of software to predict the success of start-up companies. The software will be referred to for short by its name I3SP(Information System for Start-ups Success Prediction), or in translation "information system for predicting the success of start-ups".

5.1 Requirements for Start-up Success-Prediction Software

The software for predicting the success of start-ups the success of technology start-ups must meet certain functional requirements to be both useful to users and in sync with research goals and outcomes.

Minimum application requirements are as follows:

- Compatibility with success prediction models (from data mining software products) and ability to update the model. At a minimum, classification tree-type models should be supported.
- Compatibility with Google Disk. Survey data that users complete, is stored in Google Drive and is used both to predict success and to update prediction models in the future. Support for GoogleDrive interfaces will allow the application to directly download the actual data from completed surveys to predict the chances of success of companies.
- Multi-platform – the application must be compatible with different devices and different operating systems. The software must be compatible at a minimum with the following browsers: GoogleChrome (version 30 and higher), MozillaFirefox (version 27 and higher), Opera (version 19 and higher) and Microsoft Internet Explorer (version 9 and higher), which will ensure compatibility with future browsers and versions.
- Low implementation and maintenance costs - this will make the application accessible to the target group - start-ups, most often small companies that have limited budgets for their activity.
- The software should be easy to work - no technical competences are required to make a forecast of the companies' success.
- The software should make a forecast for the success of the companies, which is rationally and comprehensibly explained to the user.
- The software would do well to make additional recommendations related to the result of the analysis – for example, factors that can be improved to make the start-up more successful.
- The software should be easy to develop and update. For this purpose, it should work simultaneously with the surveys from Google Drive and the models from the data mining software. When enough data is accumulated in Google Drive, a new classification model will be prepared in text file format and it should be easily uploaded to the success prediction software.

Based on the functional requirements for the software, appropriate technologies and tools are selected for its implementation, and its architecture is designed.

5.2 Design of the Start-up Success-Prediction Software

This point will consider the design of the software product - its architecture, work algorithm and modules.

The Information System for start-up Success Prediction (I3SP) provides a prediction of company success based on input from a survey completed by the company owner and a success prediction model (Fig. 47).

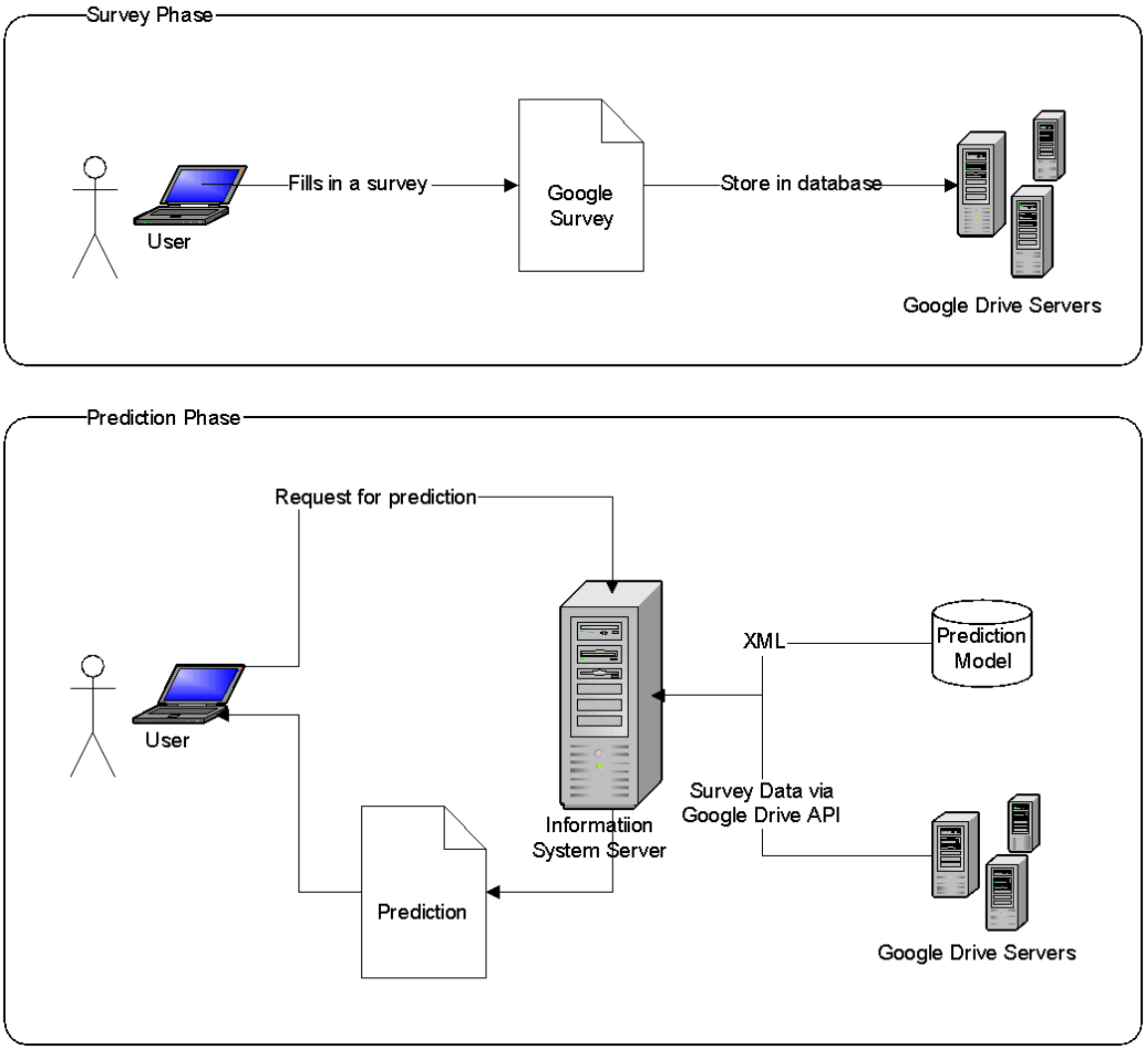


Fig. 47Data flow diagram in I3SP (created by the author)

First, the user fills out a survey and the answers are saved in Google disk (survey phase). The user then requests an analysis (prediction phase) and I3SP generates and returns a result of the company's probability of success and a completed classification tree. I3SP matches the data filled in the survey by the user and the classification tree previously created by software for extracting knowledge from data.

5.2.1 General application architecture

The software I3SP will adhere to an architecture implementing the three-layer model **MVC (Model - View - Controller) (Fig.** 48). MVC is a software architecture design pattern used in the implementation of user interfaces that represents a software application as three interrelated parts, separating the internal representation of information from the information presented to the user. Applications implemented using MVC are easier to develop and

maintain, especially by a team of designers and developers. This leads to better team coordination and fewer mistakes.

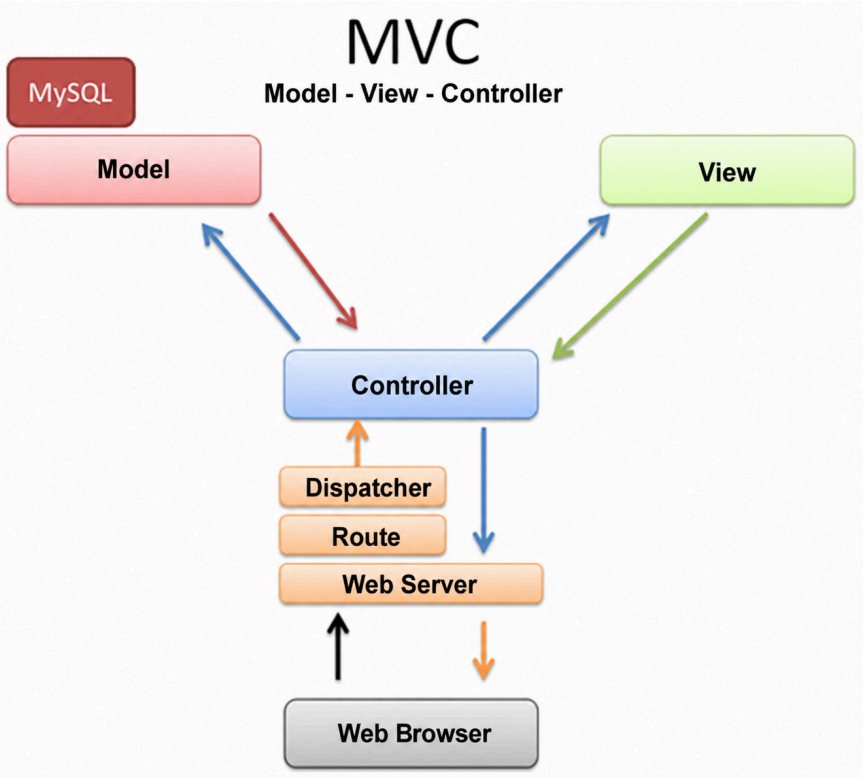


Fig. 48Three-layer model Model - View - Controller for a web-based application implemented with the PHP programming language and a MySQL database (adapted by the author from a figure from betterexplained.com)

The controller (Controler) serves to receive and process incoming requests to the web application. It communicates with both the view and the model. The success prediction software uses a technique called single entry point, which results in all requests sent from the user's web browser being passed from the web server to the same application program file. This file contains the logic for dispatching requests inside the application and, based on the parameters passed in the request, decides which data processing method to run. A single entry point architecture is considered a good practice because all files containing the business logic of the application remain hidden from external users. For the correct implementation of this approach, some specific settings are made on the web server. Rewrite rules are created and declarations and files are added to restrict access to certain directories on the web server.

The model (Model) contains the main part of the business logic of the application and builds the functionalities of the individual modules. It communicates with the database and with third-party APIs.

The view (View) takes care of presenting the data created by the model to the user. It includes presentational and graphical templates and styles for presenting data in different types of pages, and contains almost no business logic. Within the I3SPapplication, the presentation layer also has an additional very important function that does not fit strictly into the three-layer architecture methodology. For individual modules of the application, the

presentation layer contains programming code (JavaScript) that establishes a connection with a third-party programming interface and sends and receives data from it, which is necessary due to a specific requirement for the application - to work with the programming interface of Google disk.

In Fig. 49 shows schematically the architecture of the forecast application, the connections between individual functionalities and the interaction with external applications and program interfaces.

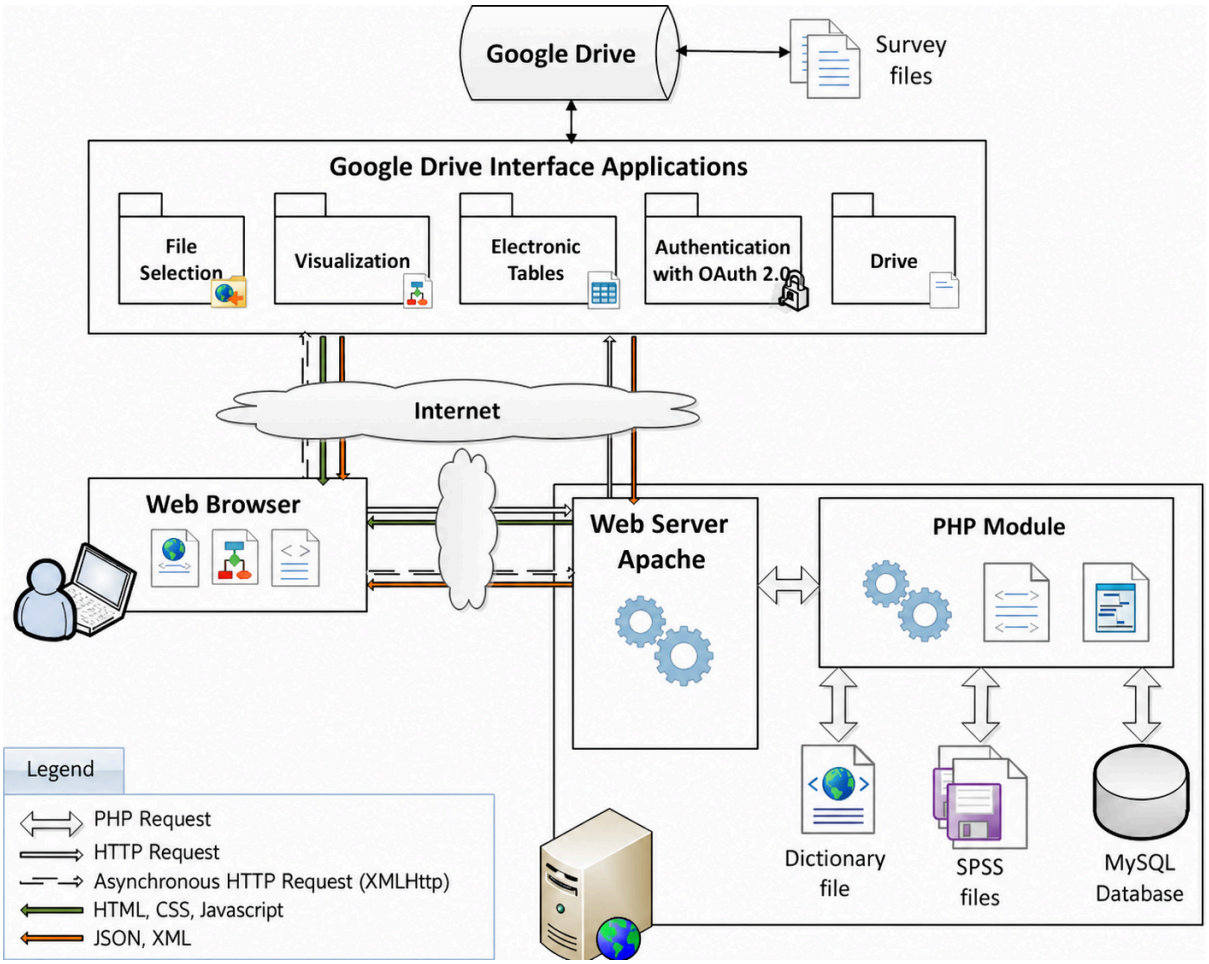


Fig. 49Schematic presentation of the application architecture and functionalities (created by the author)

5.2.2 Application modules

The software part of I3SP will be a client-server web application that is built from several modules with different type of functionality. Each of these modules consists of one or more program code files. The connection between the individual modules is made through a navigation controller. For the implementation of parts of the code with greater complexity and the need for better structuring, the object-oriented model, implemented through the PHP syntax, is used. For the remaining parts, the procedural model of PHP is used, which allows faster development of simple functionalities from the application. The main modules in I3SP are listed in the following subsections.

5.2.2.1 Login module

The login to the system will be done through an external authentication method – through Google account. In order for a user to use the I3SPapplication, they will need to have a valid Google account, as the user is required to access data in a survey file that resides on Googleservers. The user's account can be used to access the application and the survey file at the same time.

The solution that will be implemented in the prediction software is to enter the name and password only once - when the application is initially opened. After successful authentication with a Googleaccount, the application will receive from the program interface of Google an access key (access token), which will subsequently be used to access the survey file, without the need to enter a username and password again.

Restricting access to the application to only certain users is required. To achieve this goal, after the user is added to a list of registered users and after approval by an administrator, his profile is activated. The user list is managed by users who have administrative rights. They can approve, modify, delete and delegate rights to other user accounts.

5.2.2.2 Administration module

Users in the system who have administrative rights have access to advanced functionality. Administrators receive a notification by mail when a new user requests access to the application. They have the ability to view a list of all users, can activate pending approval and deactivate those who no longer need to use the application or have violated the terms of use. User activation is required to restrict access to the application.

Users with administrative rights take care of the input and updating of classification trees obtained by data mining software products. The classification trees will be updated periodically as more company data is collected, and this change will be done manually. In this way, the accuracy of the forecast will increase. Administrators have the option to update the dictionary file, which is one for the entire application and serves to make a connection between the survey data and the classification trees.

5.2.2.3 Setup Module

Once a user is approved by an administrator, they are given the right to use the system. The first thing any user should do after successfully logging in is to set up the two main configurations – survey file selection and prediction algorithm (classification tree).

The survey file is located on a remote server - Googledrive. This file can only be accessed after successfully authenticating a user with their Googleaccount. To display on the user's screen a list of all survey (or spreadsheet) files from his Googledrive, a spreadsheet interface will be used. Through this interface, the user will be given the ability to select a survey file and perform analysis (predictions) on selected rows (survey responses) from this file. Implemented functionality that checks if the selected file is valid.

The user needs to select a classification tree (algorithm to make a prediction) before starting to work with the application. The application offers a choice between different classification trees generated by data mining software IBC SPSS Modeler, Weka or other. This allows users to compare the results obtained for the same data when applying different classification trees. Each user will have a default classification tree selected.

5.2.2.4 Survey Data Module

Before generating a success prediction, the user must select which data will be analyzed using the survey data module. This is done by viewing the contents of the selected survey file (in

tabular form) and selecting a row for analysis by clicking on it. Since the survey file is saved on a remote server (Google disk) andthe application does not have direct access to it, the programming interface of Googlefor working with spreadsheets. This interface allows direct requests to the remote server and receiving data from the selected survey file. Communication takes place directly between the client browser and Googledrive via an asynchronous call, relying on the fact that the user has authenticated that they have access to the file in question.

5.2.2.5 Analysis and Forecast Module

The Analysis and Forecast module is the most important part of the software for predicting the success of start-ups. The module will accept as input parameters:

- identification key of the query file,
- selected row/index from the same file
- parameter for the classification tree that will be used by the calculation algorithm.

For the module to function correctly, these parameters must correspond to existing objects and refer to the current user in the system. Analysis of a file that does not belong to the current user will not be allowed. The survey file data is not available in the local database and the application generates a request to Googledrive to access it. For this purpose, the Googleprogram interface for working with spreadsheets is used. Because the data retrieved from GoogleDrive must be processed on the application server,a mechanism is needed to call the Google API from the application server code.

Once the data is received from the remote server,it must be converted into a format convenient for the application and calculation algorithm to work with. Data conversion will happen using the dictionary-file. The algorithm that makes the success prediction, needs the transformed data, from the poll order,and from the classification tree data.

The main action in the operation of the algorithm is traversal of a node of the classification tree. At each step, the algorithm decides which node to visit, depending on the data received from the polling queue. The result of the work of the prediction algorithm is to reach an end node of the classification tree. Each terminal node of the tree contains a percentage probability of success for the start-up.

The received data representing a forecast will be presented to the user in graphical form. The graph generated at the end of the algorithm's work makes it possible to track the performance of the prediction algorithm, and with its help, the key success factors of a start-up company can be discovered. The graph is drawn in the form of a tree with all data from the survey line plotted. Each node in the tree contains the data pertaining to a specific question and answer from the survey and data from the selected classification tree.

The analysis module also contains functionality to save analysis results, so that analyzes of different companies or analyzes of one company at different points in its development can be compared. There is also the possibility of recalculating an already made prediction with the same survey data, but a different classification tree, in order to be able to make a comparative analysis of the results of the prediction algorithm with different classification trees.

5.3 Development of the Success-Prediction Software

This point deals with the actual implementation of the application. Basic concepts in its construction are presented and the way in which the selected technologies and algorithms are applied to implement the designed functionality is explained. Code snippets are quoted to give a clearer idea of the implementation of specific functionalities. The software was developed as a dissertation under the supervision of the author.

5.3.1 Prediction Algorithm

The main purpose of I3SP is to provide a possibility to predict the probability of success of start-up companies.

In order to make a forecast, we must have a completed survey. Then, using a classification tree created by data mining software, rules are applied to classify the company as "successful", "neither successful nor unsuccessful", or "failed". The classification tree is written and stored in a file, with each line of the file representing a node of the tree. The software allows tree update. The prediction algorithm works by taking the data from the survey file, applying the rules written in the classification tree, applying the rules to transform the data, and returning a result. The sequence of the described process is illustrated by UML action diagram (Fig. 50)

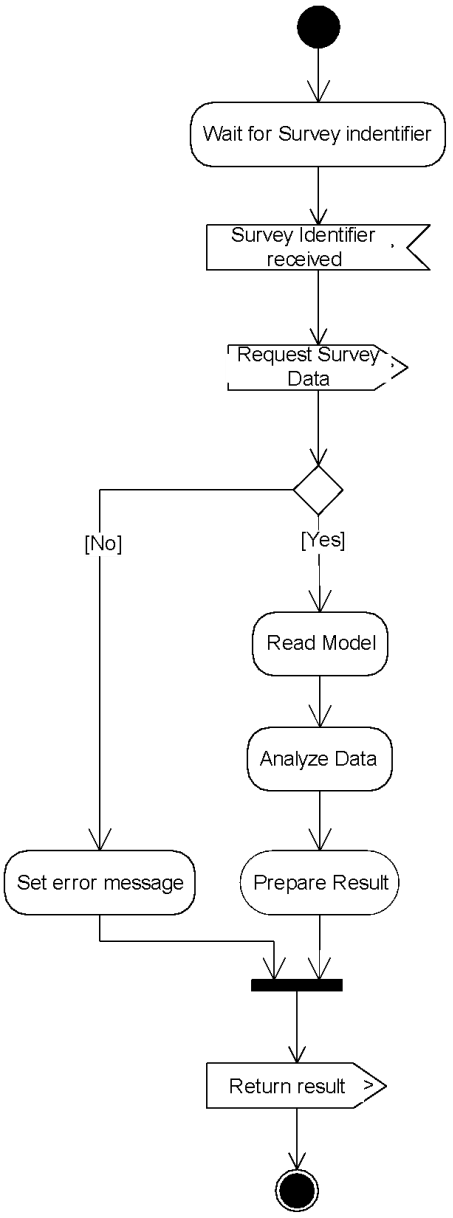


Fig. 50UML diagram of actions in predicting the success of a start-up company with I3SP (created by the author)

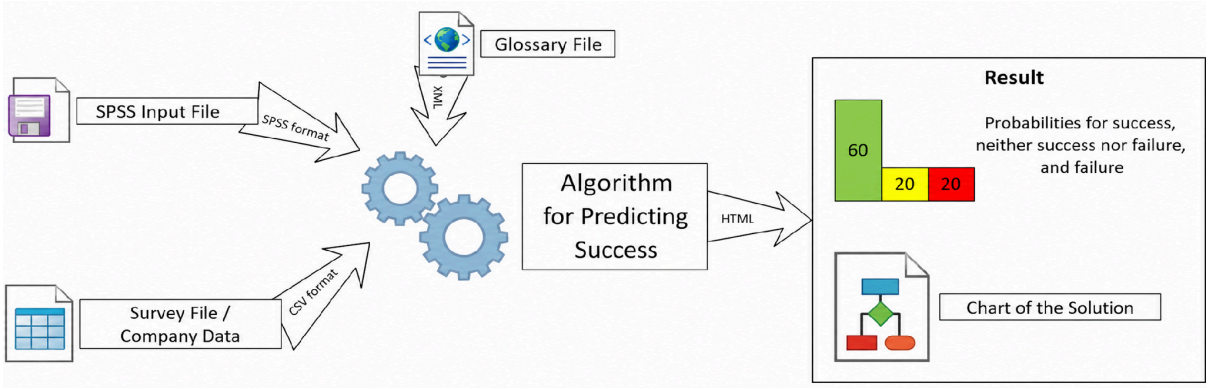


Fig. 51Scheme of the application forecast algorithm (created by the author)

The implementation of the forecast algorithm begins with loading the necessary input parameters according to the scheme shown (Fig. 51), which is carried out by calling the following methods:

```
$survey_data = FileService::convertCSVStringToAssocArray($survey_string);
$dictionary = FileService::convertDictionaryToArray();
$prediction_tree_array =
FileService::convertPredictionTreeToArrayWithTranslation($prediction_file,
$dictionary);
```

The listed methods convert necessary information from data files in associative arrays in PHP for more convenient operation. When creating the array "$prediction_tree_array" corresponding to the classification tree file from the data mining software, some additional activities are also performed - IDs are created for each node, and the predecessor node is calculated and recorded to be able to obtain a tree structure, represented in an array.

The algorithm uses a dictionary file that is used to associate the survey data with the classification tree, and which is specific to the survey questions. To the data of the classification tree , all the data from the dictionary file "$dictionary" for the corresponding nodes of the tree are added. The data from the dictionary file is also added to the responses from the survey file. This operation unifies the data and eliminates the problem of different representation of the data in the classification tree and the survey file.

The next step is traversing the classification tree. Crawling is not done recursively, but sequentially - in levels. All nodes at a given level in the tree must have the same name (contain the same survey question), which is a specific feature of the classification tree file. Nodes differ in the values they hold for different answers to this question. If we assume that at level 1 in the tree after the question "The company has a competitive advantage" there are two nodes and one of them is a branch for the answer "Yes" and the other - for the answer "No", then to classify the company it is enough to check which node contains the same value as given in the survey. After finding this node, the algorithm eliminates the remaining nodes of the same level and continues the traversal in a lower level, interested only in nodes of which

the selected node from the previous level is an ancestor. The process continues until an end node (leaf) is reached.

The endpoint data for the company's percent probability of success is actually the result of the prediction. There is also a special case where traversing the prediction tree may reach a node that does not contain a valid value according to the poll response. This is possible, for example, when the survey has a missing answer for a question involved in the classification tree. In these cases, the algorithm returns a result (success prediction) from the last crawled node and gives an indication that it has not completely completed the crawl.

5.3.1.1 Dictionary file

The I3SP software accepts data from various sources: the survey responses from the company (contained in the survey file from Google drive),the classification tree (which is generated by data mining software). The data in the survey file is in the form of text describing each question and its associated answer, and the data in the classification tree file is a combination of numbers and abbreviated strings. Some of the numbers have the same meaning as the answers in the survey file. For example, “1” means strongly disagree for questions answered with a degree of agreement or disagreement. All abbreviated strings are variable names used in data mining software and correspond to survey questions. For example, the string "Competitive advantageMybusinessp" is associated with the question "Competitive advantage [My business has a clear competitive advantage.]". In order to standardize the presentation of the data and facilitate the operation of the application, a special dictionary file in **XML format** has been created. The file is managed by the software administrators who can update it. In Fig. 52a fragment of the contents of the dictionary file is shown.

```
<dictionary>
    <element index="R" type="scale">
        <short_name>Опит_Имамуправленскиопит</short_name>
        <question>Умения и опит [Имам управленски опит.]</question>
        <answers>
            <item key="1.000" value="Напълно несъгласен"/>
            <item key="2.000" value="По-скоро несъгласен"/>
            <item key="3.000" value="Нито съгласен, нито несъгласен"/>
            <item key="4.000" value="По-скоро съгласен"/>
            <item key="5.000" value="Напълно съгласен"/>
        </answers>
    </element>
```

Fig. 52Fragment of the dictionary file used in the application (created by the author)

The dictionary file is hierarchically structured in XML format, with each "<element>" tag representing a question from the survey file and having two attributes: "index" and "type", which respectively indicate the column number in the survey file (represented by a spreadsheet) and the type of question in the survey (free text, scale, etc.). Each item has nested tags. One of them is the "<question>" tag, which contains the text of the corresponding survey question. The short representation of the question used in the classification tree is recorded in a "<short_name>" tag.

The possible answers to the survey questions are written in the "<answers>" tag containing several "<item>" elements. The <item> tag attributes contain the possible responses represented as a string (used in the poll) and a numeric value (used in the classification tree). The answers to some survey questions are in free text, in which case it is not practical for every possible string input to have a value represented, and another attribute is used for the "<item>" tag, which contains a pattern matching different strings.

5.3.1.2 Regular Expressions

The file generated by the data mining software,is text-based and contains data about the structure and nodes of the classification tree, represented by a specific syntax. This file is used as an input parameter for the prediction algorithm and is key to the operation of I3SP. To recognize the data from the file, a regular expression (Fig. 53) is used, which is implemented through the PHP function preg_match(). The variable "$parsedOK" saves the processed data.

```
$parsedOK = preg_match('/(?P<offset>\s*)(?P<name>[a-zA-Z\p{Cyrillic}0-9_-\s]+)'
    . '( in \[ )*( = )*(?P<values>[0-9. ,]+)( \])*( )*\[ Mode: (?P<currentMode>\d) \]'
    . '( => )?(?P<exitMode>[0-9,]*) \((?P<Ncount>[0-9,]+)[; ]*'
    . '(?P<Probability>[0-9,]*)\) \((?P<percent_no>[0-9.,]*);'
    . '(?P<percent_yesno>[0-9.,]*);(?P<percent_yes>[0-9.,]*)\)[ $]*/u'
,$line
,$matches);
```

Fig. 53Regular expression to extract data from a file containing a classification tree (created by the author)

5.3.2 Success-Prediction Results Graph

The main purpose of I3SPis to calculate the probability of success of start-up companies. The prediction algorithm returns as a result the probability of success, but the results presented in this way to the end user are unclear and incomplete (according to the answers of the respondents in the conducted qualitative research) because they do not contain information about the operation of the algorithm and about the factors that led to the decision. To make the application more useful to the end user, a **classification tree type graph (Fig.** 54) has been implemented, which visually shows how the algorithm traverses the nodes of the tree and why the results were obtained.

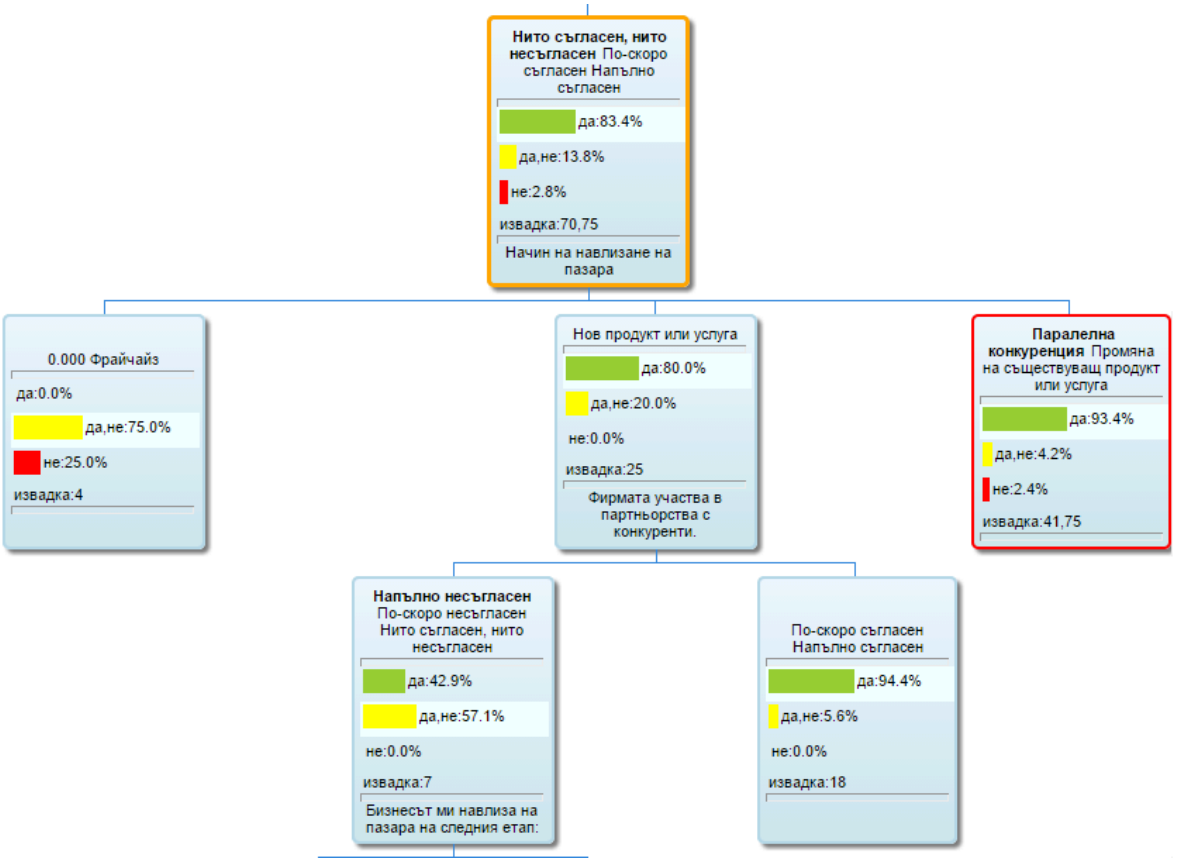


Fig. 54Fragment of the classification tree presented in the analysis page (created by the author)

For the visualization of the tree, a graph visualization tool was used, which is part of the programming interface Google Visualization API (Google Inc., 2014). It is implemented in the JavaScript language, runs in the client web browser, and is used to draw a graph with a tree structure.

The interface expects as a parameter a specially formatted object in JSON format, which contains the information about the structure of the tree (the connections between individual nodes), as well as information about the content of the specific nodes. Each node of the classification tree contains a survey question, the company's probability of success, and survey response values for the previous question at which the algorithm will reach the node. In each node, information is also shown about exactly which of the possible answers for that node was given by the user (ie, the answer from the current survey is plotted in the node).

The prediction algorithm traverses the classification tree and at each visited node that is part of the solution, node information is recorded. At the end of the crawl, a collection containing the crawl path is returned. Finally, all data is fed to a design template, which is interpreted by Smarty, and which generates the HTML content of the particular node. Construct a JSON string that contains information to draw the entire tree, taking the data from traversing all nodes. The JSON string is written directly to the source code of the page as a JavaScript variable. The following is a snippet of the JavaScript code (from the template file "templates/survey_data_analyse.tpl") that takes care of drawing the chart:

```
var data = google.visualization.arrayToDataTable({$jsonTree});
var chart = new google.visualization.
OrgChart(document.getElementById('pnlChart'));
chart.draw(data, {allowHtml:true});
```

A JavaScript object ”chart”is created, which receives a JSON string with instructions to draw the tree, and the draw method “chart.draw” is called.

5.3.3 Google APIs

Google APIs (Google Inc., 2014) enable various software applications to interact with applications created by Google. Examples of such program interfaces are Googledrive (Google Drive), Googlemaps (GoogleMaps), social network (Google+) and others. The specificity of I3SP requires working with similar program interfaces, since the survey results, used by the software are stored on remote servers of Google disc. The program interfaces that will provide access to this data are:

- **Google Drive API-** for working with the files stored in Google drive,
- **Google Spreadsheet API**– for working with spreadsheets stored in Google disk,
- **Google Visualization API** – programming interface, which enables displaying the data from Google disk in various graphic formats.

5.3.3.1 Google Drive API

Google Drive (Google Inc., 2014) enables Google users to create, modify and store various file types in the Googlecloud arrays by providing convenient ways to share, protect, and back up that data. The survey file that I3SPprocesses is created using Googlesurveys (Google Forms) and is stored in Google disk space.

The Google DriveAPI will allow the application to work with the survey file. It offers options to create new files in different formats, to download a file,to access the content and metadata of a file.

In I3SPthere is a need for a mechanism that allows users to view the contents of their own files in Google Drive and select an appropriate survey file (created by them or shared by another user). To implement this part of the application functionality,the file selection and browsing functionalities of the user interface Google drive will be used.

The program interface Googledrive can be used only after a successful login with oAuth 2.0 protocol (OAuth community, 2014) and access authorization.

5.3.3.2 Google Sheets API

The GoogleSpreadsheet API (Google Inc., 2014) provides a wide range of functionality that can be applied to spreadsheets stored in GoogleDrive. This interface makes it possible to extract the contents of an entire spreadsheet, extract the contents of only a single worksheet (worksheet), extract the data for specific rows of that worksheet, and also extract data from a specific cell in a selected worksheet. Queries written in a special language similar to SQL to extract and combine specific information **from** a spreadsheet, sort and group by certain criteria are also supported.

The application for predicting the success of start-up companies has a need to retrieve a particular row from the survey file spreadsheet and sort the rows of that table, which is achieved with appropriate queries in the API.

The program interface Googlespreadsheet can be applied only after a successful login with oAuth 2.0 protocol and access authorization.

5.3.3.3 Google Charts API

The program interface of Googlefor working with graphics tools accepts data sent by the client browser in the form of XML and JSON, processes it and sends ready HTMLcode that visualizes this data using the client browser. It provides the ability to visualize the data in various graphical objects, such as tables, charts, trees and many others. I3SPuses this API to create a tree graph that depicts the performance of the prediction algorithm.

Using the Google API for working with graphical tools does not require authentication.

5.3.4 Data Storage

I3SPrelies on three different data sources for its work. These are data stored in a local database, in local files saved on the server's file system, as well as data stored on remote servers and accessible only through a program interface. The use of all three different data storage methods used in the application should be considered.

5.3.4.1 Database structure

Databases provide a very fast and convenient way to store and retrieve data. For the purposes of I3SP, a MySQL database will be used. The data that will be recorded in the database is related to the users of the application, the classification tree (generated by software for extracting knowledge from data), as well as data of already performed analyses. The information in the database will be stored in Bulgarian and English, therefore collation "utf8_general_ci" was chosen, which supports UTF8 character encoding, including Cyrillic, Latin and other alphabets and symbols. Fig. 55illustrates the database structure used by the application. The diagram shows that the application will use three tables: users, prediction_tree and analyses.

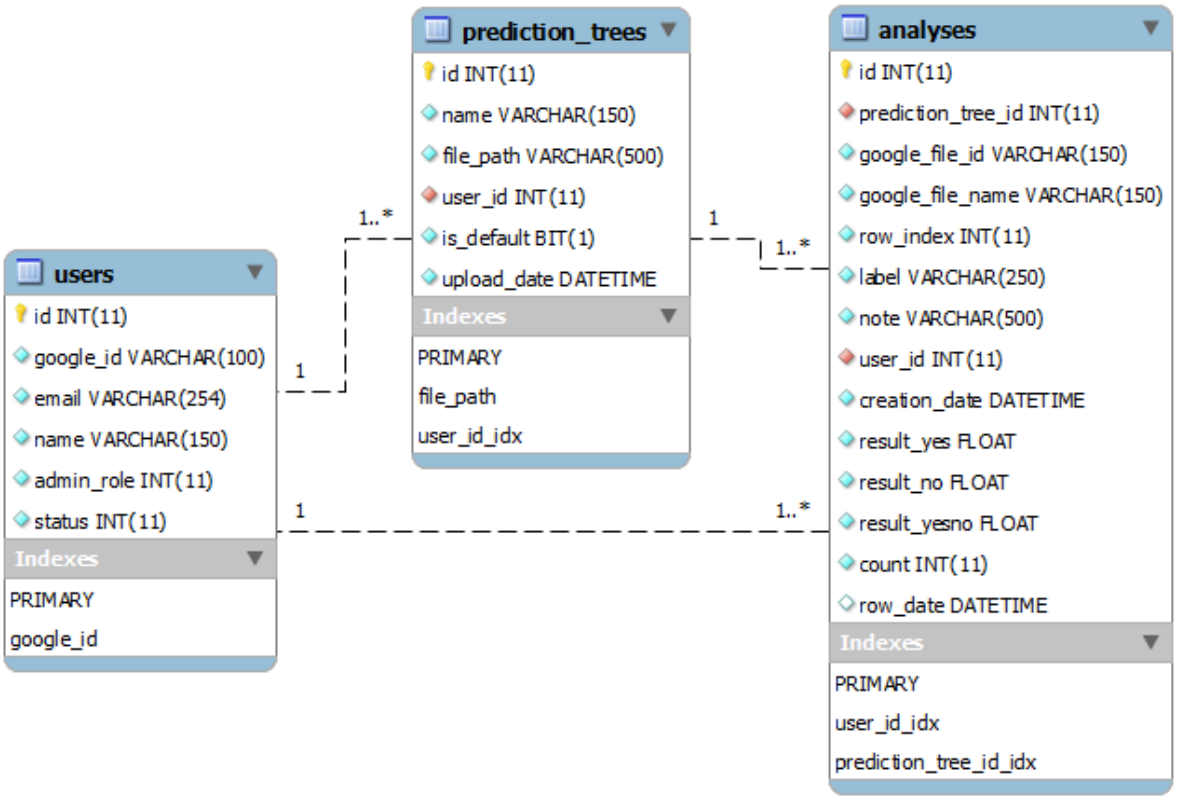


Fig. 55Application database structure diagram (created by the author)

Users tablecontains locally stored information about user data. Much of this information will be copied from Googleaccounts. The table has a primary key created (the "id" column) that is used in the other two tables and serves as a link to them (relation). In users, there is a restriction that emails are unique (unique constraint). The admin_role column contains information about the rights of a given user in the system - whether he is an administrator or a regular user.

The **table prediction_trees**stores information about various classification trees that the application uses. Only tree metadata generated by IBM SPSS Modeler or other data mining applications is stored in the table. The trees themselves are stored as text files in the file system, and the table stores the path to the files. It has a unique primary key that will be used as a link to the analyzes table, as well as a key to the users table indicating which user created the entry in the prediction_treestable.

Analyses tablecontains an archive of information about the results of success predictions made by different users for different companies. It contains fields for storing the name and identifier of the Googledisk file (survey file) to which the prediction refers, as well as an index (number) of a line of the file. The table also contains external links to the tables users and prediction_trees, which store data, respectively, about the user who saved the prediction and about the algorithm used in its generation. The table only stores the results of the success predictions (percent probability of success) and does not contain data from the survey file and the tree.

5.3.4.2 Local files

The application uses local files to store some specific types of data - a dictionary file and a classification tree.

The dictionary file is key to the operation of the application and is structured as an XML document containing information describing the rules by which the classification tree data is mapped to the survey data. PHP offers convenient and fairly fast programming tools for working with XML.

Classification trees are tree-like algorithms created using modeling software - IBM SPSS Modeler or Weka. These are text files, formatted in a specific way,set by the data mining software. Each new line of these files represents a node of the classification tree. The data in each row has a uniform format and provides information about the particular node of the tree, the level of nesting, and the corresponding decision for the success of the company.

```
1  КонкурентнопредимствоМоятбизнесп in [ 1.000 2.000 ] [ Mode: 2 ] => 2,0 (6; 0,833)
2  КонкурентнопредимствоМоятбизнесп in [ 3.000 4.000 5.000 ] [ Mode: 3 ] (101)
3  » НМАктив_изграденабизнесрепут in [ 1.000 ] [ Mode: 1 ] (10,1)
4  » » Индустр_Имабариеринавлизане in [ 1.000 ] [ Mode: 2 ] => 2,0 (2,1; 1,0)
5  » » Индустр_Имабариеринавлизане in [ 2.000 3.000 4.000 5.000 ] [ Mode: 1 ] (8)
6  » » » Опит_Имамтехническиуменията затехно in [ 1.000 2.000 3.000 4.000 ] [ Mode: 1 ] => 1,0 (4; 1,0)
7  » » » Опит_Имамтехническиуменията затехно in [ 5.000 ] [ Mode: 3 ] => 3,0 (4; 0,75)
8  » НМАктив_изграденабизнесрепут in [ 2.000 3.000 4.000 5.000 ] [ Mode: 3 ] (90,9)
9  » » ФУспех_Външнасреда = 1,000 [ Mode: 2 ] (6)
10 » » » Страт_Стратегиятакоятосъмизбрале in [ 1.000 2.000 ] [ Mode: 2 ] => 2,0 (3; 1,0)
11 » » » Страт_Стратегиятакоятосъмизбрале in [ 3.000 4.000 5.000 ] [ Mode: 1 ] => 1,0 (3; 0,667)
12 » » ФУспех_Външнасреда = 2,000 [ Mode: 3 ] (84,9)
```

Fig. 56Fragment of an IBM SPSS Modeler source file (created by the author)

In Fig. 56shows a fragment of a file containing a classification tree generated using IBM SPSS Modeler. As part of the start-up success prediction analysis, the tree will be visualized and available for the user to view. For this purpose, a mechanism will be created through which the tree will be converted into a format understandable by the software. To implement this mechanism, a regular expression will be used to extract the data from the tree and a dictionary-file,through which the data is converted from one format to another, according to the needs of the application.

5.3.4.3 Files Located on Google Drive

I3SP needs data (survey file) to analyze and on which to create a success prediction. The survey created as part of the prediction model was implemented using a Google, tool called Google Surveys (*Google Forms*) (Google Inc., 2014). The format in which data submitted through the survey is recorded is rows of a spreadsheet. An example fragment of the query file is shown in Fig. 57.

	A	B	C	D	E	F	G
1	Timestamp	Личност и ценности [Аз съм уверен в себе си.]	Личност и ценности [Аз съм предприемчив, инициативен.]	Личност и ценности [Аз съм амбициозен.]	Личност и ценности [Склонен съм да поемам рискове.]	Умения и опит [Имам умения по мениджмънт.]	Умения и опит [Имам умения по маркетинг.]
2	6/23/2013 9:05:51	По-скоро съгласен	Напълно съгласен	Напълно съгласен	По-скоро съгласен	По-скоро съгласен	По-скоро съгласен
3	6/27/2013 6:16:47	Напълно съгласен	Напълно съгласен	Напълно съгласен	Напълно съгласен	Напълно съгласен	Напълно съгласен
4	6/27/2013 23:10:36	Нито съгласен, нито несъгласен	По-скоро съгласен	По-скоро съгласен	По-скоро съгласен	Напълно съгласен	По-скоро съгласен
5	6/28/2013 11:06:40	Напълно съгласен	По-скоро съгласен	По-скоро съгласен	По-скоро съгласен	Нито съгласен, нито несъгласен	Нито съгласен, нито несъгласен

Fig. 57Fragment of Survey file (created by the author)

Each line of the survey file represents a response of a company,that completed the survey. The file contains more than one hundred columns (corresponding to survey questions), and the row values represent response data to those questions.

To predict the success of a company, the datamust be pulled from the spreadsheet and processed by the application. They are not saved locally, but stored in Google drive and updated when completing surveys with Google surveys, and accessed by I3SPwhenever needed. Direct access will not generate a lot of traffic that will make it difficult for the application, since the volume of transmitted information is not large. The data access mechanism is through the Google programming interfaces.

In order to convert the data into a format understandable by the software, a dictionary-file is used, which turns the survey data into user-friendly forms for the application.

5.3.5 Functionality Using Google APIs

Requirements for I3SPrequire the use of survey files located on a remote server Googledisk, authentication of users with Google Account and graphical visualization of the data – functionalities for which Googleprogramming interfaces are applied.

5.3.5.1 Authentication

In order to use a given Google program interface, certain requirements must be met. It is mandatory to provide an identification key as a parameter when requesting the interface. Thus, the interface verifies that the given application (using a user Google account) has rights to the requested resources (in this case – a file located on Google Drive).

The identification key is obtained after successful authentication with oAuth 2.0 protocol and is valid within 60 minutes. It is stored in a session variable. Calling a GoogleAPI is done by requesting access to it, for example "http://spreadsheets.google.com/tq". The communication between the application and the interfaces is done via the HTTP protocol.

To facilitate developers, libraries and examples are provided with each API. These libraries are implemented in various programming languages, such as PHP, ASP. NET, Java, JavaScript. Using the libraries is optional, but recommended as it allows for easier and faster development. In the development of the success prediction software, a library for working with Google disk,written in PHP was used. The Google APIs are accessed from both the server side and the client side. The server part of the application implements the "GoogleServices" class to work with the Google APIs and to retrieve data from Googledisk. The following code snippet shows the constructor of the "GoogleServices" class:

```
public function __construct() {
```

```
$this->client = new Google_Client();
$this->client->setApplicationName(APP_NAME);
$this->client->setClientId(CLIENT_ID);
$this->client->setClientSecret(CLIENT_SECRET);
$this->client->setRedirectUri(REDIRECT_URI);
$this->client->setScopes(array(
'https://www.googleapis.com/auth/drive',
'https://www.googleapis.com/auth/drive.file',
'https://www.googleapis.com/auth/userinfo.email',
'https://www.googleapis.com/auth/userinfo.profile',
'https://spreadsheets.google.com/feeds'));
}
```

In the constructor, an instance of the "Google_Client" class (defined in the Google library) is created and all necessary parameters are set to work with the interfaces correctly. This instance is a field in the GoogleService class ($this->client) through which the interface is accessed in the other methods of the class. The constants "CLIENT_ID" and "CLIENT_SECRET" are obtained from the developer when registering the application with Google and are stored in the configuration file of the application.

5.3.5.2 Retrieve User Profile Information

The following snippet is a method of the Google_Client class, part of the I3SP code, which interfaces with the Google API. to work with users and retrieve information about the Google application user profile:

```
function getLoggedUserInfo() {
$token_decoded = json_decode($this->client->getAccessToken(),true);
$query = 'https://www.googleapis.com/oauth2/v1/userinfo?access_token=' .
$token_decoded["access_token"];
$user_info = file_get_contents($query);
return json_decode($json,false);
}
```

5.3.5.3 Working with Google Drive

“GoogleServices” is a PHP class that has methods to retrieve data (single row or all data) from a file in Googledrive. The following program fragment, shows a class method retrieving the entire contents of a file from Google drive.

```
function getSurveyDataWhole($fileId){
```

```
$driveService = new Google_DriveService($this->client);
// Retrieve metadata for the file specified by $fileId.
$file = $driveService->files->get($fileId);
// Get the contents of the file.
$response = $this->client->getIo()->authenticatedRequest($request);
$Data = $response->getResponseBody();

return $Data;
}
```

The other way the application connects to the Google API is directly from the client browser - using the JavaScript programming language. One or several script files are loaded into the page directly from Google's servers - depending on the specifics of the program interface. This is done in the front-end part of the application, due to the specifics of the JavaScript programming language, which is interpreted by the client browser. The following is a snippet of JavaScript code that demonstrates working with the Google API using JavaScript.

```
this.picker = new google.picker. PickerBuilder().
addView(google.picker. ViewId. SPREADSHEETS).
setAppId(this.clientId).
setOAuthToken(this.accessT).
setCallback(this._pickerCallback.bind(this)).
build().
setVisible(true);
```

The quoted code calls the Googledrive file picker interface (Google Picker API). When the shown program code is executed, a client browser dialog opens and all files (spreadsheets) from GoogleDrive that the user owns are loaded inside.

I3SPalso works with the Google spreadsheet interface via JavaScript. Its characteristic is that it is possible to construct queries (similar to SQL) through which specific information is extracted from the spreadsheet. An example of such a query is:

```
query.setQuery('select * order by A desc limit ' + $('#last_rows').val());
```

The result of this query is the last lines of the selected survey file.

The data received from the GoogleAPIs can be in various formats,with JSON and XML being the most commonly used.

5.3.6 Design templates

In I3SPthe Smarty template tool is used, which separates the code implementing the business logic of the application (developed in PHP) from the code implementing the visual part of the application (developed in HTML, CSS, JavaScript). This helps separate the view layer from the model layer of the three-layer architecture (MVC). Smarty is implemented in the PHP programming language and its code is stored in the main application directory. In addition, several service folders are created, among which is the "templates" folder. By default, it stores all templates used by the application. All views that are served to the end user are implemented with Smarty design templates. Data passed from the logic implementation layer to the template generator,is encapsulated and visible only within the template. There is no need to mix PHP and HTML code, which improves the design of the application.

In Fig. 58 shows a code snippet of a design template that contains part of the front-end code of the application's prediction algorithm settings page.

```
{foreach name=outer item=prediction from=$predictions}
<tr>
 <td>{$prediction["name"]}</td>
 <td>{$prediction["file_path"]}</td>
 <td>{if $prediction["is_default"] eq 1}Да{else}Не{/if}</td>
 <td>
  <a target="_blank" href="{$prediction["file_path"]}">преглед</a>
  {if $prediction["is_default"] eq 0}
  | <a href="{$SCRIPT_NAME}?
  p=admin_prediction_tree&setdefault=&file_id={$prediction["id"]}">по подразбиране</a>
  <a href="{$SCRIPT_NAME}?
   p=admin_prediction_tree&delete=&file_id={$prediction["id"]}">изтрий</a>
  {else}{/if}
 </td>
</tr>
{foreachelse}
```

Fig. 58Snippet of design template code in Smarty (created by the author)

5.3.7 Security and reliability

The security and reliability of I3SPare essential from a user perspective and become increasingly important with promotion of the application. In order to ensure the stable and trouble-free operation of the application, as well as the security of user data, special measures are required to be taken, which will be discussed in this point.

5.3.7.1 Compliance with good practices

Frequent options for submitting incorrect data are:

- injecting deceptive SQL code through the GET and POST methods of the HTTP protocol (**SQL injection**). It is most commonly accomplished through manipulated values in text fields and URL parameters;
- uploading files with incorrect names, which under certain circumstances can lead to saving and running a file outside the regulated locations on the server (**file injection**);
- injecting script code into the text fields that arbitrarily collects and sends data from the client browser to other servers (**XSS - cross site scripting**);

- extracting information about session variables (**session hijacking**) and/or manipulating their values (**session fixation**), usually for the purpose of obtaining unauthorized access.

The I3SPapplication has mechanisms to prevent such common attacks. To protect against SQL injection, the application uses a standard PHP function mysqli_real_escape_string(). For XSS protection, data from text fields is always filtered through the PHP function htmlspecialchars(), which replaces by encoding dangerous HTML characters.

5.3.7.2 Data Validation, Filtering and Preparation

Part of the data that I3SPworks with, including files and database content,is submitted by the users of the application. For various reasons - intentional and unintentional, this data may be incorrect, incomplete, in an incorrect format, or manipulated. In such cases, the data could cause unforeseen consequences in the operation of the application, if the practice of checking and preparing them before their use is not followed.

All the data submitted by the user, before their use, are processed, through the programming tools provided in the language PHP. For example, the data that will be written in the database is prepared with PHP the "mysqli_real_escape_string()" function, which converts them to work-safe ones. User data that is used in the frontend (the generated HTML documents) is converted with PHPthe htmlspecialchars() function, which replaces the HTMLspecial characters with codes.

User-uploaded files are saved with a custom system-generated name to avoid unwanted execution. The application also has methods for validating the content of the survey file and the classification tree file generated by the data mining software. Before running the prediction algorithm, it is always checked that the file-dictionary is valid. Validation is also applied to the classification tree file - when it is uploaded, a method is launched that, using regular expressions, verifies that the contents of the file represent the information necessary for the application to work.

In case of incorrect data input, the system processes the data, and displays on the user's screen an appropriate error message that describes what problem occurred.

5.4 Testing the Success-Prediction Software

This point describes the operation of the start-up success prediction software by using examples, graphs, test scenarios and comments. Screenshots of the main functionalities are shown, representing the user interface and how to use the application and modules.

5.4.1 System login

The software can be used by users who have valid Googleaccounts. When a user initially accesses the application, a short description page and a login button are loaded (Fig. 59). Traditional user registration functionality is not provided.

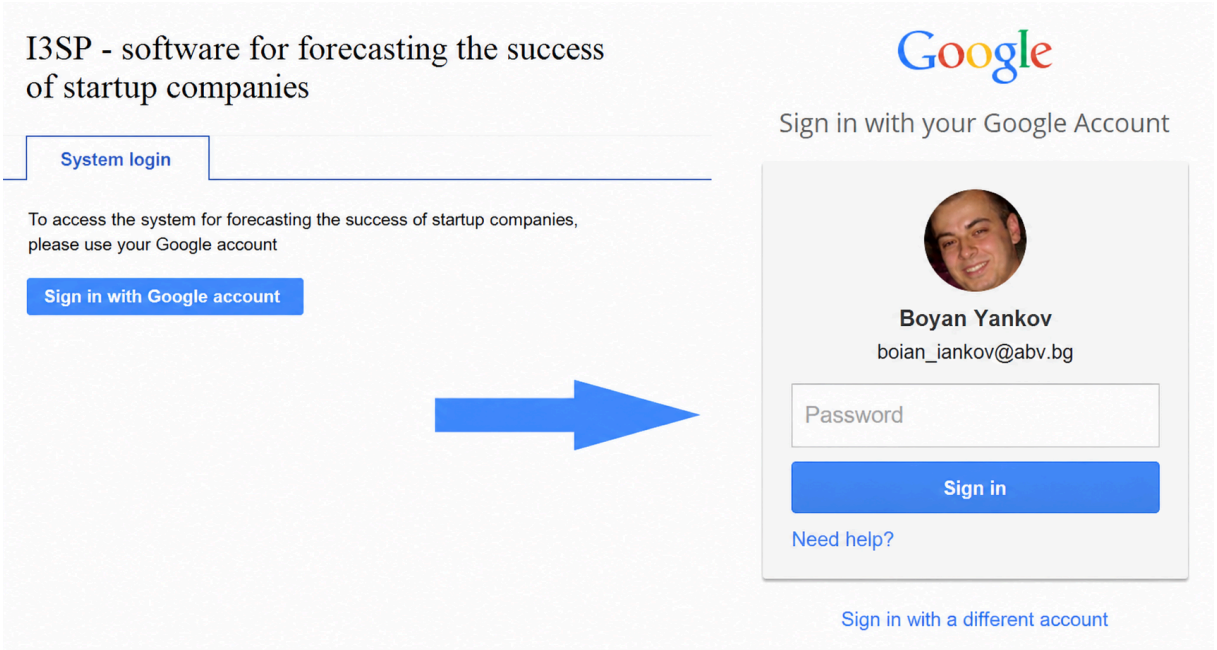


Fig. 59Login to the application with a Google account (created by the author)

After pressing the login button, if the user does not have an active Google session for this browser (session cookie),opens a window for entering a name and password. After entering the correct name and password, or if there is an active session, the user is redirected to a dialog box for allowing access to Googledrive (Fig. 60). This window explains to the user that the application requires access to certain resources from the user's Google account: account information and access to his Google drive files.

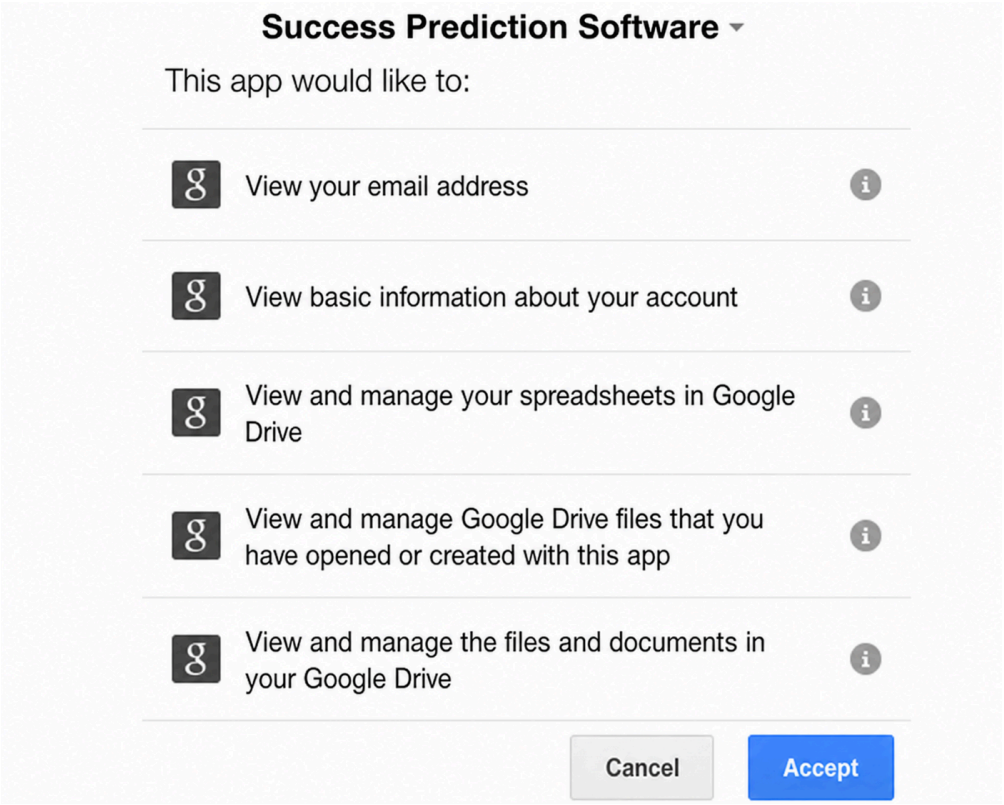


Fig. 60Allow Google Drive access dialog (created by the author)

By pressing the "Accept" button,the user agrees to the terms and allows access to the application. If the user is already enabled by an administrator, he directly enters the application and goes to the settings page. If it is not activated, the user receives a message explaining why he cannot access the application at the moment (Fig. 61) and waits for activation by an administrator.

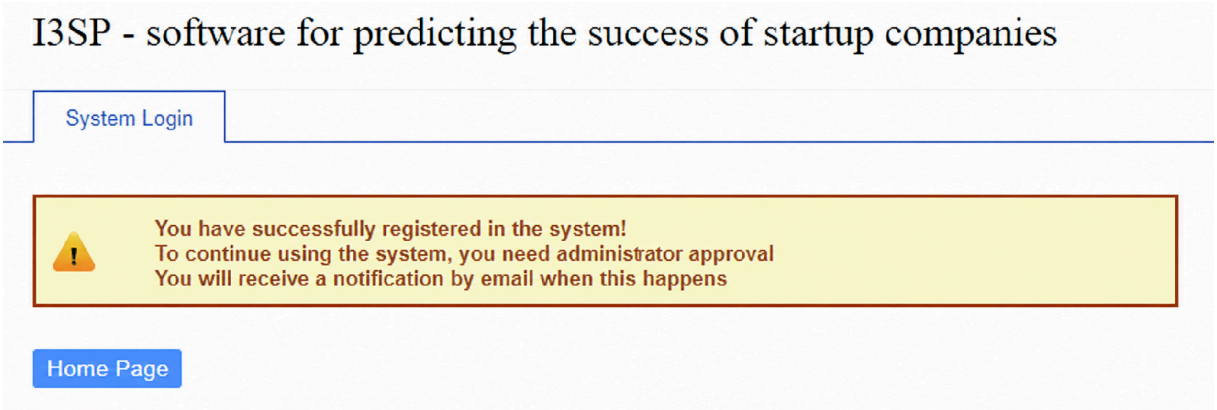


Fig. 61Waiting for approval after system access request sent (author created)

The following test scenarios were successfully executed:

1. Invalid user check: Attempt to login with invalid Google account
2. Valid user check: It is checked that after entering a valid account the user sees the access window.
3. Pressing the button "Cancel″ in the access window: It is checked, that after successfully logging in system,if the user clicks the "Cancel″ *button*, he cannot access the forecast system.
4. Login with an already activated user: Checks that after logging in an already activated user, he successfully logs into the application.
5. Login with an inactive user: Checks means that after a user logs in for the first time, an activation message is displayed and an email is sent to the administrator.

5.4.2 Application Settings

Immediately after a successful login, the user is shown the settings page,from which he can select a survey file and classification tree (algorithm) for analysis (Fig. 62). The default algorithm is selected in advance by the system administrator, but the current user can change this setting.

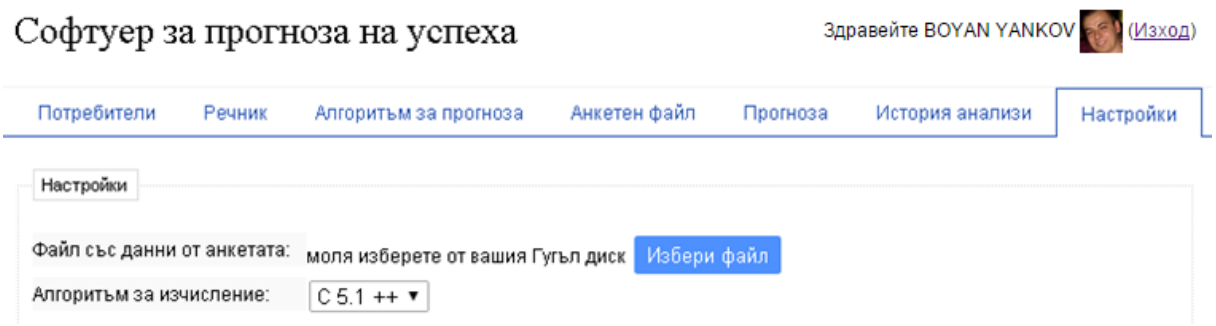


Fig. 62Application settings page (created by the author)

The survey file selection setting is mandatory for the operation of the application in order to be able to make analyzes about the success of the company. When pressing the "Choose file" button. a dialog box is displayed containing the names of all files (of spreadsheet type) from the user's Googledisk space (Fig. 63).

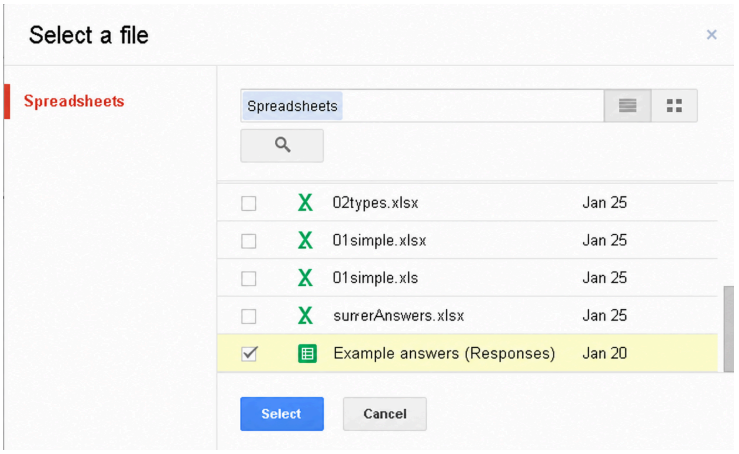


Fig. 63Dialog for selecting a survey file from Google Drive (created by the author)

After selecting a file, the name of the selected file is displayed and a link providing an option to change it.

The following test scenarios were successfully executed:

1. File Selection: Check that a file can be selected after pressing the Poll File button and selecting a file from the dialog box
2. Changing the Poll File: Checking that the file can be changed after there is already a file selected.
3. Canceling File Selection: Checking that after clicking the Cancel button in the dialog box, the poll file does not change.
4. Changing the calculation algorithm: Selecting a new algorithm, and verifying that the algorithm setting is saved successfully.

5.4.3 Viewing the contents of a survey file

After the user has selected a survey file from the application settings, the page for viewing the contents of the selected file becomes available to him. A fragment of the page is shown in Fig. 64.

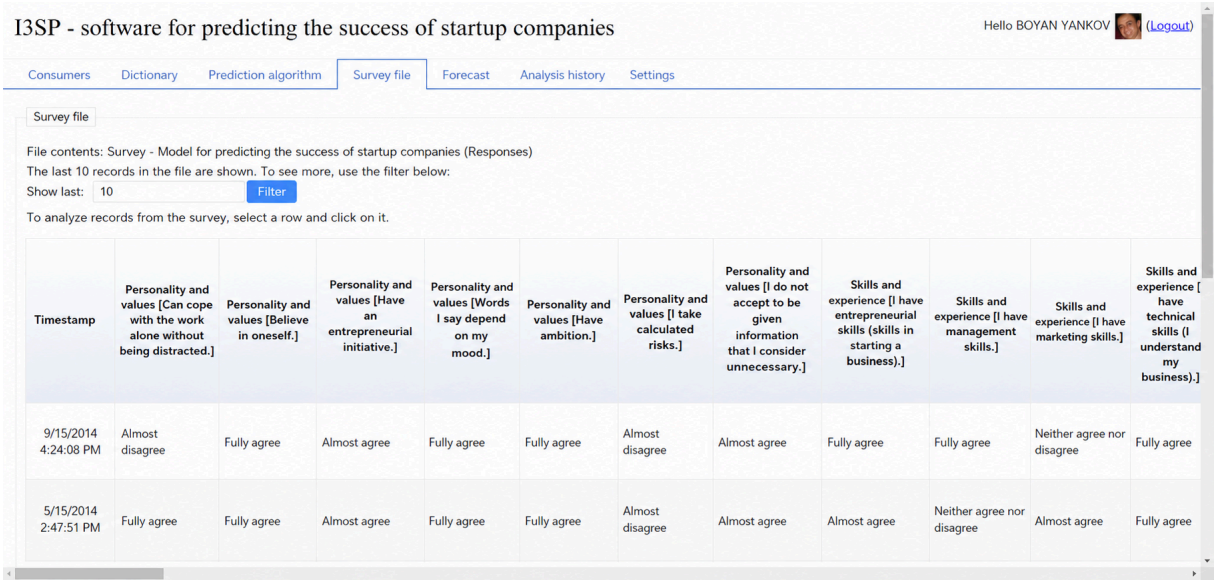


Fig. 64Survey file content preview page (created by author)

Survey file data is displayed in tabular form. Since the file contains more than 100 columns of data (corresponding to survey questions), most of which contain multiple words, a horizontal slider (scroll) is used to visualize the entire rows. A filter is provided on the page, which by default displays the last 10 lines of the survey file. After viewing the data, the user can select the desired row for analysis by clicking once with the mouse on it, which redirects him to the prediction page.

The following test scenarios were successfully performed:

1. Check for displayed data: Checking that the last 10 rows of the query file are displayed, and if it is less than 10 rows, that all rows.
2. Apply filter: Verify that after applying a filter, the correct number of rows are displayed and that they match those in the survey file.
3. Select a row from the survey file: Verify that after selecting a row from the survey file, the user is redirected to the prediction page for the corresponding completion of the survey.
4. Correct message when no survey file is selected: Checking that an appropriate message is displayed when there is no survey file or it has incompatible content.

5.4.4 Success prediction

The most important part of the application I3SP is the success prediction module. The user accesses the forecast page after selecting a row from the analysis survey file that corresponds to a survey completion from a specific company. If no file line is selected, a field is displayed where the user can enter a line number. After selecting a line, results of the analysis for the selected company are displayed (Fig. 65).

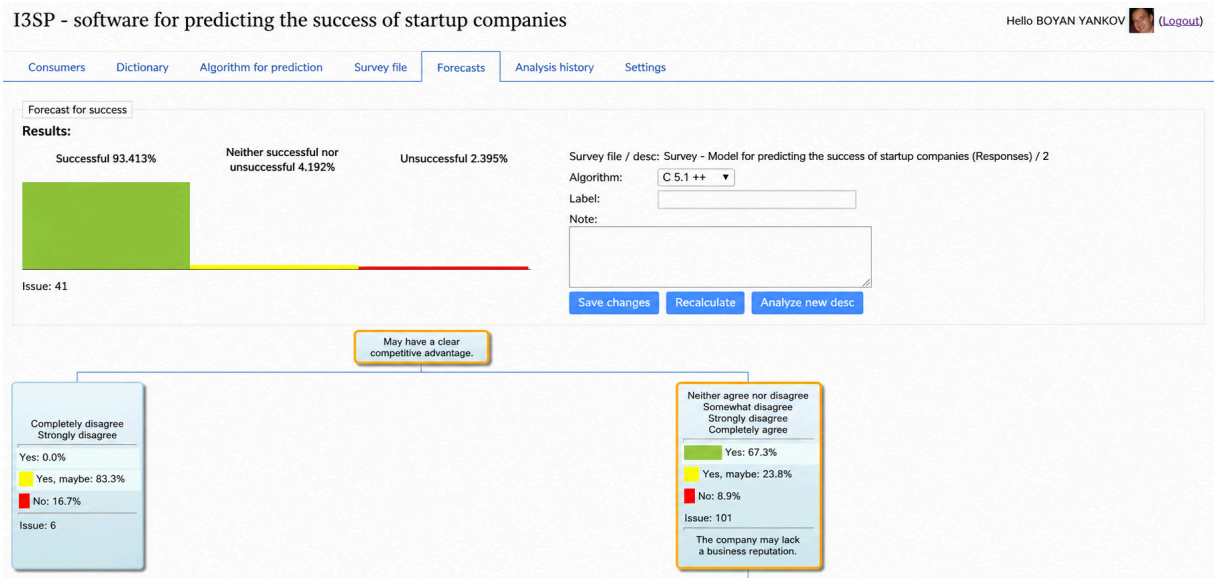


Fig. 65Excerpt from the results page of a company success forecast (created by the author)

The results page contains the company's probability of success displayed as a bar chart with numerical data attached to it. A customized classification tree for the specific company is also drawn. Each node of the tree contains information about the possible responses to the query by which it is reached, the probability of success in reaching that node, and the question of branching to subsequent nodes. The traversal path is shown in the tree by enclosing the active nodes and the terminal node in color. The page also contains a form to save forecast data.

When saving forecast data, a saved data page is displayed (Fig. 66), which can also be accessed from the application menu. Data on previous analyzes are presented in a table in which the most important fields are marked (analysis label, percentage probability of success of the company, survey file, order, calculation algorithm used, date). The probability of success is represented by a compact graph. At the end of each row there is a "details" button, when pressed, the page with the analysis for the company is loaded again.

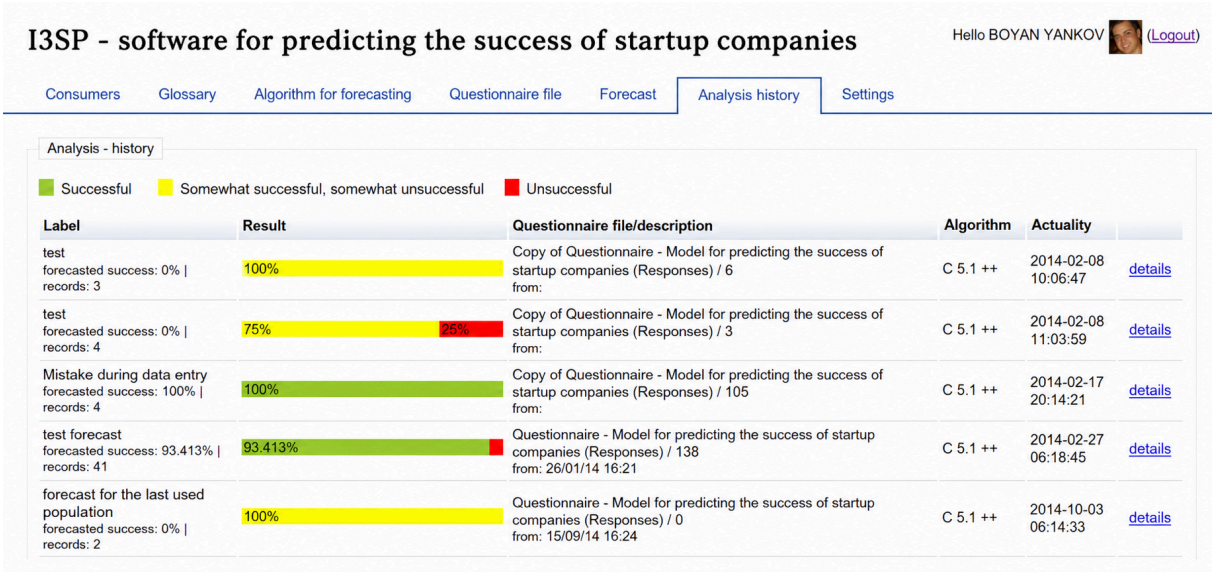


Fig. 66Saved analysis results (created by the author)

The following test scenarios were successfully executed:

1. Classification tree verification: Verification of matching of the classification tree drawn on the results page against the supplied text file with definition of the tree.
2. Verifying the interpretation of the survey data: Select a row from the data file and verify that the data and path in the tree match those of the selected row, and that the company's success prediction matches the one given at the end node of the tree.
3. Show the tree when the survey row is incomplete: In the absence of an answer to a question from the algorithm, only the answers before that question are colored in the tree, the probability of success up to that question is shown, and a message that the survey data is incomplete is displayed.
4. Analysis history: Verify that analysis data can be saved, and saved analyzes can be revisited on the analysis page.
5. Non-existing row entry: Verify that once entered non-existent row, an appropriate message is displayed.
6. Check for unselected or invalid survey file or unselected algorithm: Checking that if no survey file is selected, the survey file does not conform to the format, or no algorithm is selected, an appropriate message is displayed.

5.4.5 Administration Module

For users, possessing administrative rights, the application additionally provides the following administrative functionalities:

- Management of application users,
- Dictionary file settings,
- Uploading and setting up classification tree files derived from data mining software.

All three types of functionality are particularly important for the correct operation of the application and visible only to the administrator, who is supposed to have a deeper knowledge of working with the system.

The user settings page (Fig. 67) contains a list of all users in the system and allows to allow and terminate the access of a given user with the "Enable" and "Prohibit" buttons. When a user first visits the application,the application administrator is sent an email offering to approve the new user with a link to the user settings page. Upon activation of a user, an email is also sent to him, informing him that he can now use the system.



Fig. 67User Management (created by the author)

Working with I3SPaccumulates survey data that will be used for future improvements to the success prediction model through the use of data mining software. For this reason, I3SRincludes functionality for uploading new classification tree files to the system (Fig. 68). I3SP can work with more than one classification tree, allowing flexibility. It is possible to apply different trees to different data sets – for example, for technology companies and for all start-ups. The administrator chooses which algorithm to use by default, and it applies to all users of the application, unless they have explicitly chosen another.

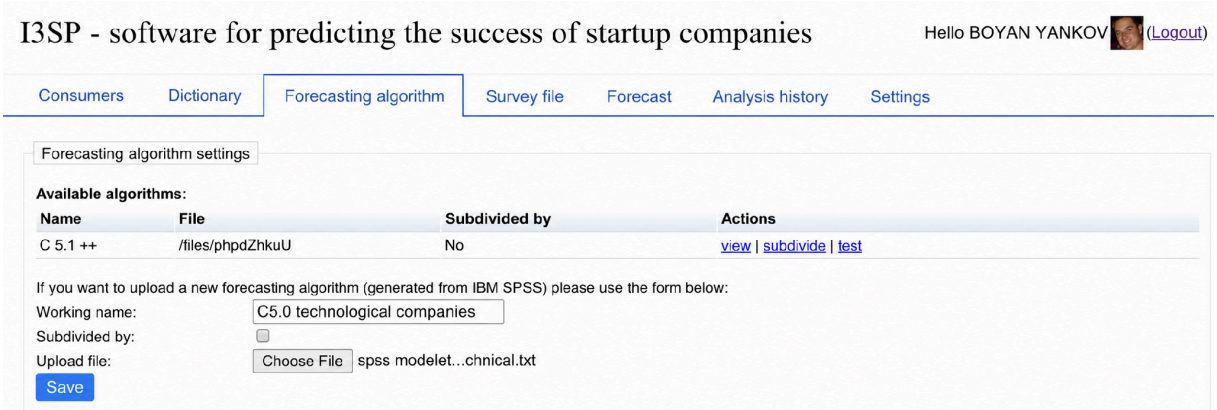


Fig. 68Uploading a new algorithm (classification tree) to the system (created by the author)

The administrative settings module also contains the possibility to upload and view a dictionary file for the application. The dictionary file is one per application and uploading a new one overwrites the existing one. It is available in the application from the beginning, but a change option is provided, which is necessary if the content of the survey that users complete changes.

The following test scenarios were successfully performed:

1. New user activation: Authentication of a new user in the system, activation of the user by an administrator, login with the already active user profile.
2. Change user status: Verify that after a user is disabled, they can no longer log in and their status is "Forbidden", and after reactivating the profile, their profile gets access rights again.
3. Upload prediction algorithm: Verify that a new prediction algorithm can be successfully uploaded.
4. Delete Algorithm: Verify that a prediction algorithm can be deleted.
5. Select Default Algorithm: Verify that only one algorithm can be selected by default by the administrator and that it is enabled for other users.
6. View Algorithm: Verify that an algorithm opens in a new window after clicking the button "review"
7. Dictionary file change: Checking that the dictionary file can be successfully updated by the administrator.

5.4.6 Results, Findings, and Conclusions

From the examined test scenarios for the various functionalities of I3SP it can be concluded that the software functions correctly and is ready for deployment in a working environment.

5.5 Evaluation and Validation of the Software Product

The software product I3SPwas tested by applying it to data of 6 companies and interviews with their owners. For each of the companies, data was collected in 2013 and again in 2015 to track its development and whether the prediction made was correct. The results of the software validation are shown in Table 15, where, at the request of the respondents, the names of the companies have been replaced with numbers, and the names of their representatives - with initials, in order not to violate the trade secret.

Table 15 Software product testing results I3SP

Company, representative, field of activity	**Predicted results obtained from the data in 2013 year**	**Actual result in 2015**
Company 1, P. P., medical equipment	Successful	Successful (increased turnover)

Company 2, I. M., web design	Neither successful nor unsuccessful	Successful (has increased the number of employees and turnover)
Company 3, L. M., bio agriculture	Successful	Successful (increased turnover)
Company 4, P. M., software development	Successful	Successful (increased the number of employees and turnover)
Company 5, V. V., internet trade	Neither successful nor unsuccessful	Neither successful nor unsuccessful (continues activity, but has not increased the number of employees and turnover)
Company 6, SC, education	Neither successful nor unsuccessful	Neither successful nor unsuccessful (continues activity, but has not increased the number of employees and turnover)

The results presented in the table **confirm the effectiveness of the software**, with the predicted results matching the real ones in 5 out of 6 cases. Only the result for Company 2 does not match the forecast. From the cases presented in the table, Company 2 and Company 6 will be examined in detail, taking into account the situation from the perspective of the owners of the companies.

According to **S. C., owner of Company 6, working in the field of education,** the state of the company at the present moment (2015) compared to the situation two years ago (2013) is unchanged – the same number of employees, approximately the same turnover and no visible growth. This coincides with the forecast prepared by I3SP based on the data in 2013, namely that the company will continue its activities but will not report growth.

According to the forecast model of I3SP, limiting factors **for the growth of Company 6were** the lack of partnerships with competitors combined with the lack of customer concentration in the market. Unfortunately, such partnerships have not been built in the last two years, but S. C considers the need to build more partnerships with other educational institutions.

The situation with Company 6 at the moment (2015) has changed - it no longer develops an innovative product, but participates in the market in parallel competition with other companies. The company has more intangible assets - intellectual property. The success factors, according to the model, in this case are different and the prediction is that Company 6 will increase its scale.

According to I. **M., the owner of Company 2 engaged in web design,** as of now (2015) the company has increased its scale compared to the situation two years ago (2013) – a significantly larger number of employees, greater turnover. This result does not coincide with the forecast prepared by I3SP based on the data in 2013, namely that the company will continue its activities but will not report growth.

According to the model, a **limiting factor for the growth of Company 2was** the lack of a clear competitive advantage. I. M. realizes this fact, in part with the help of the company's

success forecast prepared in 2013, and shares the following: "Over the past two years, the company has more clearly defined its competitive advantage." It is likely that the model for predicting success, by identifying the limiting factors for growth, has helped in the development of the company.

The situation with Company 2 at the moment has also changed with the new analysis prepared in 2015 I. M. indicates that the company has a clear competitive advantage. I3SP 's forecast is now for company growth.

In general, users expressed satisfaction with the obtained analyzes and showed interest in their interpretation and in applying the obtained knowledge.

5.6 Results, Findings, and Conclusions

The requirements for the software for predicting the success of start-up companies were discussed I3SP, the process of designing and developing the software, as well as testing, testing and validating the finished product, was described. The obtained results showed that the product is useful for the target group and in a large percentage of cases it gives correct predictions for the success of companies.

In I3SP the possibility of updating the model when accumulating more data is provided. Data from all users of the product is periodically (over several months) reanalyzed with data mining software and new success prediction models are generated (in text file format). This file is uploaded to I3SP and the software can now apply the new model. This allows for the selection of several models, as well as the creation of specialized models (for example, for technology start-ups).

A possible development of I3SP is the development of future versions to include a shorter survey with only the key questions from the success prediction models, as some users of the product expressed dissatisfaction that the survey took too long to complete (about 15 minutes). A disadvantage of this option would be that completing a shorter survey could not be used to develop the model.

Promotion of the software is forthcoming - a process that has already begun with the creation of a website for the product - http://byankov.com/prediction/, as well as publications in scientific journals, in the author's website http://byankov.com/ and in social networks. Involving more entrepreneurs in the project will allow further development of the success prediction models and the software product.

CONCLUSION

In this dissertation, the current state **in our country and in the world in the sources of** scientific information in the field of predicting the success of start-up companies was studied, researched and analyzed. A detailed literature review of the problem area and of the already known methods, approaches and models for solving the task of predicting the success of start-up companies was carried out. The success factors of start-up companies in the world and in Bulgaria were studied, researched and analyzed. Software products were reviewed and compared to predict their success.

A **company creation process model was proposed.** A theoretical model was created to predict **the success of start-up companies from Bulgaria.** The model was adapted for the environment by conducting qualitative research **with entrepreneurs** and owners of Bulgarian companies. The model contains the following main categories for determining the company's success: entrepreneurial team, business strategy, industry structure, resources. The proposed model was developed in detail with subcategories and success factors of start-up companies.

Based on the proposed model for predicting the success of start-up companies, an experiment **was** conducted, including quantitative research **of** start-up companies from Bulgaria. Through factor **analysis of** the quantitative research data, the categories and subcategories of the factors of the proposed success prediction model were validated. The data were analyzed with two software products for data **mining – IBM** SPSS Modeler and Weka. Different success **prediction models were** synthesized. After analyzing the models, those that offered the highest accuracy in predicting the company's success were selected. The main factor in the success of start-ups turned out to be the presence of a clear competitive advantage. Models specializing in predicting the success of technology **start-ups** have also been created. The best models obtained are of the classification tree type. In the **cross-validation analyses** performed, the synthesized models predicted with high accuracy the success scores of the companies from the analyzed test data set.

Based on the prepared models for predicting success, a prototype of information system **for predicting success of start-up companies** (I3SP) was implemented.

MAIN CONTRIBUTIONS OF THE DISSERTATION

Main contributions of the dissertation are described in Table 16.

Table 16Main contributions of the dissertation work

Contribution	Publications
1. Research, research and analysis of the state-of-the-art in our country and the world in libraries, scientific databases, research papers and the Internet of start-up success prediction models, start-up success factors, data mining software products and their application in classification models for success prediction.	• Yankov, B.: Overview of Success Prediction Models for New Ventures, International Conference Automatics and Informatics'12, ISSN 1313-1850, pp. 13-16 (2012)
2. Proposal of a theoretical model for start-up success prediction by Bulgaria, which has been adapted for the environment and for Bulgarian entrepreneurs through qualitative research.	• Yankov, B.: A Model for Predicting the Success of New Ventures, Vth International Scientific Conference e-Governance, ISSN 1313-8774, pp. 128-135 (2013) • Yankov, B.: Research of Success Factors for Start-up Companies, Doctoral Conference in Mathematics, Informatics and Education, MIE 2013, ISBN 978-954-07-3596-2, pp. 119-124 (2013)
3. The theoretical model proposed by the author has been validated through quantitative research and through software products for extracting knowledge from data. Models for predicting success, adapted for companies from Bulgaria, were created by applying quantitative methods, as well as a model for predicting the success of technological companies.	• Yankov, B., Haralampiev, K., Ruskov, P.: Start-up Companies Predictive Models Analysis, Vanguard Scientific Instruments in Management '2013 (VSIM:13), ISSN 1314-0582, pp. 275-285 (2013) • Yankov, B.: Synthesis of Predictive Models for Start-up Companies, International Conference Automatics and Informatics'13, ISSN 1313-1850, (2013) - accepted for publication • Yankov, B., Ruskov, P., Haralampiev, K.: Models and Tools for Technology Start-Up Companies Success Analysis, Journal Economic Alternatives 2014/3, ISSN 1312–7462, pp. 15-24 (2014) • Yankov, B.: Comparison of classification models for start-up companies. Vanguard Scientific

	Instruments in Management '2014 (VSIM:14), ISSN 1314-0582, (2014) - accepted for publication publication
4. A prototype of an information system for predicting the success of start-up companies was designed and implemented, which allows future upgrading and improving the algorithm based on the collected data.	● Yankov, B., Vitanov, N.: Information System for Forecasting the Success of Bulgarian Start-up Companies, Doctoral Conference in Mathematics, Informatics and Education, MIE 2014, ISBN 978-954-07-3759-1, pp. 95-102 (2014) ● Yankov, B., Ruskov, P., Haralampiev, K.: Models and Tools for Technology Start-Up Companies Success Analysis, Journal Economic Alternatives 2014/3, ISSN 1312–7462, pp. 15-24 (2014)

PUBLICATIONS RELATED TO THE DISSERTATION TOPIC

The author has the following publications on the topic of the dissertation:

1. Yankov, B.: Overview of Success Prediction Models for New Ventures, International Conference Automatics and Informatics‘12, ISSN 1313-1850, pp. 13-16 (2012)
2. Yankov, B.: A Model for Predicting the Success of New Ventures, Vth International Scientific Conference e-Governance, ISSN 1313-8774, pp. 128-135 (2013)
3. Yankov, B.: Research of Success Factors for Start-up Companies, Doctoral Conference in Mathematics, Informatics and Education, MIE 2013, ISBN 978-954-07-3596-2, pp. 119-124 (2013)
4. Yankov, B., Haralampiev, K., Ruskov, P.: Start-up Companies Predictive Models Analysis, Vanguard Scientific Instruments in Management ‘2013 (VSIM:13), ISSN 1314-0582, pp. 275-285 (2013)
5. Yankov, B.: Synthesis of Predictive Models for Start-up Companies, International Conference Automatics and Informatics'13, ISSN 1313-1850, (2013) - accepted for publication
6. Yankov, B., Ruskov, P., Haralampiev, K.: Models and Tools for Technology Start-Up Companies Success Analysis, Journal Economic Alternatives 2014/3, ISSN 1312–7462, pp. 15-24 (2014)
7. Yankov, B., Vitanov, N.: Information System for Forecasting the Success of Bulgarian Start-up Companies, Doctoral Conference in Mathematics, Informatics and Education, MIE 2014, ISBN 978-954-07-3759-1, pp. 95-102 (2014)
8. Yankov, B.: Comparison of classification models for start-up companies. Vanguard Scientific Instruments in Management '2014 (VSIM:14), ISSN 1314-0582, (2014) - accepted for publication publication

REFERENCES

Abbas, A. (2008). *An Assessment Methodology for Predicting the Success of Technological Enterprises.*

Achour, M. B. (13 11 2014). PHP *Manual. Retrieved* from PHP: http://php.net/manual/en/

AppSplit - Crowdsourcing for Apps. (July 14, 2014). Retrieved from http://appsplit.com

Astebro, T. (2000). Key Success Factors for Technological Entrepreneurs' R&D Projects. *IEEE Transactions on Engineering Management, Vol. 51, No. 3,* pp. 314-321.

Avramov, J. (2012). Some Problems in the Operation of the Management of Mis-Funding for Modernization and Innovation for Industry Under OP "Competitiveness". *IV International Science Conference "E-governance",* (pp. 227-238).

Bailetti, T. (2012). Technology Entrepreneurship: Overview, Definition, and Distinctive Aspects. *Technology Innovation Management Review*.

Baron, R. A. (2005). *Entrepreneurship: A process Perspective, First Edition.* Mason. OH: Thomas corp.

Biggadike, R. C. (1976). Corporate diversification: entry, strategy, and performance. *Cambridge, MA: Harvard, University Press*.

Blank, S. (2005). *The Four Steps to the EpiphanySteve Blank.* K&S Ranch Press.

Blank, S. (22 11 2014). Retrieved from Steve Blank - Entrepreneurship and Conservation: http://steveblank.com/

Blueprint New Business start-up Tester. (July 15, 2014). Retrieved from Blueprint Business Plans and Consultancy Services: http://q1.prototype.thinkblueprint.com.au/new-business-success-calculator/

Bouckaert, R. F. (2013). *WEKA Manual for Version 3-6-10.* Hamilton, New Zealand : University of Waikato.

Briggs, I. M. (1995). *Gifts Differing: Understanding Personality Type.* Davies-Black Publishing.

Brockhaus, R. (1982). The Psychology of the Entrepreneur. By D. S. Kent, Encyclopedia *of Entrepreneurship In C. A.* (pp. 39-56).

Brüderl, J. P. (April 1992). Survival Chances of Newly Founded Organizations. *American Sociological Review, Vol. 57, No. 2*, pp. 227–242.

Carland, J. C. (1996). The Theoretical Bases and Dimensionality of the Carland Entrepreneurship Index. *Proceedings of the RISE'96 Conference* (pp. 1-24). University of Jyväskylä, Finland.

Carland, J. C. (Oct 2000). A new venture creation model. Journal *of Business and Entrepreneurship;,* page 29.

Chandler, G. &. (1994). Resource-based capabilities, venture strategies, and venture performance. *Journal of Business Venturing,* pp. 331–349.

Chandler, G. H. (1994). Resource-based capabilities, venture strategies, and venture performance. *Journal of Business Venturing, 9,* pp. 331-349.

Chapman P., C. J. (2000). *CRISP-DM 1.0 - Step-by-step data mining guide.*

Chrisman, J. B. (1998). The Determinants of New Venture Performance: An Extended Model. *Entrepreneurship, theory and practice : ET & P Vol. 23.,* pp. 5-30.

Compass - Benchmark Your Business Metrics From 30+ Data Sources. (July 15, 2014). Retrieved from https://www.compass.co

Davis, A. M. (2005). The Rise and Fall of a Software start-up. *Journal of Information Technology Case and Application Research, Vol. 7, No. 2,* pp. 31-48.

Ebben, J. J. (2005). Efficiency, flexibility, or both? Evidence linking strategy to performance in small firms. *Strategic Management Journal, 26*, pp. 1249–1259.

Elsevier. (November 20, 2014). Retrieved from ScienceDirect.com - a leading full-text scientific database: http://www.sciencedirect.com/

Emerald. (November 20, 2014). Retrieved from Emerald Insight: http://emeraldinsight.com/

Ensley, M. (1999). *Entrepreneurial teams as determinants of new venture performance.* New York: Garland Publishing Inc.

EquityNet - The Leading Equity Crowdfunding Platform. (July 14, 2014). Retrieved from https://www.equitynet.com/

European Commission. (2014). HORIZON 2020, The New EU Framework Programme for Research and Innovation 2014-2020.

Eusostat, European Commission. (2009). *Business Demography: employment and survival.*

Florida, U. o. (9 November 2007). National Colligate Invention and Innovators Alliance (NCIIA) Participants Guide. Invention *to Venture Workshop in Technology Entrepreneurship Workshop.* NCIIA Publishing.

Fundly - Online Fundraising Websites to Raise Money for Anything. (July 14, 2014). Retrieved from https://fundly.com

Gartner, W. B. (1988). "Who is an Entrepreneur?" Is the Wrong Question. *American Journal of Small Business, Spring, 12(4)*, pp. 11-32.

GiveForward - The Number 1 Way to Raise Money for a Loved One. (July 14, 2014). Retrieved from http://www.giveforward.com/

GoFundMe - Number One for Crowdfunding and Fundraising Websites. (July 14, 2014). Retrieved from http://www.gofundme.com/

Google Inc. (November 20, 2014). Retrieved from Google Scholar: http://scholar.google.bg/

Google Inc. (12 08 2014). Google *APIs Explorer.* Retrieved from Google Developers: https://developers.google.com/apis-explorer/#p/

Google Inc. (12 08 2014). Google *Forms.* Retrieved from Google Docs: http://www.google.com/google-d-s/createforms.html

Google Inc. (November 15, 2014). *Google Sheets API version 3.0*. Retrieved from Google Developers: https://developers.google.com/google-apps/spreadsheets/

Google Inc. (12 08 2014). Google *Spreadsheets API.* Retrieved from Google Developers: https://developers.google.com/google-apps/spreadsheets/

Google Inc. (November 15, 2014). *Google Visualization API Reference*. Retrieved from Google Developers: https://developers.google.com/chart/interactive/docs/reference

Google Inc. (n.d.). Google *Trends. Downloaded* on 25 05 2014 from http://www.google.bg/trends/explore#q=entrepreneurship%2C%20%2Fm%2F02nwq&cmpt=q

Grabenwarter, U. (2005). *Exposed to the J-curve: Understanding and Managing Private Equity Fund Investments.* Euromoney Investors Plc.

Gruber, M. (2004). Marketing in new ventures: theory and empirical evidence. *Schmalenbach Business Review, Vol. 56 no. 2*, pp. 164-99.

Hall, M. F. (Volume 11, Issue 1 2009). The WEKA Data Mining Software: An Update. SIGKDD *Explorations.*

HP Vertica. (11 01 2015). Retrieved from Real-Time Analytics Platform: http://www.vertica.com/

IBM. (2010). *Predictive Modeling with IBM SPSS Modeler - Student Guide* . IBM Corp.

IBM. (November 22, 2014). *IBM SPSS Statistics*. Retrieved from http://www-01.ibm.com/software/analytics/spss/products/statistics/

IBM Corporation. (2012). *IBM SPSS Modeler 15 User's Guide.* IBM Corporation.

IBM SPSS Statistics. (01 11 2014). Retrieved from IBM: http://www-01.ibm.com/software/analytics/spss/products/statistics/

Indiegogo - Crowdfunding for What Matters to You. (July 14, 2014). Retrieved from https://www.indiegogo.com/

Jelinek, M. (1996). International Journal of Technology Management. 799-813.

Jung, C. G. (1971). Psychological *Types - Collected Works of C. G. Jung, Volume 6.* Princeton University Press.

Kaiser, H. R. (1974). Little jiffy, mark iv. *Educational and Psychological Measurement*, 111-117.

Keeley, R. R. (1989). Determinants of new venture success before 1982 and after a preliminary look at two eras. *IESE DI-173-E*.

KickStarter - The World's Largest Funding Platform for Creative Projects. (July 14, 2014). Retrieved from https://www.kickstarter.com

Lawrence, P. S. (1998). *The PageRank Citation Ranking: Bringing Order to the Web.*

Learning, P. f. (1993). *Quinlan, J.* San Mateo CA: Morgan Kaufmann.

Likert, R. (1932). *A technique for the measurement of attitudes.* New York: The Science Press.

Luo, Y. (1999). Environment-strategy-performance relations in small businesses in China: A case of township and village enterprises in Southern China. *Journal of Small Business Management*, pp. 37–52.

Lussier, R. N. (1995). There Are Few Differences Between Successful and Failed Small Businesses. *Journal of Small Business Strategy, Vol. 6, No. 1*.

Manev, I. M. (6 2012). Social Capital and Strategy Effectiveness: An Empirical Study of Entrepreneurial Ventures in a Transition Economy. Competitive *Strategies and Tactics,* pp. 57-70.

McDougall. (1987). An Analysis of Strategy, Entry Barriers, and Origin as Factors Explaining New Venture Performance. *Unpublished Doctoral Dissertation*. University of South Carolina.

McGill, U. (2014). *McGill University*. Retrieved from McGill University: http://www.mcgill.ca/

Microsoft. (November 13, 2014). ASP. *NET.* Retrieved from http://www.asp.net/

Microsoft. (January 13, 2015). *Explore SQL Server 2012-2014*. Retrieved from http://www.microsoft.com/en-us/server-cloud/products/sql-server/

Microsoft Excel - Spreadsheet Software. (01 11 2014). Retrieved from Microsoft Office: http://office.microsoft.com/bg-bg/excel/

Moore, G. A. (1991). Crossing *the chasm: Marketing and selling technology products to mainstream customers.* New York, N. Y.: HarperBusiness.

NCIIA. (2007). *National Colligate Invention and Innovators Alliance (NCIIA) Participants Guide.* University of Central Florida, NCIIA Publishing .

Nelson, R. (1999). *The Handy History Answer Book.* Visible Ink Press.

Nicholas, S. A. (2003). Engineering Entrepreneurship: does entrepreneurship have a role in engineering education? *Antennas and Propagation Magazine, IEEE*.

OAuth community. (November 15, 2014). Retrieved from oAuth Community Site: http://oauth.net/

Oracle Corp. (12 11 2014). Retrieved from Java: https://www.java.com/en/

Park, M. L. (2000). A New Venture Performance Model in the Korean Information and Telecommunications Industry. *ETRI Journal, Volume 22, Number 4*, 51-65.

Peng, M. L. (2000). Managerial ties and firm performance in a transition economy: The nature of a micro-macro link. *Academy of Management Journal, 43 (3)*, pp. 486–501.

Peng, M. W. (2001). How entrepreneurs create wealth in transition economies. *Academy of Management Executive 15 (1),* pp. 95-108.

Poindexter, J. B. (1975). *The Efficiency of financial Markets: The Venture Capital Case.* New York University, Graduate School of Business Administration.

ProQuest. (November 20, 2014). *UMI Dissertation Express.* Retrieved from http://dissexpress.umi.com/dxweb/search.html

Ries, E. (2011). The *Lean start-up.* Crown Business.

Ries, E. (2011). The *Lean start-up: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses.* Crown Business Publishing.

Ries, E. (22 11 2014). Retrieved from The Lean start-up - The Movement That is Transforming How New Products Are Built and Launched: http://theleanstartup.com/

Robertson, M. C. (2003). Barriers to start-up and their effect on aspiring entrepreneurs. *Education + Training, Vol. 45 Iss: 6*, pp. 308–316.

RocketHub - Crowdfunding. (14 07 2914). Retrieved from http://www.rockethub.com/

RuleQuest. (December 11, 2014). *Data Mining Tools See5 and C5.0*. Retrieved from RuleQuest Research Data Mining Tools: http://www.rulequest.com/see5-info.html

Ruskov, P. H. (2012). Online Investigation of SMEs Competitive Advantage. *MEB 2012, 10th International Conference on Management, Enterprise and Benchmarking*, (pp. 143-159).

Sandberg, W. R. (1986). *New venture performance: The role of strategy and industry structure.* Lexington Books.

Sell a Band - Where Fans Invest in Music. (July 14, 2014). Retrieved from https://www.sellaband.com/

Shepherd, D. (May 1999). Venture Capitalists' Assessment of New Venture Survival. *Management Science, Vol. 45, No. 5*, pp. 621-632.

Shepherd, D. E. (September–November 2000). New venture strategy and profitability: A venture capitalist's assessment. *Journal of Business Venturing, Volume 15, Issues 5–6*, pp. 449–467.

Shirokova, G. S. (2010). Factors of new venture performance in Russia. *Management, Research Review, Vol. 33 Iss: 5*, pp. 484–498.

Small Business Services (SBS), Small Firms: Big Business! (2002). London.

start-up Risk Calculator. (July 15, 2014). Retrieved from EquityNet - The Leading Equity Crowdfunding Platform: https://www.equitynet.com/crowdfunding-tools/start-up-risk-calculator.aspx

Surveys, W. T.-W. (July 18, 2014). *Usage Statistics and Market Share for Operating Systems for Websites, July 2014*. Retrieved from http://w3techs.com/technologies/overview/operating_system/all

Teal, E. a. (2003). The determinants of new venture success: strategy, industry structure, and the founding entrepreneurial team. *The Journal of Private Equity, Vol. 6 No. 4*, 38-51.

Teal, E. J. (1998). The determinants of new venture success: Strategy, industry structure, and the founding entrepreneurial team. *Unpublished doctoral dissertation*. Athens, Greece: The University of Georgia.

The R Foundation. (11 01 2015). Retrieved from The R Project for Statistical Computing: http://www.r-project.org/

Tranfield, D. D. (2003). Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *British Journal of Management 14(3),* pp. 207-222.

Tyebjee, T. B. (1984). A Model of Venture Capitalist Investment Activity. *Management Science, Vol. 30, No. 9*, pp. 1051-1066.

Waikato, M. L. (11 09 2014). Weka *3: Data Mining Software in Java.* Retrieved from Weka: http://www.cs.waikato.ac.nz/ml/weka/index.html

Wells, W. A. (1974). *Venture Capital Decision Making, unpublished doctoral dissertation.* Carnegie-Mellon University.

Yankov, B. (2012). Overview of Success Prediction Models for New Ventures. *International Conference Automatics and Informatics'12, ISSN 1313-1850*, (pp. 13-16). Sofia.

Yankov, B. (2013). A Model for Predicting the Success of New Ventures. *Vth International Scientific Conference e-Governance, ISSN 1313-8774,* , (pp. 128-135). Sozopol.

BAS, et al. (2012). *Dictionary of the Bulgarian language. T. I–XIV.* Sofia.

National Statistical Institute - Republic of Bulgaria. (2014). Retrieved from National Statistical Institute - Republic of Bulgaria: http://www.nsi.bg/bg

NSI, National Statistical Institute - Republic of Bulgaria. (2013). Demographics *of enterprises as of 31.12.2012.*

Stoicheva, M. E. (2013). Practical training in the field of entrepreneurship. *VII National Conference Education in the Information Society*, (pp. 207-216). Plovdiv.

Kharalampiev, K. (2012). *IBM SPSS - Statistical solutions to applied research problems. (Second revised and enlarged edition).*

DECLARATION OF ORIGINALITY

I declare that the dissertation on the topic: "Methodology for predicting the success for technological start-up companies in Bulgaria" is my work and in its development no foreign publications and developments were used in violation of their copyrights.

Citations of all sources of information, text, illustrations, tables, images and others are indicated according to standards.

The results and contributions of the dissertation research conducted are original and not borrowed from research and publications in which I have no part.

APPENDICES

APPENDIX 1: GLOSSARY

Term / Abbreviation	Definition
AJAX (Asynchronous JavaScript And XML)	A group of techniques in web programming that allow the creation of asynchronous web applications using the JavaScript and XML languages.
BI (Business Intelligence)	A set of technologies and tools to transform unstructured data into meaningful and useful information for business analysis. The purpose of BI is to enable easy interpretation of large volumes of data, thanks to which new opportunities and business strategies can be identified and created.
CLR (Common Language Runtime)	Virtual machine, a component of the Microsoft . NET Framework that manages the execution of . NET programs.
CLI (Common Language Infrastructure)	A specification developed by Microsoft that describes the executable code and environment and is implemented in Microsoft . NET Framework.
CRISP-DM (Cross Industry Standard Process for Data Mining)	A methodology that describes a standard of the data mining process and was created by a consortium of companies as a project of the European Union in 1997. CRISP-DM is described as a hierarchical process model containing groups of tasks described at four levels of abstraction (from general to particular): phases, general tasks, specialized tasks, and process instances.
CSV (Comma-separated Values)	Translated as "comma-separated values", a file format for storing tabular data in the form of text defined in the RFC4180 recommendation.
CSS (Cascading Style Sheets)	Language for a stylesheet specified and maintained by the W3C that describes the presentation of a document written in a markup language (most commonly HTML, xHTML, XML).
DB (database)	Database - a structured set of data stored on a computer
DBMS (database management system)	A database management system and software application that retrieves and analyzes data from a database by interacting with the user and other applications.
DM (Data Mining)	Translated as "data mining", the process of discovering meaningful correlations, dependencies, repeating patterns, trends and anomalies in large data sets, by using techniques and algorithms from the fields of machine learning, pattern recognition, statistics, neural networks, and data visualization.

DOM (Document Object Model)	An object model used to represent XML-based documents, also used in HTML and xHTML.
GPL (GNU General Public License)	The GNU General Public License is a license issued by the Free Software Foundation, in order for it to be used to license software as "free". The GPL guarantees users of computer programs the following rights: ● Freedom to use the program for any purpose, ● Freedom to study how the program works and change it, ● Freedom to distribute copies, ● Freedom to improve the program and give the public access until improvements.
HTML (HyperText Markup Language)	Translated as "hypertext markup language", the main markup language for describing and designing web pages, standardized and supported by the W3C.
HTTP (HyperText Transfer Protocol)	Translated as "hypertext transfer protocol", a network protocol, from the application layer of the OSI model, for transferring information in computer networks. The latest version of the protocol is described in RFC 2616.
IBM SPSS Modeler	A software platform for extracting knowledge from data and creating success prediction models supporting decision-making by individuals, groups, systems and companies.
I3SP (Information System for Start-ups Success Prediction)	In translation "information system for predicting the success of start-up companies", software an information system developed for this dissertation that serves to predict the probability of success of start-up companies and visualize a classification tree with the success factors.
Java	Object-oriented programming language. Code written in Java is compiled into language-specific bytecode. Programs written in Java run in a virtual machine (Java Virtual Machine).
JavaScript	A programming language that is interpreted and executed by most web browsers. Supports object-oriented and procedural programming style.
JSON (JavaScript Object Notatio)	A text-based open standard designed to exchange data objects consisting of key-value pairs. JSON is derived from JavaScript but is language independent.
JSP (Java Server Pages)	Translated as JavaServer Pages, a web server application development technology using the Java programming language.

Microsoft . NET Framework	An application development platform created by Microsoft that provides a programming model, class library, and programming environment. code.
MVC (Model-View-Controller)	Translated as model-view-controller, a software architecture design pattern used in the implementation of user interfaces. MVC divides the software application into three interconnected parts, separating the internal representation of information from the information presented to the user.
NCIIA (The National Collegiate Inventors and Innovators Alliance)	Translated from "National University Association of Inventors and Innovators", a US organization that trains entrepreneurs and funds student companies.
NVP (New Venture Performance)	Translated as "start-up company productivity", the end result of the process of creating and organizing a new business that develops, creates and markets products and services to meet market needs for profit and growth.
oAuth	An open protocol that enables secure authorization for web, mobile, and desktop applications. The protocol provides a standard framework by which third-party applications can gain limited access to the HTTP service.
OS (Operating System)	Translated as "operating system (OS)", a basic part of computer system software that manages and coordinates the operation of the processor and devices in the computer system, as well as serves the operation of application software.
OSI (Open Systems Interconnection) model	Translated as "open model for connecting systems", a theoretical model describing the principle of communication and the construction of computer networks. Seven layers are used as the main building block, and each layer provides an interface and services to the upper layer, while at the same time receiving services from the layer below it.
PERL	A general-purpose, cross-platform scripting programming language distributed under the GNU license.
RIA (Rich Internet Application)	Translated as "internet applications with rich functionality", web applications that have many of the characteristics of desktop applications.
SME (Small and medium-sized enterprises)	Translated as "small and medium-sized companies". SMEsare companies whose number of employees falls within certain limits (up to 250 employees for the European Union), defining them as micro, small and medium-sized companies.

W3C (World Wide Web Consortium)	Translated as "world Internet consortium", the main global organization standardizing the Internet, also known as world wide web, www and w3.
Weka	A set of machine learning algorithms that serve to solve problems related to extracting knowledge from data. Weka is written in the programming language Java and algorithms can be called directly or from Java code.
xHTML (eXtensible HyperText Markup Language)	Translated as "extended hypertext markup language", a language in the family of XML-based languages that inherits HTML.
XML (eXtensible Markup Language)	Translated as "extended markup language", a language defining rules for creating specialized markup languages, as well as the syntax to which these languages must obey. XMLspecifies how to structure a document by tagging its content, making it a metalanguage.

APPENDIX 2: RESULTS OF ANALYSIS WITH DATA MINING SOFTWARE APPLICATIONS

Results of factor analysis with the software product IBM SPSS Statistics

The output data of the factor analysis prepared by IBM SPSS Statistics are presented in detail.

Summary of the factor analysis:

Notes		
Output Created		15-Nov-2014 11:23:36
Comments		
Input	Data	survey data spss 12 fa.sav
	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	142
Missing Value Handling	Definition of Missing	MISSING=EXCLUDE: User-defined missing values are treated as missing.
	Cases Used	MEAN SUBSTITUTION: For each variable used, missing values are replaced with the variable mean.
Syntax		FACTOR /VARIABLES Lichn_Mogadasespravatamvrabotasib Personal_Itrusteveryone Personal_I am enterprising Personal_I am ambitious Personal_I am risk taking Experience_I have entrepreneurship school Experience_Imamumeniyapomanagementment Experience_Imchangesinmarketing Experience_ImchangesinhumanresourcesH Experience_There are changes in investment Attempt_I have a previous start-up attempt Experience_There is a question similar to the position Experience_Imamopitvshatshatsferanaindus Experience_I have managerial experience Team_Companyhascomplete Team_Teamhasknowledge Team_Team has skills Team_Team has positive Plan_Planswide Plan_I plan deeply Plan_Make formal written strategies Plan_I plan often Plan_Planirampofunctionalarea Plan_Preplanninglooksmmno GoalsThe goals I set for the business The target market of the company is Strat_I manage to stick to the camper Strat_Strategythatwehavechosen Strat_MarkedStrategyEpodr Strat_I have a meaty strategy for the business The business is entering the market next Competitive advantageMybusinessp Strategic PartnershipsThe Company Customer_Partner Supplier_Partner Partner_Other

	Industry_MoreProfit Industry_IndustryGDP Industry_IndustryDemand Industry_IndustryGrowth Competition_Competitorsintheindustrywith Competition_Company is exposed to competition MActive_Inventories MActive_Real Estate MActive_Money NMActive_recognizebrand NMActive_softwaresystemisau NMActive_builtbusinessreput NMActive_Licenses NMActive_builtbusinessaccounts NMActive_channelsdistributed NMActive_marketlabor NMActive_capital markets NMActive_materialsdelivery /MISSING MEANSUB /ANALYSIS Personal_I can handle my job Personal_I am self-confident Personal_I am enterprising Personal_I am ambitious Personally_Averse to taking risks Experience_I have entrepreneurship school Experience_Imumumeniapomanagement Experience_Imumumeniapomarketing Experience_There are changes in human resourcesH Experience_There are changes in investment Attempt_I have a previous start-up attempt Experience_There is a question similar to the position Experience_Imamopitvshatshatsferanaindus Experience_I have management experience Team_The company has complete Team_TeamhasKnowledge Team_TeamHasSkill Team_Team has positive Plan_I plan wide Plan_I plan deep Plan_MakeFormalWritingStrategies Plan_I plan often Plan_Planirampofunctionalarea Plan_Preplanninglooksmmno GoalsThe goals I set for the business The target market of the company is Strat_I manage to stick to the camper Strat_Strategythatwehavechosen Strat_MarkedStrategyEpodr Strat_I have a meaty strategy for the business The business is entering the market next Competitive advantageMybusinessp Strategic PartnershipsThe Company Customer_Partner Supplier_Partner Partner_Other Industry_MoreProfit Industry_IndustryGDP Industry_IndustryDemand Industry_IndustryGrowth Competition_IndustryCompetitors Competition_Company is exposed to competition Inventory_asset Real Estate_asset MActive_pairnumber NMActive_recognitionbrand NMActive_softwaresystemsau

		NMActive_builtbusinessreput NMActive_licenses NMActive_builtbusinessaccounts NMActive_channelsdistribute NMActive_marketlabor NMActive_capital markets NMActive_materialssupply /PRINT INITIAL KMO AIC EXTRACTION ROTATION /FORMAT BLANK(.5) /PLOT EIGEN /CRITERIA FACTORS(11) ITERATE(25) /EXTRACTION PC /CRITERIA ITERATE(25) /ROTATION VARIMAX /METHOD=CORRELATION.
Resources	Processor Time	00 00:00:03.500
	Elapsed Time	00 00:00:02.782
	Maximum Memory Required	328872 (321.164K) bytes

List of variables excluded from the factor analysis (full labels of the variables) are written:

- Personality and values [What happens to me in life depends on me.]
- Personality and values [I don't worry about making decisions without the necessary information.]
- Skills and Experience [I have technical skills (for technology related to my business).]
- What education do you have graduates?
- In what field did you graduate?
- What is your age group?
- Are your parents entrepreneurs / business owners?
- Planning [I often do competitive analysis.]
- The target annual profit of the company is (in BGN)
- How to enter market
- The product I offer has been accepted by the market to the following extent:
- My business has the following main competitive advantage:
- Strategic Partnerships [The firm engages in partnerships with the government.]
- Characteristics of the industry in which the firm operates [There are barriers to entry for new firms in the industry.]
- Characteristics of the industry in which the firm operates [There are barriers (barriers) to exiting the business.]
- Characteristics of the industry in which the firm operates [There are many bankruptcies in the industry.]
- Characteristics of the industry in which the firm operates [The industry sector in which the firm operates is large compared to other sectors.]
- Characteristics of the industry in which the company operates [There is a crisis in the industry.]
- Characteristics of the industry in which the company operates [The industry is predictable.]
- I believe the following factors are key to success for my business:Innovation
- I consider the following factors key to success for my business:Marketing and Advertising
- I consider the following factors key to success for my business:Team
- I believe the following factors are key to success for my business: Knowledge
- I believe the following factors are key to success for my business: Business Adaptability
- I believe the following factors are key to success for my business: Strategy and Planning
- I consider the following factors to be key to success for my business: External environment
- I consider the following factors to be key to success for my business: Customer service
- I believe the following factors are key to success for my business: Price
- The company operates in the following industry sector:
- The industry in which the company operates is in a stage of:
- Competitiveness of the industry [There is strong competition in the industry.]
- Competitiveness of the industry [Small firms are also present in the industry.]
- Characteristics of customers [There is a concentration of customers in the industry.]
- Customer characteristics [Customers are different types.]

- Tangible assets [The company owns equipment / machinery.]
- Intangible assets [The company owns patents / trademarks.]
- Intangible assets [The company owns its own developed products.]
- Founders' own funds - The main source used for initial funding of the company is:
- Funds from family / friends - The main source used for initial funding of the company is:
- Scholarship / small business fund - The main source used for initial funding of the company is:
- Loan / credit - The main source used for seed funding of the firm is:
- Business incubator - The main source used for seed funding of the firm is:
- Individual investor (business angel) - The main source used for seed funding of the firm is:
- Venture capital (VC) investment fund - The main source used for seed funding of the firm is:
- Barter - The main source used for initial financing of the company is:
- Partnership - The main source used for initial financing of the company is:
- Individual investor (business angel) - The main source used for initial financing of the company is:

Results of the factor analysis:

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.697
Bartlett's Test of Sphericity	Approx. Chi-Square	3154.260
	df	1431
	Sig.	.000

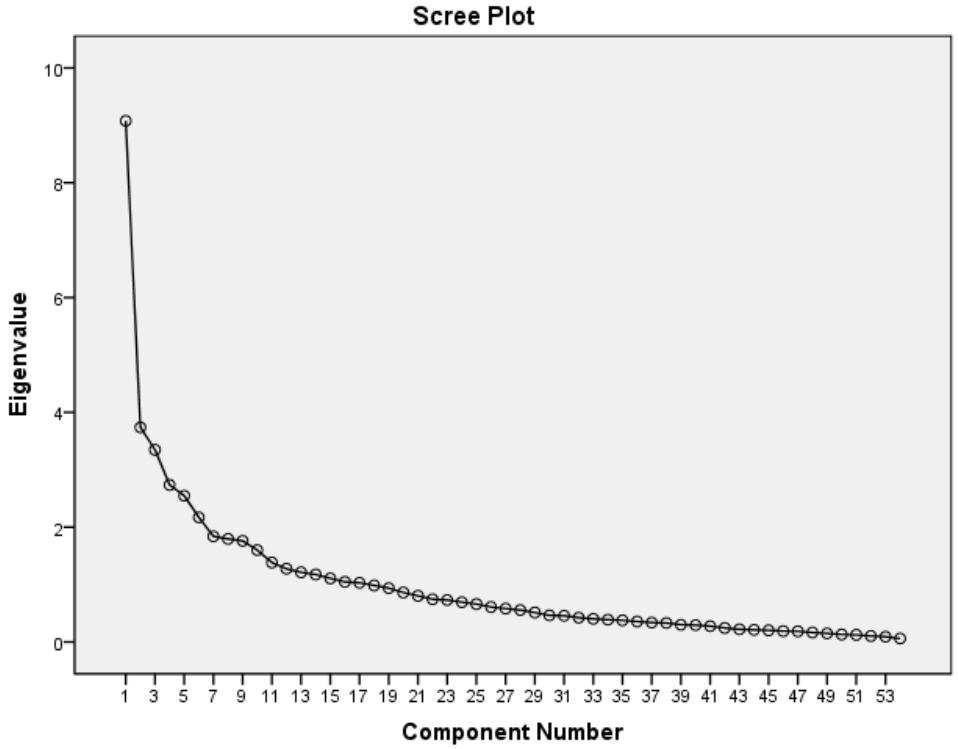
Communalities		
	Initial	Extraction
Personality and values [I can manage my work without the need for assistance.]	1.000	.237
Personality and values [I am self-confident.]	1.000	.649
Personality and values [I am enterprising, initiative.]	1.000	.699
Personality and values [I I am ambitious.]	1.000	.645
Personality and values [I am a risk taker.]	1.000	.569
Skills and experience [I have entrepreneurial skills (business start-up skills).]	1.000	.657
Skills and experience [I have skills in management.]	1.000	.733
Skills and experience [I have marketing skills.]	1.000	.743
Skills and experience [I have human resources (HR) skills.]	1.000	.677
Skills and experience [I have investment skills.]	1.000	.612
Skills and experience [I have previous experience in starting a business.]	1.000	.659
Skills and experience [I have experience in a similar position (to the one currently held).]	1.000	.759
Skills and experience [I have experience in the same industry.]	1.000	.663
Skills and experience [I have managerial experience.]	1.000	.656
Teamwork [The company has a complete team.]	1.000	.650
Teamwork [The team has the necessary knowledge.]	1.000	.853
Teamwork [The team has the necessary skills.]	1.000	.830
Teamwork [The team has the necessary positive attitude.]	1.000	.699
Planning [I plan covering a wide area.]	1.000	.492
Planning [I plan thoroughly.]	1.000	.632
Planning [I make a formal (written) strategic plan.]	1.000	.546
Planning [I plan often.]	1.000	.481
Planning [I plan by functional areas / departments.]	1.000	.506
Planning [I consider many alternatives when planning.]	1.000	.513
Goals [The goals I set for my business are ambitious.]	1.000	.475
The target market share of the company is:	1.000	.281
Strategic direction [I manage to stick to the strategy I initially chose.]	1.000	.492
Strategic direction [The strategy I have chosen is aggressive.]	1.000	.521

Strategic direction [The strategy I have chosen is detailed.]	1.000	.718
Strategic direction [I have a clear strategy for my business.]	1.000	.671
My business enters the market at the following stage:	1.000	.546
Competitive advantage [My business has a clear competitive advantage.]	1.000	.547
Strategic partnerships [The firm engages in partnerships with competitors.]	1.000	.386
Strategic partnerships [The firm engages in partnerships with customers.]	1.000	.675
Strategic partnerships [The firm engages in partnerships with suppliers.]	1.000	.652
Strategic partnerships [The firm engages in partnerships with other stakeholders.]	1.000	.608
Characteristics of the industry in which the company operates [Most companies in the industry have a good profit.]	1.000	.564
Characteristics of the industry in which the firm operates [The industry in which the firm operates generates a large portion of gross domestic product (GDP).]	1.000	.489
Characteristics of the industry in which the firm operates [The industry has stable demand.]	1.000	.572
Characteristics of the industry in which the firm operates [The industry is in growth period.]	1.000	.606
Industry competitiveness [Competitors in the industry are aggressive.]	1.000	.750
Industry competitiveness [The company is exposed to competitive attacks.]	1.000	.661
Tangible assets [The company owns inventory.]	1.000	.616
Tangible assets [The company owns real estate.]	1.000	.542
Tangible assets [The company owns cash.]	1.000	.409
Intangible assets [The company owns a recognizable brand.]	1.000	.694
Intangible assets [The company owns business process management software systems (ERP, CRM).]	1.000	.527
Intangible assets [The company has an established business reputation.]	1.000	.622
Intangible assets [The company has licenses.]	1.000	.511
Intangible assets [The company has established business contacts (social network).]	1.000	.568
Intangible assets [The company has access to distribution channels.]	1.000	.584
Intangible assets [The company has access to labor markets.]	1.000	.549
Intangible assets [The company has access to capital markets.]	1.000	.433
Intangible assets [The company has access to materials and suppliers.]	1.000	.573
Extraction Method: Principal Component Analysis.		

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	9.078	16.812	16.812	9.078	16.812	16.812	4.677	8.661	8.661
2	3.739	6.924	23.735	3.739	6.924	23.735	3.621	6.705	15.367
3	3.347	6.199	29.934	3.347	6.199	29.934	3.304	6.118	21.485
4	2.736	5.067	35.001	2.736	5.067	35.001	3.110	5.760	27.245
5	2.548	4.718	39.719	2.548	4.718	39.719	2.792	5.170	32.415
6	2.171	4.020	43.738	2.171	4.020	43.738	2.747	5.087	37.502
7	1.839	3.406	47.144	1.839	3.406	47.144	2.594	4.804	42.306
8	1.795	3.324	50.468	1.795	3.324	50.468	2.481	4.595	46.901
9	1.762	3.262	53.730	1.762	3.262	53.730	2.392	4.430	51.331
10	1.600	2.963	56.694	1.600	2.963	56.694	2.273	4.208	55.540
11	1.384	2.563	59.256	1.384	2.563	59.256	2.007	3.717	59.256
12	1.275	2.362	61.618						
13	1.213	2.247	63.865						
14	1.176	2.178	66.043						
15	1.108	2.051	68.094						
16	1.047	1.939	70.033						

17	1.031	1.910	71.943						
18	.985	1.823	73.766						
19	.938	1.736	75.503						
20	.862	1.596	77.098						
21	.803	1.487	78.585						
22	.745	1.379	79.964						
23	.729	1.351	81.315						
24	.694	1.284	82.599						
25	.660	1.222	83.822						
26	.609	1.128	84.949						
27	.582	1.077	86.027						
28	.557	1.031	87.058						
29	.513	.950	88.007						
30	.463	.857	88.864						
31	.456	.845	89.710						
32	.423	.784	90.493						
33	.403	.747	91.240						
34	.388	.718	91.958						
35	.376	.697	92.655						
36	.356	.660	93.315						
37	.337	.623	93.938						
38	.331	.612	94.551						
39	.301	.557	95.108						
40	.292	.541	95.649						
41	.275	.510	96.159						
42	.243	.450	96.608						
43	.220	.408	97.016						
44	.211	.392	97.408						
45	.202	.375	97.783						
46	.189	.350	98.133						
47	.185	.342	98.475						
48	.164	.304	98.779						
49	.149	.277	99.056						
50	.130	.241	99.297						
51	.124	.229	99.526						
52	.104	.192	99.719						
53	.092	.171	99.890						
54	.060	.110	100.000						
Extraction Method: Principal Component Analysis.									



Component Matrix[a]											
	Component										
	1	2	3	4	5	6	7	8	9	10	11
Personality and values [I can do my job without assistance.]											
Personality and values [I am self-confident.]											
Personality and values [I am enterprising, proactive.]											
Personality and values [I am ambitious.]											
Personality and values [I tend to take risks.]											
Skills and Experience [I have entrepreneurial skills (business start-up skills).]	.576										
Skills and Experience [I have management skills.]	.534										
Skills and Experience [I have marketing skills.]					-.512						
Skills and Experience [I have HR skills.]											
Skills and Experience [I have investment skills.]	.619										
Skills and Experience [I have previous experience in starting a business.]											
Skills and Experience [I have experience in a similar position (relative to the one held in moment).]						.557					
Skills and experience [I have experience in the same industry.]						.594					
Skills and experience [I have management experience.]											
Teamwork [The company has a complete team.]				-.539							
Teamwork [The team has the necessary knowledge.]											
Teamwork [The team has the necessary skills.]											
Teamwork [The team has the necessary positive attitude.]											
Planning [I plan covering a wide area.]											
Planning [I plan thoroughly.]	.531										
Planning [I make a formal (written) strategic plan.]											

Planning [I plan often.]											
Planning [I plan by functional areas / departments.]	.541										
Planning [I consider multiple alternatives when planning.]	.509										
Goals [The goals I set for my business are ambitious.]	.514										
The target market share of the company is:											
Strategic direction [I manage to stick to the strategy I initially chose.]											
Strategic direction [The strategy I have chosen is aggressive.]											
Strategic direction [I have a clear strategy for my business.]	.659										
Strategic direction [I have a clear strategy for my business.]	.656										
My business enters the market at the following stage:											
Competitive advantage [My business has a clear competitive advantage.]											
Strategic Partnerships [The firm engages in partnerships with competitors.]											
Strategic Partnerships [The firm engages in partnerships with customers.]											
Strategic Partnerships [The firm engages in partnerships with suppliers.]								.616			
Strategic partnerships [The firm engages in partnerships with other stakeholders.]											
Characteristics of the industry in which the firm operates [Most companies in the industry have a good profit.]											
Characteristics of the industry in which the firm operates [The industry in which the firm operates generates a large portion of gross domestic product (GDP).]											
Characteristics of the industry in which the firm operates [The industry has stable demand.]											
Characteristics of the industry in which the firm operates [The industry is in a period of growth.]											
Competitiveness of the industry [Competitors in the industry are aggressive.]									.506		
Competitiveness of the industry [The firm is exposed to competitive attacks.]											
Tangible assets [Company holds inventory.]											
Tangible assets [The company owns real estate.]											
Tangible assets [The company owns cash.]											
Intangible assets [The company has a recognizable brand.]											
Intangible assets [The company has software systems for managing business processes (ERP, CRM).]											
Intangible assets [The company has an established business reputation.]											
Intangible assets [The company owns licenses.]											
Intangible assets [The company has established business contacts (social network).]	.659										
Intangible assets [The company has access to distribution channels.]											
Intangible assets [The company has access to labor markets.]											
Intangible assets [The company has access to capital markets.]											
Intangible assets [The company has access to materials and suppliers.]											
Extraction Method: Principal Component Analysis.											
a. 11 components extracted.											

Rotated Component Matrix[a]											
	Component										
	1	2	3	4	5	6	7	8	9	10	11
Personality and values [I can do my job without assistance.]											
Personality and values [I am self-confident.]		.754									

Personality and values [I am enterprising, initiative.]		.761									
Personality and values [I am ambitious.]		.683									
Personality and values [I tend to take charge. risks.]		.643									
Skills and experience [I have entrepreneurial skills (business start-up skills).]											
Skills and experience [I have management skills.]				.775							
Skills and experience [I have marketing skills.]				.784							
Skills and experience [I have management skills. human resources (HR).]				.777							
Skills and experience [I have investment skills.]				.546							
Skills and experience [I have previous experience in starting a business.]					.645						
Skills and experience [I have experience in a similar position (relative to the one held in the moment).]					.782						
Skills and experience [I have experience in the same industry.]					.748						
Skills and experience [I have management experience.]					.695						
Teamwork [The company has a complete team.]			.743								
Teamwork [The team has the necessary knowledge.]			.883								
Teamwork [The team has the necessary skills.]			.886								
Teamwork [The team has the necessary positive attitude.]			.775								
Planning [I plan covering a wide area.]	.635										
Planning [I plan in depth.]	.768										
Planning [I make a formal (written) strategic plan.]	.653										
Planning [I plan often.]	.633										
Planning [I plan by functional areas / departments.]	.659										
Planning [I consider multiple alternatives when planning.]	.659										
Goals [The goals I set for my business are ambitious.]											
The target market share of the company is:											
Strategic direction [I manage to stick to the initially chosen strategy.]											
Strategic direction [The strategy I have chosen is aggressive.]											
Strategic direction [The target strategy is detailed.]	.704										
Strategic direction [I have a clear strategy for my business.]	.653										
My business enters the market at the following stage:										-.515	
Competitive advantage [My business has a clear competitive advantage.]											
Strategic partnerships [The company engages in partnerships with competitors.]											
Strategic Partnerships [The firm engages in partnerships with customers.]						.689					
Strategic Partnerships [The firm engages in partnerships with suppliers.]						.728					
Strategic Partnerships [The firm engages in partnerships with other stakeholders. countries.]						.699					
Characteristics of the industry in which the firm operates [Most companies in the industry have a good profit.]									.684		
Characteristics of the industry in which the firm operates [The industry in which the firm operates generates a large portion of gross domestic product (GDP).]									.609		
Characteristics of the industry in which the firm operates [The industry has stable demand.]									.634		
Characteristics of the industry in which the firm operates [The industry is in a period of growth.]									.704		
Competitiveness of the industry [The competitors in the industry are aggressive.]											.821
Industry competitiveness [The company is exposed to competitive attacks.]											.700

Tangible assets [The company owns inventory.]								.770			
Tangible assets [The company owns real estate properties.]								.685			
Tangible assets [The company has cash.]								.504			
Intangible assets [The company has a recognizable brand.]							.736				
Intangible assets [The company has business process management software systems (ERP, CRM).]							.647				
Intangible assets [The company has an established business reputation.]							.619				
Intangible assets [The company owns licenses.]							.636				
Intangible assets [The company owns established business contacts (social network).]											
Intangible assets [The company has access to distribution channels.]											
Intangible assets [The company has access to labor markets.]										.536	
Intangible assets [The company has access to capital markets.]											
Intangible assets [The company has access to materials and suppliers.]								.539			
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.											
a. Rotation converged in 10 iterations.											

Component Transformation Matrix											
Component	1	2	3	4	5	6	7	8	9	10	11
1	.557	.392	.295	.345	.246	.306	.257	.140	.087	.282	.054
2	-.262	-.482	-.157	.080	.276	.291	.408	.470	-.058	.046	.338
3	-.345	.198	.522	-.355	-.384	.118	.160	.284	.410	.083	.023
4	-.231	.302	-.601	.012	.200	.159	-.109	-.014	.586	.249	-.113
5	.568	-.369	-.116	-.620	-.022	.288	-.036	-.063	.229	-.020	-.049
6	-.267	-.071	.393	-.098	.656	.191	.019	-.420	.128	-.256	-.185
7	.029	.554	-.189	-.413	.233	.004	.056	.315	-.295	-.473	.136
8	-.085	.032	.056	.183	-.181	.644	-.605	-.021	-.087	-.172	.322
9	.106	-.029	.146	-.074	.262	-.466	-.291	-.002	.290	.089	.706
10	.034	-.143	.162	.003	.274	-.136	-.517	.607	-.029	.132	-.452
11	.185	-.117	-.020	.381	-.123	-.110	.080	.176	.479	-.710	-.089
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.											

Results of synthesizing success prediction models with the software product IBM SPSS Modeler

start-up Success Prediction Model

Detailed results of synthesizing a model for predicting the success of start-up **companies from Bulgaria (from** the entire data set), using the C5.0 algorithm with the application of cross-validation.

Analysis

Tree depth: 7

Cross Validation

Mean: 60,3

Standard Error: 5,3

- Analysis of survey data spss 12.sav (2014-9-5 8:31:31)
  - Number of records: 117
  - Analysis Accuracy: 83,761%

Fields

- Target
  - Company failed
- Inputs
  - Competitive AdvantageMyBusinessp
  - FSuccess_ExternalEnvironment
  - Experience_I have a similar position
  - NMActive_builtbusinessreputation
  - How to enter the market
  - Industry_MoreProfit
  - Partner_Other
  - Client_Industryhasconcentration
  - NMActive_recognizebrand

Build Settings

- Use partitioned data: true
- Partition: Partition
- Calculate predictor importance: true
- Calculate raw propensity scores: false
- Calculate adjusted propensity scores: false
- Use weight: false
- Output type: Decision tree
- Group symbolics: false
- Use boosting: false
- Cross-validate: true
- Number of folds: 10
- Mode: Simple
- Favor: Accuracy
- Expected noise (%): 0
- Use misclassification costs: false

Training Summary

- Algorithm: C5
- Model type: Classification

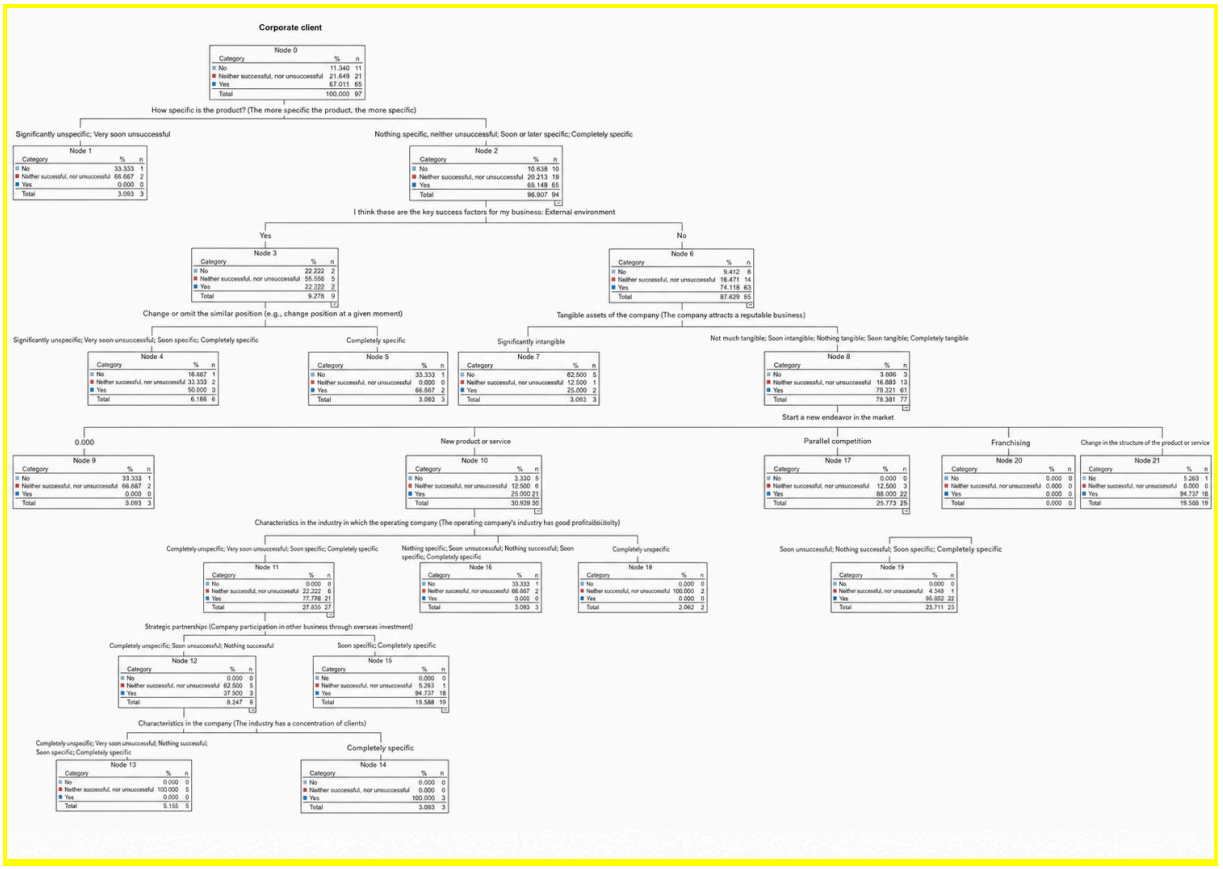
Stream: Z:\Boyan-2014-08-15.str

User: Gunshot

Date built: 14-9-5 8:30

Application: IBM® SPSS® Modeler 14.2

Elapsed time for model build: 0 hours, 0 mins, 1 secs



Competitive AdvantageMyBusinessp in [1 2] [Mode: 2] => 2.0 (3; 0.667)

Competitive AdvantageMyBusinessp in [3 4 5] [Mode: 3] (94)

FSuccess_ExternalEnvironment = 1 [Mode: 2] (9)

Experience_Improspectlikepositionspr in [1 2 3 4] [Mode: 2] => 2.0 (6; 0.833)

Experiment_Improspectlikepositionspr in [5] [Mode: 3] => 3.0 (3; 0.667)

FSuccess_ExternalEnvironment = 2 [Mode: 3] (85)

NMActive_builtbusinessreput in [1] [Mode: 1] => 1.0 (8; 0.625)

NMActive_builtbusinessreput in [2 3 4 5] [Mode: 3] (77)

Mode of market entry = 0 [Mode: 2] => 2.0 (3; 0.667)

Mode of market entry = 1 [Mode: 3] (30)

Industry_MoreProfit in [1 2 3 4] [Mode: 3] (27)

Partner_others in [1 2 3] [Mode: 2] (8)

Client_Inindustryhasconcentration in [1 2 3 4] [Mode: 2] => 2.0 (5; 1.0)

Client_Industry has concentration in [5] [Mode: 3] => 3.0 (3; 1.0)

Partner_others in [4 5] [Mode: 3] => 3.0 (19; 0.947)

Industry_MoreProfit in [5] [Mode: 2] => 2.0 (3; 0.667)

Market Entry = 2 [Mode: 3] (25)

NMActive_brand recognition in [1] [Mode: 2] => 2.0 (2; 1.0)

NMActive_brand recognition in [2 3 4 5] [Mode: 3] => 3.0 (23; 0.957)

Mode of market entry = 3 [Mode: 3] => 3.0 (0)

Mode of market entry = 4 [Mode: 3] => 3.0 (19; 0.947)

A model for predicting the success of technology start-up companies

Detailed results of synthesizing a model for predicting the success of technology **start-ups from Bulgaria** (from the data concerning only technology companies), by algorithm C5.0 by applying cross validation.

Analysis

Tree depth: 7

Cross Validation

Mean: 60,0

Standard Error: 6,2

Analysis of survey data spss 12 tech.sav (2014-11-16 10:01:41)

Number of records: 72

Analysis Accuracy: 87,5%

Fields

Target

CompanySuccessful

Inputs

Industry_MoreProfit

FSuccess_Price

Experience_InventoryManagement

SUCCESS_MARKETINGIREC

Strategy_StrategythatIhavechosen

Competition_IntheIndustryHasStrongConcon

NMAactive_builtbusinessreput

Strategic PartnershipsThe Firm

Build Settings

Use partitioned data: true

Partition: Partition

Calculate predictor importance: true

Calculate raw propensity scores: false

Calculate adjusted propensity scores: false

Use weight: false

Output type: Decision tree

Group symbolics: false

Use boosting: false

Cross-validate: true

Number of folds: 10

Mode: Simple

Favor: Accuracy

Expected noise (%): 0

Use misclassification costs: false

Training Summary

Algorithm: C5

Model type: Classification

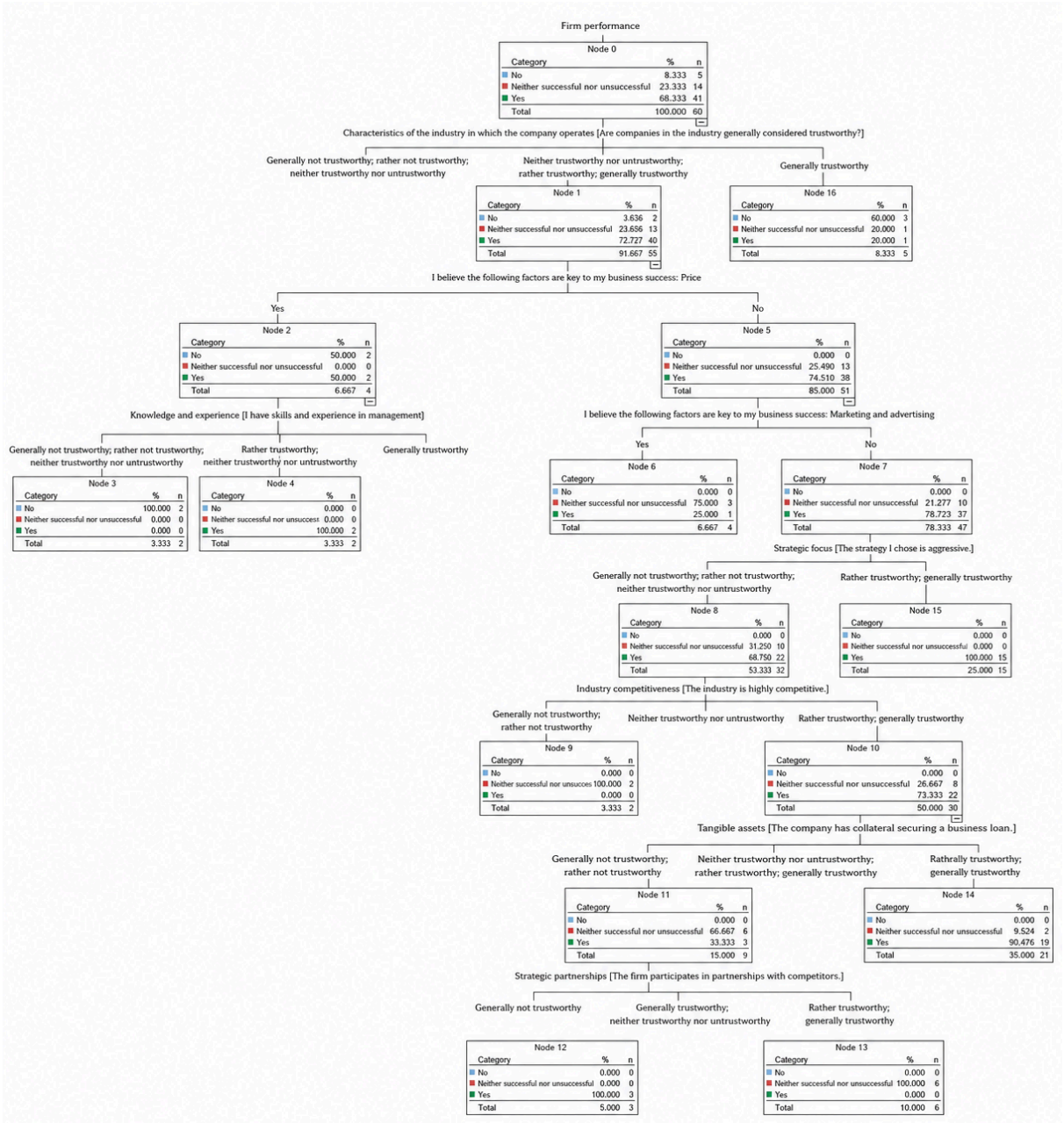
Stream: F:\Boyan-2014-11-12 tech.str

User: Gunshot

Date built: 14-11-16 10:00

Application: IBM® SPSS® Modeler 14.2

Elapsed time for model build: 0 hours, 0 mins, 33 secs



Industry_MoreProfit in [1 2 3 4] [Mode: 3] (55)

FSuccess_Price = 1 [Mode: 1] (4)

Experience_Inventory_Management in [1 2 3] [Mode: 1] => 1.0 (2; 1.0)

Experience_Inventory_Management in [4 5] [Mode: 3] => 3.0 (2; 1.0)

FU_Price = 2 [Mode: 3] (51)

FU_Marketing = 1 [Mode: 2] => 2.0 (4; 0.75)

FU_Marketing = 2 [Mode: 3] (47)

Strategy_Strategythatwehavechosen in [1 2 3] [Mode: 3] (32)

Industry_Competitionisstrongerthanconku in [1 2] [Mode: 2] => 2.0 (2; 1,0)

Competition_IntheIndustryHasStrongOnKunku in [3 4 5] [Mode: 3] (30)

NMActive_BuiltBusinessReput in [1 2 3] [Mode: 2] (9)

Strategic partnershipsThe company in [1] [Mode: 3] => 3.0 (3; 1.0)

Strategic partnershipsThe company in [2 3 4 5] [Mode: 2] => 2.0 (6; 1.0)

NMActive_builtbusinessreput in [4 5] [Mode: 3] => 3.0 (21; 0.905)

Strate_Strategythatwehavechosen in [4 5] [Mode: 3] => 3.0 (15; 1.0)

Industry_MoreProfit in [5] [Mode: 1] => 1.0 (5; 0.6)

Program code of the algorithm C5.0

A sample of the program code of the algorithm C5.0,used to synthesize models of the classification tree type is presented. The presented fragment is the function "int main(int Argc, char *Argv[])" from the code of the Linux version of the algorithm, which is implemented in the language C and distributed under open Gnu GPL license. The complete code is available for download on the manufacturer's website http://www.rulequest.com/download.html

```
/*************************************************************************/
/*                                                       */
/* Copyright 2010 Rulequest Research Pty Ltd.                    */
/*                                                       */
/* This file is part of C5.0 GPL Edition, a single-threaded version     */
/* of C5.0 release 2.07.                                 */
/*                                                       */
/* C5.0 GPL Edition is free software: you can redistribute it and/or    */
/* modify it under the terms of the GNU General Public License as */
/* published by the Free Software Foundation, either version 3 of the   */
/* License, or (at your option) any later version.               */
/*                                                       */
/* C5.0 GPL Edition is distributed in the hope that it will be useful,  */
/* but WITHOUT ANY WARRANTY; without even the implied warranty of */
/* MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU     */
/* General Public License for more details.                      */
/*                                                       */
/* You should have received a copy of the GNU General Public License    */
/* (gpl.txt) along with C5.0 GPL Edition. If not, see        */
/*                                                       */
/* <http://www.gnu.org/licenses/>.                       */
/*                                                       */
/*************************************************************************/



/*************************************************************************/
/*                                                       */
/*    Main routine, C5.0                                 */
/*    ------------------                                 */
/*                                                       */
```

```
/*****************************************************************************/


#include "defns.i"
#include "extern.i"
#include <signal.h>

#include <sys/unistd.h>
#include <sys/time.h>
#include <sys/resource.h>

#define SetFOpt(V)       V = strtod(OptArg, &EndPtr);\
                  if ( ! EndPtr || *EndPtr != '\00' ) break;\
                  ArgOK = true
#define SetIOpt(V)       V = strtol(OptArg, &EndPtr, 10);\
                  if ( ! EndPtr || *EndPtr != '\00' ) break;\
                  ArgOK = true


int main(int Argc, char *Argv[])
/* ---- */
{
into;
extern StringOptArg, Option;
char*EndPtr;
BooleanFirstTime=true, ArgOK;
doubleStartTime;
FILE*F;
CaseNoSaveMaxCase;
AttributeAtt;

struct rlimit RL;

/* Make sure there is a largish runtime stack */

getrlimit(RLIMIT_STACK, &RL);

RL.rlim_cur = Max(RL.rlim_cur, 20 * 1024 * 1024);

if ( RL.rlim_max > 0 )/* -1 if unlimited */
{
      RL.rlim_cur = Min(RL.rlim_max, RL.rlim_cur);
}

setrlimit(RLIMIT_STACK, &RL);


/* Check for output to be saved to a file */

if ( Argc > 2 && ! strcmp(Argv[Argc-2], "-o") )
{
      Of = fopen(Argv[Argc-1], "w");
      Argc -= 2;
}

if ( ! Of )
{
      Of = stdout;
}

KRInit = time(0) & 07777;
```

```
PrintHeader("");

/* Process options */

while ( (o = ProcessOption(Argc, Argv, "f+bpv+t+sm+c+S+I+ru+egX+wh")) )
{
      if ( FirstTime )
      {
      fprintf(Of, T_OptHeader);
      FirstTime = false;
      }

      ArgOK = false;

      switch (o)
      {
      case 'f': FileStem = OptArg;
            fprintf(Of, T_OptApplication, FileStem);
            ArgOK = true;
            break;
      case 'b': BOOST = true;
            fprintf(Of, T_OptBoost);
            if ( TRIALS == 1 ) TRIALS = 10;
            ArgOK = true;
            break;
      case 'p': PROBTHRESH = true;
            fprintf(Of, T_OptProbThresh);
            ArgOK = true;
            break;
#ifdef VerbOpt
      case 'v': SetIOpt(VERBOSITY);
            fprintf(Of, "\tVerbosity level %d\n", VERBOSITY);
            ArgOK = true;
            break;
#endif
      case 't': SetIOpt(TRIALS);
            fprintf(Of, T_OptTrials, TRIALS);
            Check(TRIALS, 3, 1000);
            BOOST = true;
            break;
      case 's': SUBSET = true;
            fprintf(Of, T_OptSubsets);
            ArgOK = true;
            break;
      case 'm': SetFOpt(MINITEMS);
            fprintf(Of, T_OptMinCases, MINITEMS);
            Check(MINITEMS, 1, 1000000);
            break;
      case 'c': SetFOpt(CF);
            fprintf(Of, T_OptCF, CF);
            Check(CF, 0, 100);
            CF /= 100;
            break;
      case 'r': RULES = true;
            fprintf(Of, T_OptRules);
            ArgOK = true;
            break;
      case 'S': SetFOpt(SAMPLE);
            fprintf(Of, T_OptSampling, SAMPLE);
            Check(SAMPLE, 0.1, 99.9);
```

```
            SAMPLE /= 100;
            break;
      case 'I': SetIOpt(KRInit);
            fprintf(Of, T_OptSeed, KRInit);
            KRInit = KRInit & 07777;
            break;
      case 'u': SetIOpt(UTILITY);
            fprintf(Of, T_OptUtility, UTILITY);
            Check(UTILITY, 2, 10000);
            RULES = true;
            break;
      case 'e': NOCOSTS = true;
            fprintf(Of, T_OptNoCosts);
            ArgOK = true;
            break;
      case 'w': WINNOW = true;
            fprintf(Of, T_OptWinnow);
            ArgOK = true;
            break;
      case 'g': GLOBAL = false;
            fprintf(Of, T_OptNoGlobal);
            ArgOK = true;
            break;
      case 'X': SetIOpt(FOLDS);
            fprintf(Of, T_OptXval, FOLDS);
            Check(FOLDS, 2, 1000);
            XVAL = true;
            break;
      }

      if ( ! ArgOK )
      {
      if ( o != 'h' )
      {
            fprintf(Of, T_UnregnizedOpt,
                  Option,
                  ( ! OptArg || OptArg == Option+2 ? "" : OptArg ));
            fprintf(Of, T_SummaryOpts);
      }
      fprintf(Of, T_ListOpts);
      Goodbye(1);
      }
}

if ( UTILITY && BOOST )
{
      fprintf(Of, T_UBWarn);
}

StartTime = ExecTime();

/* Get information on training data */

if ( ! (F = GetFile(".names", "r")) ) Error(NOFILE, "", "");
GetNames(F);

if ( ClassAtt )
{
      fprintf(Of, T_ClassVar, AttName[ClassAtt]);
}
```

```
NotifyStage(READDATA);
Progress(-1.0);

/* Allocate space for SomeMiss[] and SomeNA[] */

SomeMiss = AllocZero(MaxAtt+1, Boolean);
SomeNA = AllocZero(MaxAtt+1, Boolean);

/* Read data file */

if ( ! (F = GetFile(".data", "r")) ) Error(NOFILE, "", "");
GetData(F, true, false);
fprintf(Of, TX_ReadData(MaxCase+1, MaxAtt, FileStem));

if ( XVAL && (F = GetFile(".test", "r")) )
{
      SaveMaxCase = MaxCase;
      GetData(F, false, false);
      fprintf(Of, TX_ReadTest(MaxCase-SaveMaxCase, FileStem));
}

/* Check whether case weight attribute appears */

if ( CWtAtt )
{
      fprintf(Of, T_CWtAtt);
}

if ( ! NOCOSTS && (F = GetFile(".costs", "r")) )
{
      GetMCosts(F);
      if ( MCost )
      {
      fprintf(Of, T_ReadCosts, FileStem);
      }
}

/* Note any attribute exclusions/inclusions */

if ( AttExIn )
{
      fprintf(Of, "%s", ( AttExIn == -1 ? T_AttributesOut : T_AttributesIn
));

      ForEach(Att, 1, MaxAtt)
      {
      if ( Att != ClassAtt &&
             Att != CWtAtt &&
             ( StatBit(Att, SKIP) > 0 ) == ( AttExIn == -1 ) )
      {
            fprintf(Of, " %s\n", AttName[Att]);
      }
      }
}

/* Build decision trees */

if ( ! BOOST )
{
      TRIALS = 1;
}
```

```
InitialiseTreeData();
if ( RULES )
{
      RuleSet = AllocZero(TRIALS+1, CRuleSet);
}

if ( WINNOW )
{
      NotifyStage(WINNOWATTS);
      Progress(-MaxAtt);
      WinnowAtts();
}

if ( XVAL )
{
      CrossVal();
}
else
{
      ConstructClassifiers();

      /* Evaluation */

      fprintf(Of, T_EvalTrain, MaxCase+1);

      NotifyStage(EVALTRAIN);
      Progress(-TRIALS * (MaxCase+1.0));

      Evaluate(CMINFO | USAGEINFO);

      if ( (F = GetFile(( SAMPLE ? ".data" : ".test" ), "r")) )
      {
      NotifyStage(READTEST);
      fprintf(Of, "\n");

      FreeData();
      GetData(F, false, false);

      fprintf(Of, T_EvalTest, MaxCase+1);

      NotifyStage(EVALTEST);
      Progress(-TRIALS * (MaxCase+1.0));

      Evaluate(CMINFO);
      }
}

fprintf(Of, T_Time, ExecTime() - StartTime);

#ifdef VerbOpt
Cleanup();
#endif

return 0;
}
```

Model synthesis results for predicting success with the software product Weka

Detailed model synthesis results for predicting the success of **start-up companies from Bulgaria**(from the whole data set), by algorithm J48 with application of cross-validation.

=== Run information ===

Scheme:weka.classifiers.trees. J48 -C 0.25 -M 2

Relation: survey to spss 12 for weka-weka.filters.unsupervised.attribute.
Remove-R102,104-107

Instances: 117

Attributes: 103

[list of attributes omitted]

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

Competitive AdvantageMyBusinessp <= 2: 2.0 (6.0/1.0)

Competitive AdvantageMyBusinessp > 2

| FSuccess_Cost <= 1

| | Change_ManagementExperience <= 3: 1.0 (2.0)

| | Change_Management_Experience > 3: 3.0 (2.0)

| FSuccess_Cost > 1

| | FSuccess_ExternalEnvironment <= 1

| | | Experience_I have an experience similar to the position <= 4: 2.0 (5.0)

| | | Experience_I have an experience similar to the position > 4: 3.0 (3.0/1.0)

| | FSUCCESS_ExternalEnvironment > 1

| | | NMactive_builtbusinessreput <= 1

| | | | The product we offer accepted <= 2: 3.0 (3.03/1.03)

| | | | The product we offer accepted > 2: 1.0 (4.04/0.04)

| | | NMactive_builtbusinessreput > 1

| | | | NMAactive_recognizebrand <= 2

| | | | | Working with one basic client Basic <= 1: 3.0 (2.0)

| | | | | Working with one basic client Basic > 1

| | | | | Personal_Don't bother me I'll take it <= 3: 2.0 (9.15/1.0)
| | | | | Personal_Don't bother I'll take it > 3: 3.0 (4.0/1.0)
| | | | NMActive_recognitionbrand > 2: 3.0 (76.78/13.78)

Number of Leaves : 11

Size of the tree : 21

Time taken to build model: 0.25 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances 78 66.6667 %
Incorrectly Classified Instances 39 33.3333 %
Kappa statistic 0.2537
Mean absolute error 0.2018
Root mean squared error 0.3835
Relative absolute error 77.7491 %
Root relative squared error 107.3047 %
Total Number of Instances 117

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class
0.323 0.116 0.5 0.323 0.392 0.551 2.0
0.893 0.595 0.728 0.893 0.802 0.68 3.0
0.091 0.038 0.2 0.091 0.125 0.265 1.0
0 0 0 0 0 ?
Weighted Avg. 0.667 0.416 0.618 0.667 0.63 0.607

=== Confusion Matrix ===

a b c d <-- classified as

10 18 3 0 | a = 2.0

7 67 1 0 | b = 3.0

3 7 1 0 | c = 1.0

0 0 0 0 | d =

APPENDIX 3: PROGRAM CODE OF THE INFORMATION SYSTEM I3SP

The appendix presents fragments of the program code of the information system "I3SP", developed for the dissertation. The code of the main classes implementing the business logic and some of the frontend templates is presented.

GoogleService class code:

```
class GoogleService {
public $client;

public function __construct() {
$this->client = new Google_Client();
$this->client->setApplicationName(APP_NAME);
$this->client->setClientId(CLIENT_ID);
$this->client->setClientSecret(CLIENT_SECRET);
$this->client->setRedirectUri(REDIRECT_URI);
$this->client->setAccessType('online');
$this->client->setScopes(array(
'https://www.googleapis.com/auth/drive',
'https://www.googleapis.com/auth/drive.file',
'https://www.googleapis.com/auth/userinfo.email',
'https://www.googleapis.com/auth/userinfo.profile',
'https://spreadsheets.google.com/feeds'));
}

//get logged user info
function getLoggedUserInfo()
{
$token_decoded = json_decode($this->client->getAccessToken(),true);
$query = 'https://www.googleapis.com/oauth2/v1/userinfo?access_token=' .
$token_decoded["access_token"];
$user_info = file_get_contents($query);
/*
* $json should be in format
{ "id": "xx",
"name": "xx",
"given_name": "xx",
"family_name": "xx",
"link": "xx",
"picture": "xx",
"gender": "xx",
"locale": "xx",
"email" : "xx" }
*/
return json_decode($user_info,false);
}

//get access token json decoded
function getAccessTokenDecoded()
{
$token_decoded = json_decode($this->client->getAccessToken(),true);
return $token_decoded["access_token"];
}

//check for access user token
function isAccessTokenValid()
{
$token_decoded = json_decode($this->client->getAccessToken(),true);
```

```
$query = 'https://www.googleapis.com/oauth2/v1/tokeninfo?access_token=' .
$token_decoded["access_token"];
$ch = curl_init($query);
curl_setopt($ch, CURLOPT_RETURNTRANSFER, true);
curl_setopt($ch, CURLOPT_SSL_VERIFYPEER, false);

$json = '';
if( ($json = curl_exec($ch) ) === false)
{
return false;
}
else
{
$json = json_decode($json,true);
if(!isset($json) || array_key_exists('error', $json))
{
return false;
}
/*
* $json should be in format
{
"audience":"8819981768.apps.googleusercontent.com",
"user_id":"123456789",
"scope":"profile email",
"expires_in":436
}
*/
return true;
}
}

//get survey data row by file id and row_index
function getSurveyDataIndex($file_id,$row_index)
{
$json_token = json_decode($this->client->getAccessToken());
$curlParams = array (
CURLOPT_RETURNTRANSFER => true,
CURLOPT_FOLLOWLOCATION => 0,
CURLOPT_FAILONERROR => false,
CURLOPT_SSL_VERIFYPEER => false,
CURLOPT_VERBOSE => false,
);

$ch = curl_init();
curl_setopt_array($ch, $curlParams);
$getURL =
'https://spreadsheets.google.com/tq?tqx=out:csv&key='.$file_id.'&tq=select+*+
limit+1+offset+' . $row_index;
curl_setopt($ch, CURLOPT_URL, $getURL);

$headers = array();
$headers[] = "Authorization:Bearer " . $json_token->access_token;
curl_setopt($ch, CURLOPT_HTTPHEADER, $headers);
curl_setopt($ch, CURLOPT_CUSTOMREQUEST, 'GET');
curl_setopt($ch, CURLOPT_USERAGENT, 'PHP Prediction App');
curl_setopt($ch, CURLOPT_AUTOREFERER, true);
$ret = curl_exec($ch);
$info = curl_getinfo($ch);
curl_close($ch);

if($info['http_code'] != '200'
```

```
|| str_replace(array('\'', '"'), '', trim($ret)) == '' )
return false;
else
{
//return string with content
return $ret;
}
}

//get survey data row by file id and timespamp
function getSurveyDataByDate($file_id,$date)
{
//get access token
$json_token = json_decode($this->client->getAccessToken());

$curlParams = array (
CURLOPT_RETURNTRANSFER => true,
CURLOPT_FOLLOWLOCATION => 0,
CURLOPT_FAILONERROR => false,
CURLOPT_SSL_VERIFYPEER => false,
CURLOPT_VERBOSE => false,
);

$ch = curl_init();
curl_setopt_array($ch, $curlParams);

//set google spreadsheet query
$query = urlencode("select * where A = datetime '$date'");

$getURL =
"https://spreadsheets.google.com/tq?tqx=out:csv&key=$file_id&tq=$query";

curl_setopt($ch, CURLOPT_URL, $getURL);

//create request to google API
$headers = array();
$headers[] = "Authorization:Bearer " . $json_token->access_token;
curl_setopt($ch, CURLOPT_HTTPHEADER, $headers);
curl_setopt($ch, CURLOPT_CUSTOMREQUEST, 'GET');
curl_setopt($ch, CURLOPT_USERAGENT, 'PHP Prediction App');
curl_setopt($ch, CURLOPT_AUTOREFERER, true);
$ret = curl_exec($ch);
$info = curl_getinfo($ch);
curl_close($ch);

if($info['http_code'] != '200'
|| str_replace(array('\'', '"'), '', trim($ret)) == '' )
return false;
else
{
//return string with content
return $ret;
}
}


//get whole file from google drive by file id
function getSurveyDataWhole($fileId)
{
$driveService = new Google_DriveService($this->client);
// Retrieve metadata for the file specified by $fileId.
```

```
$file = $driveService->files->get($fileId);

// Get the contents of the file.
//$request = new
Google_HttpRequest($file['exportLinks']['application/vnd.openxmlformats-off
icedocument.spreadsheetml.sheet']);
$response = $this->client->getIo()->authenticatedRequest($request);
$Data = $response->getResponseBody();

return $Data;
}
}
```

AnalyseService class code:

```
class AnalyseService {
/* expects array with id , name */
private $surveyFile;
private $rowIndex;
private $sql;

function __construct($sql,$surveyFile) {
$this->surveyFile = $surveyFile;
$this->sql = $sql;
}

//dispaly the survey contents page
function displaySurveyContents()
{
global $smarty;
global $googleService;


$smarty->assign('PUBLIC_API_KEY', PUBLIC_API_KEY);
$smarty->assign('CLIENT_ID_JSAPI', CLIENT_ID_JSAPI);
$smarty->assign('APP_URL', APP_URL);
$smarty->assign('row_index', 1);
$smarty->assign('access_token', $googleService->getAccessTokenDecoded());

$smarty->display('survey_data.tpl');

}

//display the analyse history page
function displayAnalysesHistory()
{
global $smarty;
$user_id = $_SESSION["user"]["id"];

//get all save analyses for this user from the DB
$this->sql->query("SELECT a.*, p.name as prediction_name FROM analyses a LEFT
JOIN prediction_trees p on p.id = a.prediction_tree_id WHERE
a.user_id=$user_id", SQL_ALL);
$smarty->assign('analyses', $this->sql->record);
$smarty->display('analyses_history.tpl');
}

//saves an analyse to the DB
private function AnalyseSave()
```

```
{
global $googleService;
if(isset($_POST["analyse_save"]) && isset($_POST["row_index"]) &&
isset($_POST["file_id"]))
{
//prepare all parameters for the DB
$row_index = $this->sql->db->real_escape_string($_POST["row_index"]);
$percent_yes = $this->sql->db->real_escape_string($_POST["percent_yes"]);
$percent_no = $this->sql->db->real_escape_string($_POST["percent_no"]);
$percent_yesno = $this->sql->db->real_escape_string($_POST["percent_yesno"]);
$count = $this->sql->db->real_escape_string($_POST["count"]);
$date = date("Y-m-d H:i:s");
$note = $this->sql->db->real_escape_string(htmlspecialchars($_POST["note"]));
$label =
$this->sql->db->real_escape_string(htmlspecialchars($_POST["label"]));
$user_id = $_SESSION["user"]["id"];
$gfileid =
$this->sql->db->real_escape_string(htmlspecialchars($_POST["file_id"]));
$gfilename =
$this->sql->db->real_escape_string(htmlspecialchars($_POST["file_name"]));
$prediction_file_id =
$this->sql->db->real_escape_string(htmlspecialchars($_POST["prediction_file_i
d"]));
$row_date =
$this->sql->db->real_escape_string(htmlspecialchars($_POST["row_date"]));

$this->sql->query("SELECT * FROM analyses WHERE user_id=$user_id and
google_file_id='$gfileid' and row_index=$row_index and
prediction_tree_id=$prediction_file_id", SQL_INIT);
$result = $this->sql->record;
//ddecide whether to insert or update the record
if(isset($result) && $result > 0)
{
$query = "UPDATE analyses SET label='$label', note='$note',
creation_date='$date', result_yes=$percent_yes, result_no=$percent_no,
result_yesno=$percent_yesno, count=$count ";
if(isset($row_date) and date_parse($row_date))
{
$parsedDate = date_parse($row_date);
$parsedYear = $parsedDate["year"];
if($parsedYear > 2000)// check for valid date year
{
$query = $query . ", row_date='$row_date' ";
}
}
$query = $query . " WHERE user_id=$user_id and google_file_id='$gfileid' and
row_index=$row_index and prediction_tree_id=$prediction_file_id";
}
else
{
$query = "INSERT INTO analyses (prediction_tree_id, google_file_id,
google_file_name, row_index, label, note, user_id, creation_date, result_yes,
result_no, result_yesno, count, row_date) "
.
"VALUES($prediction_file_id,'$gfileid','$gfilename',$row_index,'$label','$not
e',$user_id,'$date',$percent_yes,$percent_no,$percent_yesno,$count,'$row_date
')";

}
$this->sql->query($query,SQL_NONE);
}
```

```
}

//dispay the analyse page
function displayAnalyse()
{

if(isset($_POST["analyse_save"]))
{
//if save button clicked -> store the analyse into DB
$this->AnalyseSave();
//redirect to history page
redirect(APP_URL.'?p=analyses_history');
exit();
}

global $smarty;
global $googleService;

$row_index = null;
$row_date = null;

//perform various checks for the input parameters for this page
if(isset($_GET["row_index"]) )
{
if(is_numeric($_GET["row_index"]) && $_GET["row_index"] >= 0)
{
if(isset($_GET["dbdate"]) && strlen ($_GET["dbdate"]) > 0)
{
$row_date = $_GET["dbdate"];
}
$row_index = round($_GET["row_index"]);
$user_id = $_SESSION["user"]["id"];

$prediction_file_path = $_SESSION['prediction_file']["file_path"];
$prediction_file_id = $_SESSION['prediction_file']["id"];
if(isset($_GET["prediction_file_id"]) && trim($_GET["prediction_file_id"]) !=
'')
{
$prediction_file_id =
$this->sql->db->real_escape_string(trim($_GET["prediction_file_id"]));
//get selected file
$this->sql->query("SELECT * FROM prediction_trees WHERE
id=$prediction_file_id", SQL_INIT);
if(isset($this->sql->record) && !empty($this->sql->record))
{
$res = $this->sql->record;
$prediction_file_path = $res["file_path"];
}
}
$this->sql->query("SELECT * FROM prediction_trees", SQL_ALL);
$predictions = $this->sql->record;

$file_id = '';
$file_name = '';
if(!isset($_GET["from_feed"]) && isset($_GET["file_id"]) &&
trim($_GET["file_id"]) != '')
{
$file_id = trim($_GET["file_id"]);
$this->sql->query("SELECT * FROM analyses WHERE user_id=$user_id and
google_file_id='$file_id' and row_index=$row_index", SQL_INIT);
$result = $this->sql->record;
```

```
if(isset($result))
{
$file_name = $result["google_file_name"];
}
}
else if(isset($this->surveyFile))
{
$file_id = $this->surveyFile["id"];
$file_name = $this->surveyFile["name"];
}

//$googleService->getSurveyDataWhole($file_id);
//get data from google drive about the selected row by date or index
if(isset($row_date))
{
$survey_string = $googleService->getSurveyDataByDate($file_id,$row_date);
}
else
{
$survey_string = $googleService->getSurveyDataIndex($file_id,$row_index);
}


if($survey_string)
{
//convert the csv string to assoc array NO TRANSLATIONS
$survey_data = FileService::convertCSVStringToAssocArray($survey_string);

//load dictionary data into array
$dictionary = FileService::convertDictionaryToArray();

//load the prediction tree into array and use the offeset to determine the
tree levels
$prediction_tree_array =
FileService::convertPredictionTreeToArrayWithTranslation($prediction_file_pat
h,$dictionary);
//var_dump($prediction_tree_array);

//quick Survey data validation - only check if core survey data is there /
not whole file
if(FileService::validateSurveyData($survey_data,$prediction_tree_array))
{

//see the procedure declaration for more info
$dictionary_array =
FileService::getDictionaryTranslation($survey_data,$dictionary);
//var_dump($dictionary_array);

//analyse and get results
$analyseResult = $this->Analyse($prediction_tree_array,$dictionary_array);

//generate json for the three visualisation
$prediction_tree_json =
AnalyseService::convertPredictionTreeToJson($smarty,$prediction_tree_array,$d
ictionary_array);


//check if that file / row has not been previously analysed
$label = '';
$note = '';
```

```
$this->sql->query("SELECT * FROM analyses WHERE user_id=$user_id and
google_file_id='$file_id' and row_index=$row_index and prediction_tree_id =
$prediction_file_id", SQL_INIT);
$result = $this->sql->record;
if(isset($result))
{
$label = $result["label"];
$note = $result["note"];
}

//prepare the JSON that stores the info for the tree visualization
$jsonTree = json_encode( $prediction_tree_json);
$jsonTreePathIndexs = json_encode($analyseResult->path_indexes_array);

$smarty->assign('count', round($analyseResult->count));
$smarty->assign('percent_yes', $analyseResult->percent_yes);
$smarty->assign('percent_yesno', $analyseResult->percent_yesno);
$smarty->assign('percent_no', $analyseResult->percent_no);
$smarty->assign('warning', $analyseResult->warning);
$smarty->assign('jsonTree', $jsonTree);
$smarty->assign('jsonTreePathIndexs', $jsonTreePathIndexs);
$smarty->assign('survey_raw_data', $survey_data);

$smarty->assign('predictions', $predictions);
$smarty->assign('row_index', $row_index);
$smarty->assign('row_date', $row_date);
$smarty->assign('prediction_file_id', $prediction_file_id);
$smarty->assign('file_id', $file_id);
$smarty->assign('file_name', $file_name);

$smarty->assign('label', $label);
$smarty->assign('note', $note);
}
else {
$smarty->assign('notification', 10);
$smarty->assign('additional_notes', 'The selected file for analysis has
missing key data or is not a valid survey file! Analysis cannot be performed.
Please select another file from the settings section.' );
}
}
else {
$smarty->assign('notification', 10);
$smarty->assign('additional_notes', 'An error occurred while trying to access
the selected row from the query file! Please try again or enter a smaller
row/index' );
}
}
else
{
$smarty->assign('notification', 10);
$smarty->assign('additional_notes', 'Invalid survey file line/index entered!'
);
}
}

$smarty->assign('PUBLIC_API_KEY', PUBLIC_API_KEY);
$smarty->assign('CLIENT_ID_JSAPI', CLIENT_ID_JSAPI);
$smarty->assign('APP_URL', APP_URL);
$smarty->assign('access_token', $googleService->getAccessTokenDecoded());

$smarty->display('survey_data_analyse.tpl');
```

```
}

//perform the prediction tree analyse
private function Analyse($prediction_tree_array,$dictionary_array)
{
//create the result object
$result = new AnalyseResult();

//print_r($prediction_tree_array);
$current_level = 0;

//store the steps of the algorithm
$prediction_path_array = array();

for($i=0; $i < sizeof($prediction_tree_array); $i++)
{
if($current_level == $prediction_tree_array[$i]["offset"])
{
$current_iteration_name = $prediction_tree_array[$i]["name"];

//replace , with . because in the dictionary there are no values with ,
$current_iteration_values = str_replace(',','.',
$prediction_tree_array[$i]["values"]);

$pos =
strripos($current_iteration_values,$dictionary_array[$current_iteration_name]
["value"]);
if(! ($pos === false) )
{
//entering this node
$current_level++;
$prediction_path_array[] = $prediction_tree_array[$i];

//exit Mode reached
if($prediction_tree_array[$i]["exitMode"])
{
//success! - results
$result->percent_yes = $prediction_tree_array[$i]['percent_yes'];
$result->percent_yesno = $prediction_tree_array[$i]['percent_yesno'];
$result->percent_no = $prediction_tree_array[$i]['percent_no'];
$result->count = $prediction_tree_array[$i]["Ncount"];
break;
}
}
else if($dictionary_array[$current_iteration_name]["value"] == '0.000')
{
//insufficient data for the algorithm
$result->percent_yes = $prediction_tree_array[$i]['percent_yes'];
$result->percent_yesno = $prediction_tree_array[$i]['percent_yesno'];
$result->percent_no = $prediction_tree_array[$i]['percent_no'];
$result->count = $prediction_tree_array[$i]["Ncount"];
$result->warning = "The displayed results are incomplete! There are missing
data or empty fields in the survey line.";
break;
}
}
}

//check if there are results after running the analyse if not give message
if(!isset($result->percent_yes))
{
```

```
$result->percent_yes = 0;
$result->percent_yesno = 0;
$result->percent_no = 0;
$result->count = 0;
 $result->warning = "An error occurred while running the analysis. The
displayed results are incomplete or incorrect!";
}

$prediction_path_indexes = array();
for($i = 0; $i < sizeof($prediction_path_array); $i++)
{
$prediction_path_indexes[] = $prediction_path_array[$i]["index"];
}

$result->path_array = $prediction_path_array;
$result->path_indexes_array = $prediction_path_indexes;

return $result;
}

//create the JSON for the prediction tree visualisation
private static function
convertPredictionTreeToJson($smarty,$prediction_array,$dictionary_array)
{
$visual_tree_json = array();
$visual_tree_json[] = array('Column1','Column2','Column3');

if(sizeof($prediction_array) > 0)
{
$visual_tree_json[] = array(array( "v" => '0' , "f" => '<div class="pnode">'
. ($prediction_array[0]["question_name"] == '' ?
$prediction_array[0]["name_translated"] :
$prediction_array[0]["question_name"]) . '</div>' ), null, null);
}
for($i=0;$i<sizeof($prediction_array);$i++)
{
$hint='';
if($i<sizeof($prediction_array)-1 &&
isset($dictionary_array[$prediction_array[$i+1]["name"]]["hint"]) &&
strlen($dictionary_array[$prediction_array[$i+1]["name"]]["hint"]) > 0)
{
$hint = ' <b>' . $dictionary_array[$prediction_array[$i+1]["name"]]["hint"] .
'</b>';
}
 //this should return Fully True or Yes or Early or prozivolentex
if($dictionary_array[$prediction_array[$i]["name"]]["value_name"])
{
//this will return code Example: 1.000
$survey_value =
$dictionary_array[$prediction_array[$i]["name"]]["value_name"];
//we need the value after translation - $prediction_array holds it under
"values_translated" key
$survey_values = str_replace($survey_value, '<b>'.$survey_value.'</b>',
$prediction_array[$i]["values_translated"]);
}
else
{
$survey_values = $prediction_array[$i]["values_translated"];
}

if($i<sizeof($prediction_array)-1)
```

```
$next_node_name = ($prediction_array[$i+1]["question_name"] == '' ?
$prediction_array[$i+1]["name_translated"] :
$prediction_array[$i+1]["question_name"] );
else
$next_node_name = '';


$smarty->assign('values_translated', $survey_values);
$smarty->assign('percent_yes', $prediction_array[$i]["percent_yes"]);
$smarty->assign('percent_yesno', $prediction_array[$i]["percent_yesno"]);
$smarty->assign('percent_no', $prediction_array[$i]["percent_no"]);
$smarty->assign('Ncount', $prediction_array[$i]["Ncount"]);
$smarty->assign('current_mode', $prediction_array[$i]["currentMode"]);
$smarty->assign('next_node_name', ($prediction_array[$i]["exitMode"]!='' ? ''
: $next_node_name . $hint));
$visual_tree_json[] = array(
array(
"v" => (string)$prediction_array[$i]["index"] ,

"f" => $smarty->fetch('prediction_node.tpl')),
(string)$prediction_array[$i]["parent"], null);
}

return $visual_tree_json;
}
}
```

UserService class code:

```
class UserService {
private $sql;
/* is set DB array for user */
private $loggedUser;
private $isFirstLogin;
private $authUrl;

public function __construct($sql, $authUrl)
{
$this->sql = $sql;
$this->authUrl = $authUrl;
}

//authenticates google user to the system
public function authenticate($googleUser)
{
$email = $this->sql->db->real_escape_string($googleUser->email);
$name = $this->sql->db->real_escape_string($googleUser->name);
$id = $this->sql->db->real_escape_string($googleUser->id);

$firstTimeLogged = false;
//get this user and check
//what is its admin_role and status
$this->sql->query("SELECT * FROM users WHERE email = '$email'", SQL_INIT);
$loggedUser = $this->sql->record;

//insert the google user if not presented into the DB
if(!$loggedUser)
{
```

```
$res = $this->sql->query("INSERT INTO
users(google_id,email,name,admin_role,status) values
('$id','$email','$name',2,1)", SQL_NONE);

if($res)
{
$firstTimeLogged = true;
$this->sql->query("SELECT * FROM users WHERE email = '$email'", SQL_INIT);
$loggedUser = $this->sql->record;
}
else
{
return false;
}
}

//set user profile
$loggedUser["picture"] = $googleUser->picture;
$this->loggedUser = $loggedUser;
if($this->loggedUser["status"] == 2)
{
//ok
$_SESSION["user"] = $this->loggedUser;
$_SESSION["mode"] = 'normal';
}
else if ($this->loggedUser["status"] == 3)
{
//suspended
$_SESSION["mode"] = 'suspend';
}
else
{
if($firstTimeLogged)
{
$_SESSION["pending_name"] = $this->loggedUser["name"];
$_SESSION["pending_email"] = $this->loggedUser["email"];
$_SESSION["mode"] = 'pending_first_time';
}
else
$_SESSION["mode"] = 'pending';

}

return true;
}

//get user from session if exists
function setup()
{
//authenticates and populate $this->loggedUser;
if(isset($_SESSION["user"]))
{
$this->loggedUser = $_SESSION["user"];
}
}

//logout user - remove stored data from session
function logout()
{
$this->loggedUser=null;
unset($_SESSION['user']);
```

```
unset($_SESSION['access_token']);
unset($_SESSION['isFirstLogin']);
session_destroy();
}

//display default page for either not logged user or admin/normal logged
user
public function displayDefault()
{
global $smarty;
if(isset($this->loggedUser))
{
if($this->loggedUser["status"] == 2)
{
//authorised
if($this->loggedUser["admin_role"] == 1)
{
$this->displayAdminDefaultView();
}
else
{
$this->displayUserDefaultView();
}
}

}
else if(isset($_SESSION["mode"]))
{
if ($_SESSION["mode"] == 'suspend')
{
$this->displayUserSuspendedView();
$_SESSION["mode"] = '';
}
else if ($_SESSION["mode"] == 'pending')
{
$this->displayUserPendingView();
$_SESSION["mode"] = '';
}
else if($_SESSION["mode"] == 'pending_first_time')
{
$this->isFirstLogin = true;
$this->displayUserPendingView();
$_SESSION["mode"] = '';
}
else
{
$this->displayDefaultView();
}
}
else{
$this->displayDefaultView();
}
}

//display default for logged users
private function displayDefaultView()
{
global $smarty;
if(isset($this->loggedUser))
{
if($this->loggedUser["admin_role"] == 1)
```

```
{
$this->displayAdminDefaultView();
}
else
{
$this->displayUserDefaultView();
}
}
else
{
$smarty->assign('authUrl', $this->authUrl);
$smarty->assign('mode', 'public');
$smarty->display('index_public.tpl');
}
}

//pending user view
private function displayUserPendingView()
{
global $smarty;
//global $smarty;
$mode = 'pending';
if($this->isFirstLogin)
{
$mode = 'pending_first_time';
//send email to admin
$pending_user_name = $_SESSION["pending_name"];
$pending_user_email = $_SESSION["pending_email"];


$this->sql->query('SELECT * FROM users WHERE admin_role=1',SQL_INIT);
$record = $this->sql->record;
if(isset($record))
{
$smarty->assign('name',$pending_user_name);
$smarty->assign('email',$pending_user_email);
$smarty->assign('APP_URL',APP_URL);
$email_body = $smarty->fetch("email_newuser.tpl");
 $email = new Email('prediction@' . APP_DOMAIN,$record["email"],'New user at
' . APP_NAME,$email_body );
$email->send();
}
}
$smarty->assign('mode', $mode);
$smarty->display('index_public.tpl');

}

//suspend user view
private function displayUserSuspendedView()
{
global $smarty;
$smarty->assign('mode', 'blocked');
$smarty->display('index_public.tpl');
}

//display normal logged user view
function displayUserDefaultView()
{
global $smarty;
global $googleService;
```

```
//set prediction default algorithm if not any
if(!isset($_SESSION['prediction_file']) || isset($_POST['prediction_id']))
{
//if prediction algorithm changed from page
if(isset($_POST['prediction_id']))
{
$prediction_id = $this->sql->db->real_escape_string($_POST['prediction_id']);
$this->sql->query("SELECT * FROM prediction_trees WHERE id=$prediction_id",
SQL_INIT);
}
//if not set any set the default one
else if(!isset($_SESSION['prediction_file']))
{
$this->sql->query("SELECT * FROM prediction_trees WHERE is_default = 1",
SQL_INIT);
}

if(isset($this->sql->record))
{
$data = $this->sql->record;
if(!empty($data))
{
$prediction = array();
$prediction["id"] = $data['id'];
$prediction["name"] = $data['name'];
$prediction["file_path"] = $data['file_path'];
$_SESSION['prediction_file'] = $prediction;
}
}
}

//set survey file - picked from google
if(isset($_POST["file_id"]) && isset($_POST["file_name"]) )
{
if( isset($_SESSION["current_file"]) )
{
unset($_SESSION["current_file"]);
}

$_SESSION["current_file"] = array(
"id" => $_POST["file_id"],
"name" => $_POST["file_name"]
);

redirect(APP_URL . '?p=survey_data');
}



$this->sql->query("SELECT * FROM prediction_trees", SQL_ALL);
if(isset($this->sql->record))
{
$smarty->assign('predictions',$this->sql->record);
//$smarty->assign('prediction_selected',$_SESSION['prediction_file']);
}
$smarty->assign('PUBLIC_API_KEY', PUBLIC_API_KEY);
$smarty->assign('CLIENT_ID_JSAPI', CLIENT_ID_JSAPI);
$smarty->assign('APP_URL', APP_URL);
$smarty->assign('access_token', $googleService->getAccessTokenDecoded());
$smarty->display('index.tpl');
```

```
}

//admin defaul view
public function displayAdminDefaultView()
{
global $smarty;
if($this->loggedUser["admin_role"] == 1)
{
//show page list of users
$this->sql->query("SELECT * FROM users", SQL_ALL);
$allUsers = $this->sql->record;

//var_dump($allUsers);
$smarty->assign('users', $allUsers);
$smarty->display("admin_manageusers.tpl");


}
else
{
$this->displayNotAuthorised();
}
}

//not authorized view
private function displayNotAuthorised()
{
global $smarty;
$smarty->assign('mode', 'notauthorised');
$smarty->display('index_public.tpl');
}
}
```

Code of frontend template "\templates\admin_manageusers.tpl":

```
{config_load file="user.conf" section="setup"}
{include file="header.tpl" title="Application User Settings"}
{if isset($notification)}
{include file="notification.tpl" notification=$notification}
{/if}
<fieldset>
<legend>Users in the application</legend>
<br>
<table class="tabledefault">
<tr>
<th>User</th>
<th>Email</th>
<th>Administrator</th>
<th>Status</th>
</tr>
{foreach name=outer item=user from=$users}
<tr>
<td>{$user["name"]}</td>
<td>{$user["email"]}</td>
<td>{if $user["admin_role"] == 2}No{else}Yes{/if}</td>
<td>{if $user["status"] == 2}
Active {if $user["admin_role"] == 2}(<a
href="{$SCRIPT_NAME}?p=admin_manageusers&block=&user={$user["id"]}">Blocked</
a>){/if}
```

```
{elseif $user["status"] == 3}
Banned (<a
href="{$SCRIPT_NAME}?p=admin_manageusers&activate=&user={$user["id"]}">Activa
te</a>)
{else}
Pending approval (<a
href="{$SCRIPT_NAME}?p=admin_manageusers&activate=&user={$user["id"]}">Activa
te</a>)
{/if}</td>
</tr>
{/foreach}
</table>
</fieldset>
{include file="footer.tpl"}
```

Program code of frontend template "\templates\survey_data_analyse.tpl":

```
{config_load file="user.conf" section="setup"}
{include file="header.tpl" title="Survey Results"}

{if isset($smarty.session.current_file) || (isset($file_id) && $file_id !=
'') }

{if isset($row_index) }
<fieldset>
<legend>Success prediction</legend>
{if isset($warning)}<div class="pt20 pb20"><b
style="color:red">{$warning}</b></div>{/if}
<h3>Results:</h3>
<br>
<div class="fl pr20">

{include file="chart.tpl" percent_Yes=$percent_yes percent_No=$percent_no
percent_YesNo=$percent_yesno ncount=$count}
</div>
<div class="fl pl20">

<form action="{$SCRIPT_NAME}?p=survey_data_analyse" method="post">
<table >
<tr>
<td>Survey file / line: </td>
<td>{$file_name} / {$row_index}</td>
</tr>
<tr>
<td>Algorithm: </td>
<td>
<select name="prediction_id" id="prediction_id">
{foreach name=outer item=prediction_file from=$predictions}
<option value="{$prediction_file["id"]}" {if isset($prediction_file_id) &&
$prediction_file_id eq
$prediction_file["id"]}selected="selected"{/if}>{$prediction_file["name"]}<
/option>
{/foreach}
</select>
</td>
</tr>
<tr>
<td>Tag: </td>
<td><input type="text" name="label" id="label" value="{$label}"/></td>
```

```
</tr>
<tr>
<td colspan="2">
Note: <br>
<textarea name="note" id="note" rows="4" cols="45">{$note}</textarea>
</td>
</tr>
<tr>
<td colspan="2">
<input type="hidden" id="row_index" name="row_index" value="{$row_index}" />
<input type="hidden" id="row_date" name="row_date" value="{$row_date}" />
<input type="hidden" id="prediction_file_id" name="prediction_file_id"
value="{$prediction_file_id}"/>
<input type="hidden" id="file_id" name="file_id" value="{$file_id}" />
<input type="hidden" id="file_name" name="file_name" value="{$file_name}" />

<input type="hidden" id="count" name="count" value="{$count}" />
<input type="hidden" id="percent_yes" name="percent_yes"
value="{$percent_yes}" />
<input type="hidden" id="percent_no" name="percent_no" value="{$percent_no}"
/>
<input type="hidden" id="percent_yesno" name="percent_yesno"
value="{$percent_yesno}" />
<input type="hidden" id="analyse_save" name="analyse_save"/>


<button>Save changes</button>
<a class="myButton" id="btnRecalc"
href="{$SCRIPT_NAME}?p=survey_data_analyse&row_index={$row_index}&file_id={
$file_id}&prediction_file_id=">Recalculate</a>
<a class="myButton" href="{$SCRIPT_NAME}?p=survey_data_analyse">Analyze New
Row</a>
</td>
</tr>
</table>
</form>
</div>
<div class="clear">
</fieldset>

<div id="pnlChart"></div>
<script type="text/javascript"
src="//www.google.com/jsapi?key={$PUBLIC_API_KEY}"></script>
<script type="text/javascript">
{literal}
google.load('visualization', '1', {packages: ['orgchart']});
{/literal}
</script>
<script type="text/javascript">
function drawChart() {
var dataArr = {$jsonTree};
var jsonTreePathIndexs = $.parseJSON('{$jsonTreePathIndexs}');
var data = google.visualization.arrayToDataTable(dataArr);
//set different style for the path of the algorithm
data.setRowProperty(0, 'style', "border: 3px solid orange");
for(var i=0;i<jsonTreePathIndexs.length;i++)
{
//console.log(jsonTreePathIndexs[i]);
if(i != jsonTreePathIndexs.length-1)
data.setRowProperty(jsonTreePathIndexs[i], 'style', "border: 3px solid
orange");
```

```
else
data.setRowProperty(jsonTreePathIndexs[i], 'style', "border: 2px solid red");
}
var chart = new google.visualization.
OrgChart(document.getElementById('pnlChart'));
{literal}
chart.draw(data, {allowHtml:true});
{/literal}
}

$(function(){

google.setOnLoadCallback(drawChart);

$('#btnRecalc').click(function(e){
e.preventDefault();
location.href = $('#btnRecalc').attr('href') + $('#prediction_id').val();
});
});
</script>
</script>

{if isset($survey_raw_data)}
All responses to the selected row for analysis:<br>
{foreach name=outer key=question item=answer from=$survey_raw_data}
<br>{$question} : {$answer}
{/foreach}
{/if}
{else}
{if isset($notification)}
{include file="notification.tpl" notification=$notification
additional_notes=$additional_notes}
{/if}
<fieldset>
<legend>Row/Index Analysis</legend>
<div>
<form method="get" action="{$SCRIPT_NAME}?p=survey_data_analyse">
<input type="hidden" name="p" id="p" value="survey_data_analyse"/>
 Poll file: {$smarty.session.current_file.name}
<br/>
Select a row for analysis / index: <input type="text" id="row_index"
name="row_index" value="" /><button>Make prediction</button>
</form>
</div>
</fieldset>
{/if}
{else}
<fieldset>
<legend>Row/Index Analysis</legend>
<div class="warningbox">Warning: No survey file selected</div>
To view the contents of a file and subsequently analyze it, please first
select one from the <a href="{$SCRIPT_NAME}?p=config">configuration
station</a>.
</fieldset>
{/if}


{include file="footer.tpl"}
```

Exported information about the structure of the database in the format . SQL:

```
CREATE TABLE IF NOT EXISTS `analyses` (
`id` int(11) NOT NULL AUTO_INCREMENT,
`prediction_tree_id` int(11) NOT NULL,
`google_file_id` varchar(150) NOT NULL,
`google_file_name` varchar(150) NOT NULL,
`row_index` int(11) NOT NULL,
`label` varchar(250) NOT NULL,
`note` varchar(500) NOT NULL,
`user_id` int(11) NOT NULL,
`creation_date` datetime NOT NULL,
`result_yes` float NOT NULL,
`result_no` float NOT NULL,
`result_yesno` float NOT NULL,
`count` int(11) NOT NULL,
`row_date` datetime DEFAULT NULL,
PRIMARY KEY (`id`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8 ;

CREATE TABLE IF NOT EXISTS `prediction_trees` (
`id` int(11) NOT NULL AUTO_INCREMENT,
`name` varchar(150) NOT NULL,
`file_path` varchar(500) NOT NULL,
`user_id` int(11) NOT NULL,
`is_default` bit(1) NOT NULL,
`upload_date` timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP,
PRIMARY KEY (`id`),
KEY `file_path` (`file_path`(255))
) ENGINE=InnoDB DEFAULT CHARSET=utf8 ;

CREATE TABLE IF NOT EXISTS `users` (
`id` int(11) NOT NULL AUTO_INCREMENT,
`google_id` varchar(100) NOT NULL,
`email` varchar(254) NOT NULL,
`name` varchar(150) NOT NULL,
`admin_role` int(11) NOT NULL COMMENT '1-admin;2-user',
`status` int(11) NOT NULL COMMENT '1 pending; 2-authorized; 3-blocked',
PRIMARY KEY (`id`),
UNIQUE KEY `google_id` (`google_id`,`email`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8 ;
```