Сравнение на класификационни модели за стартиращи компании

Боян Янков

Comparison of Classification Models for Start-up Companies

Boyan Yankov

Резюме: *Проведено е количествено изследване на факторите за успех на стартиращи компании от България. Наборът от данни за 136 компании е анализиран с помощта на софтуерните продукти за извличане на знания от данни - IBM SPSS Modeler и Weka. Като резултат са синтезирани класификационни модели за предсказване на успеха на стартиращи компании от България. Получените модели са анализирани и сравнени, като са избрани най-точните и ефективни модели. Идентифицирани са факторите за успех на компаниите, включени в моделите, както и принципът на вземане на решение за тяхната класификация.*

Ключови думи: *предприемачество, стартиращи компании, предсказване на успеха, бизнес модел, нови компании, IBM SPSS Modeler, Weka*

Abstract: *A quantitative research of the success factors of Bulgarian start-up companies is conducted. The dataset of 136 companies is analyzed with the help of two data mining software products: IBM SPSS Modeler and Weka. As a result, classification models for Bulgarian start-ups succes prediction are synthesized. The models are then analyzed, compared and the most accurate ones are selected. The success factors contained in the models are identified and the decision taking principles are observed.*

Keywords: *entrepreneurship, start-up companies, success prediction, business model, new ventures, IBM SPSS Modeler, Weka*

JEL: *M13, C38*

1. Въведение

В условията на икономика на прехода, характерни за Централна и Източна Европа, се разчита в голяма степен на стартиращите компании, за да генерират условия за икономически растеж. В приоритетите на програмния период на Европейския Съюз от 2014 до 2020 година „Хоризонт 2020" е залегнала необходимостта от развитието на иновативния малък и средния бизнес, с цел генериране на икономически растеж и работни позиции [1].

Предсказването на успеха за стартиращи компании е възможност за увеличаване на ефективността на процеса на стартиране на бизнес и за намаляване на риска и разхода на ресурси. Създадени са модели за предсказване на успехи на стартиращи компании, които са специфични за различни пазари, държави, икономически особености, както и понякога за моментната ситуация на тяхното създаване. Настоящото изследване е фокусирано върху създаване на модели за предсказване на успеха на стартиращи компании от България.

За целите на изследването приемаме, че стартираща компания е такава, която е в началото на своето развитие и по размер е SME (микро, малка или средно голяма). На база проведени интервюта с български предприемачи, инвеститори и преподаватели по предприемачество, за успешна стартираща компания приемаме тази, която е увеличила своя размер през последните пет години.

След анализ на литературни източници, методологии и модели за предсказване на успеха от предишни изследвания, е създаден теоретичен модел за предсказване на успеха (хипотеза), който съдържа факторите за успех (както и тяхната категоризация) за стартиращи компании от България. Изготвеният модел е ревизиран и адаптиран за българската бизнес среда чрез количествено изследване – провеждане на интервюта с български предприемачи [2].

На база предложения модел за предсказване на успеха на стартиращи компании, е проведено количествено изследване. 136 предприемачи и собственици на малки и средно големи стартиращи компании от България попълниха онлайн анкета, съдържаща над 100 фактора за успех от

предложения модел. Предложената категоризация на факторите за успех от теорети1ния модел беше успешно валидирана посредством факторен анализ на събраните данни.

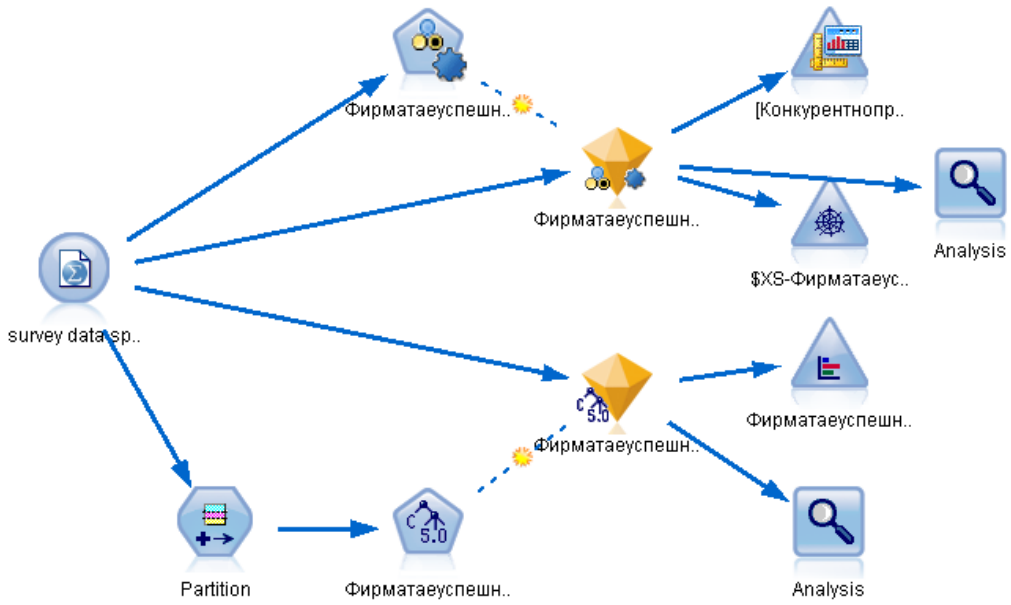
В настоящата статия ще бъдат разгледани модели за предсказване на успеха на стартиращи компании, синтезирани чрез прилагане на два софтуерни продукта за извличане на знания от данни – IBM SPSS Modeler и Weka върху набора от данни от количественото изследване.

2. Извличане на знания от данни и създаване на модел за предсказване на пуспеха с IBM SPSS Modeler

С помощта на софтуерния продукт IBM SPSS Modeler [3] са създадени модели за предсказване на успеха, които извършват класификация на компаниите - предвиждат в коя категория попадат те: „успешни“, „нито успешни, нито неуспешни“ или „неуспешни“. Софтуерният продукт предлага инструмент, който автоматично генерира множество модели, които решават поставената задача, сравнява ги и предлага най-добрите модели.

Постановка на експеримента

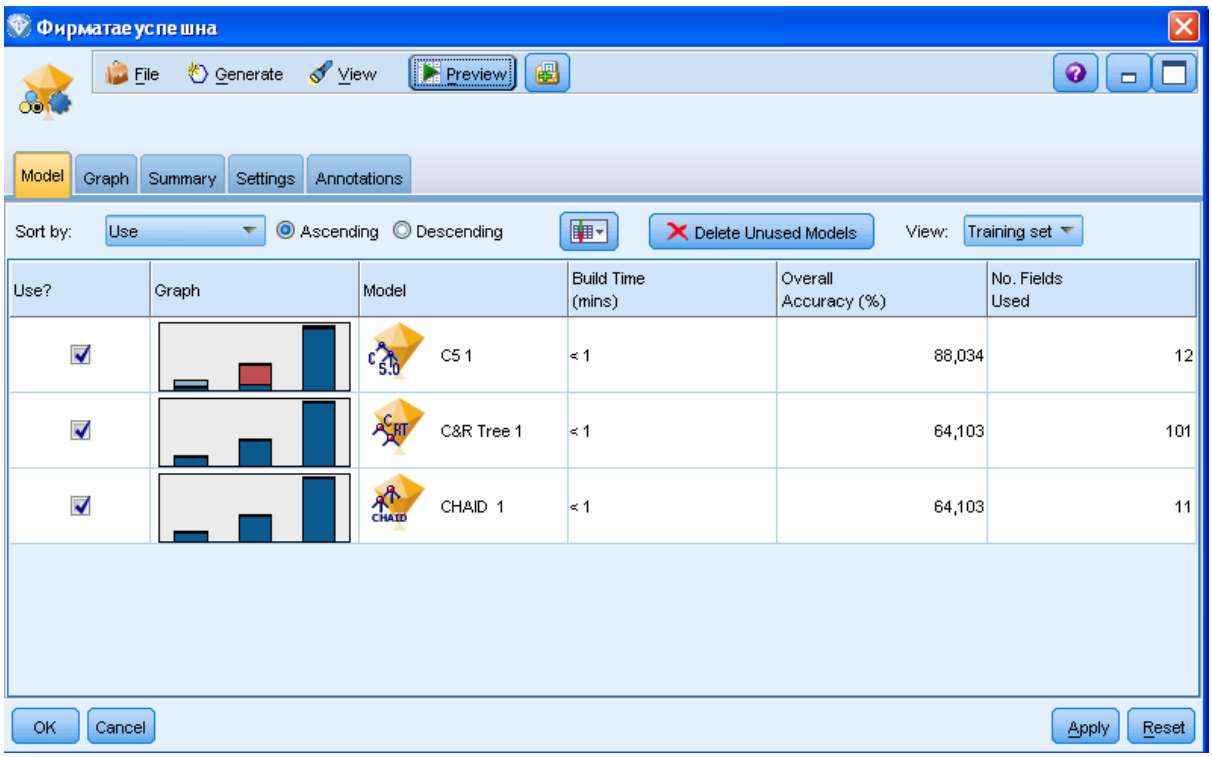
В IBM SPSS Modeler последователностите от действия, които се изпълняват могат да бъдат представени графично. На Фигура 1 са показани графично последователностите от действия за получаване на модели. Последователността в горната част на фигурата показва получаване на модели чрез автоматичния класификатор, а последователността в долната й част – получаване на модел чрез алгоритъма C5.0 с ръчни настройки и прилагане на кръстосана валидация.



Фиг. 1: Последователност от действия в IMB SPSS Modeler за получаване на класификационни модели – графично представяне

Данните от анкетите се зареждат в IMB SPSS Modeler чрез звеното „Statistics File“, разположено в лявата част на фигурата и маркирано с етикет „survey data spss“. Звеното за автоматична класификация „Auto Classifier“ синтезира, оценява и сравнява класификационни модели за избраната цел, използвайки различни алгоритми. То е настроено, като са избирани цел и входни данни. При синтезирането за модели за предсказване на успеха, за цел е избрана променливата успех на компанията, показваща в каква степен компанията е успешна. Входните данни са останалите независими променливи. Получените модели за предсказване на успеха могат да бъдат разгледани в звено с форма на диамант „model nugget“, разположено в горната част на фигурата.

Автоматичният класификатор класира получените модели и запазва най-добрите три модела за предсказване на успеха (тези с най-висока точност) за по-нататъшен анализ (Фиг. 2).



Фиг. 2: Най-добри получени модели за предсказване на успеха при автоматична класификация с IBM SPSS Modeler

Получените модели са съпоставени по своята точност (overall accuracy), като в случая не се прилага кръстосана валидация и посочената точност е по-скоро максимална, отколкото реалистична. Най-точен е моделът, получен чрез алгоритъма C5.1 – с точност 88,03%. Останалите модели са получени чрез алгоритмите C&R Tree 1 (дърво за класификация и регресия) и CHAID 1, като точността и на двата е 64,10%.

Базирайки се на получените модели, избираме алгоритъма C5.1, тъй като чрез него е получен модел с най-висока точност. Алгоритъмът ще бъде използван като отправна точка за синтезиране на модел с висока точност, чрез прилагане на кръстосана валидация. По този начин ще бъде създаден модел, чиято точност ще бъде оценена по-реалистично.

Синтезиране на модел с кръстосана валидация чрез алгоритъма C5.0

При създаване на модел за предсказване на успеха, може да се приложи методът кръстосана валидация. При кръстосаната валидация данните се

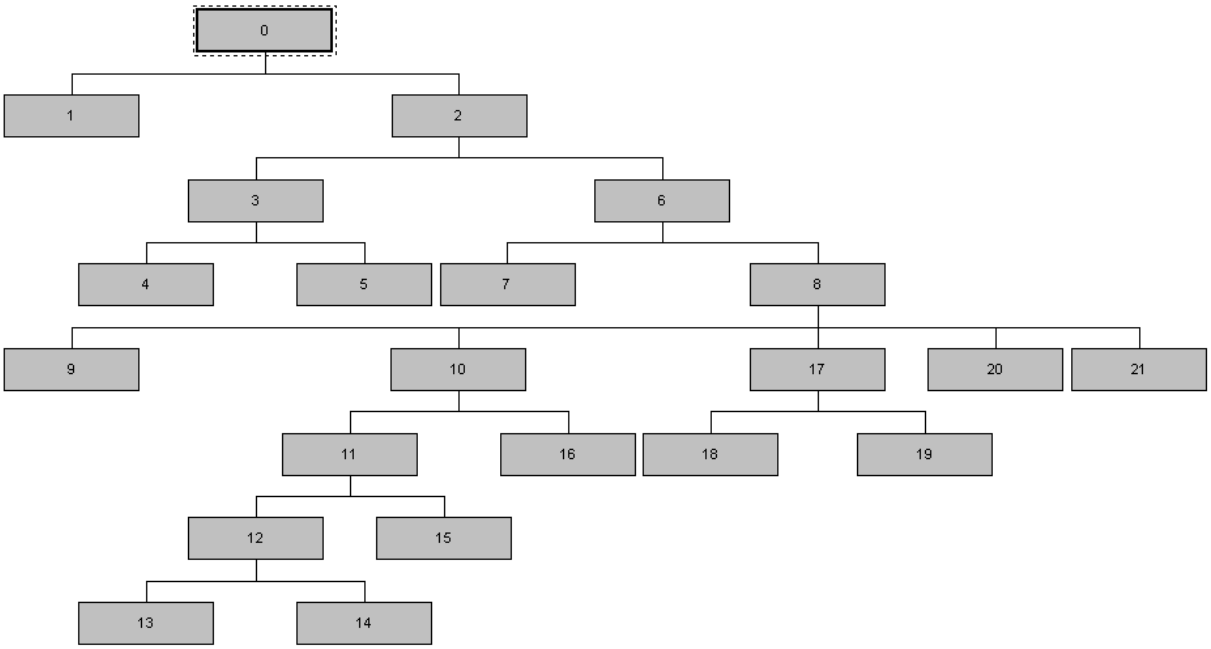
разделят на два набора: набор за обучение (training set), служещ за създаване на модела и набор за тестване (testing set), служещ за проверка на модела. Предимството на този метод е, че точността на получения модел е значително по-реалистична. Авторът прилага алгоритъма C5.0, който поддържа множество ръчни настройки и кръстосана валидация на данните. След прилагането му върху данните, се получава **модел с точност 83,76%**.

Полученият чрез кръстосана валидация модел представлява класификационно дърво. Моделът съдържа следните 9 променливи, които описват успешната стартираща компания, като след всяка променлива в скоби е изписан съответстващия номер на възел в дървото, чиито разклонения зависят от променливата:

- Конкурентно предимство: Моят бизнес притежава ясно конкурентно предимство. (възел 0)
- Смятам, че следните фактори са ключови за успеха за моя бизнес: Външна среда. (възел 2)
- Умения и опит: Имам опит на подобна позиция (спрямо заеманата в момента). (възел 3)
- Нематериални активи Компанията притежава изградена бизнес репутация. (възел 6)
- Начин на навлизане на пазара (възел 8)
- Нематериални активи: Компанията притежава разпознаваем бранд. (възел 17)
- Характеристики на индустрията, в която фирмата оперира: Повечето компании в индустрията имат добра печалба. (възел 10)
- Стратегически партньорства: Фирмата участва в партньорства с други заинтересовани страни. (възел 11)
- Характеристики на клиентите: В индустрията има концентрация на клиенти. (възел 12)

Цялостно графично представяне на класификационното дърво (карта на дървото) е показано на Фигура 3, като във всеки от възлите е изписан неговия номер. Най-горният възел (0) е основен и се разклонява на два възела

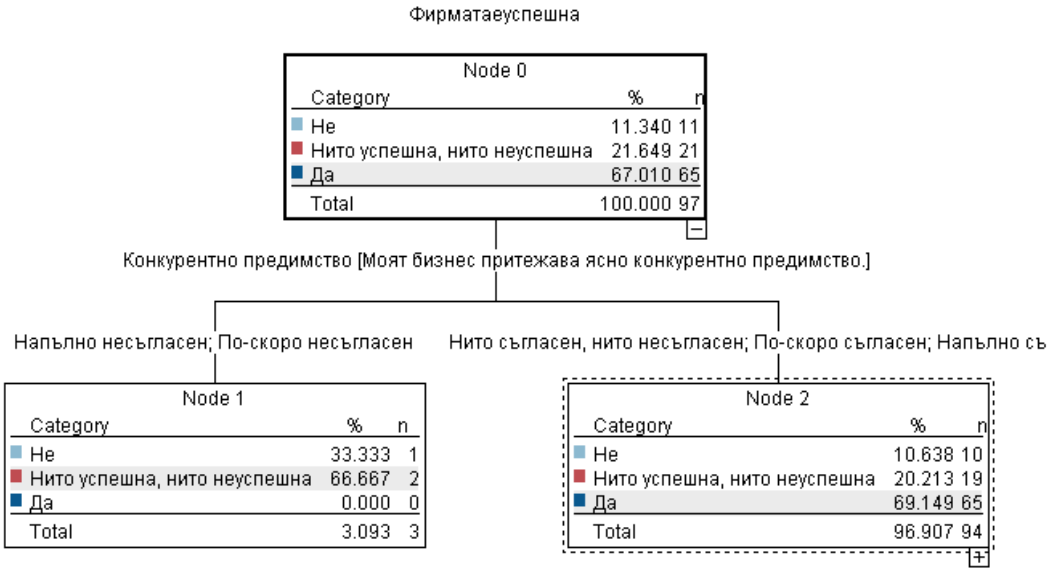
(1 и 2) в зависимост стойността за основния фактор за успех на компанията – наличието на конкурентно предимство. След това възел 2 се разклонява в зависимост от следващия фактор и така до достигането на краен възел.



Фиг. 3: Цялостно графично представяне (карта) на класификационното дърво

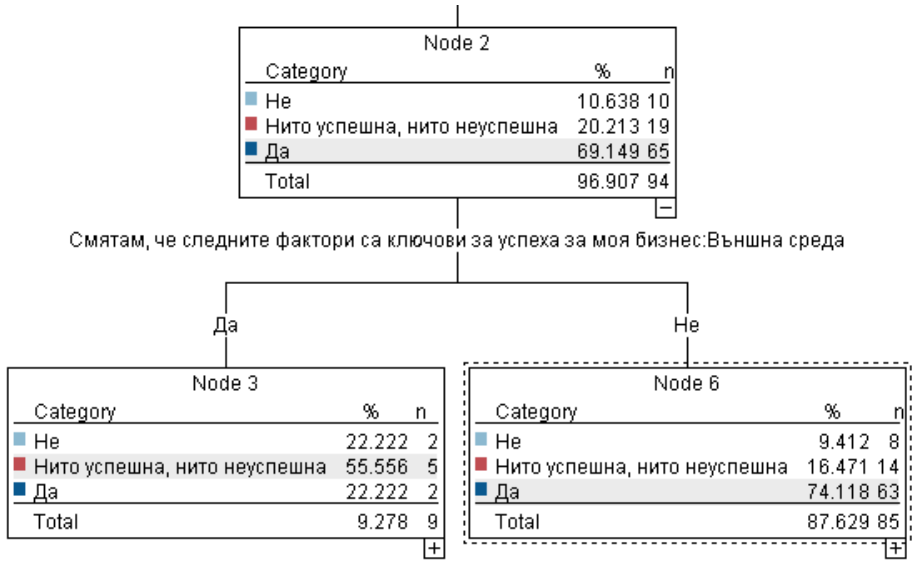
Ще бъдат разгледани някои от възлите на класификационното дърво, които съдържат най-важните фактори за успех, както и техните разклонения. Чрез разглеждане на структурата на класификационното дърво, можем да установим важността на факторите за успех, тъй като най-важните фактори се намират в горните възли на дървото и определят първоначалното разделяне на данните.

Първото ниво на дървото (възел 0) има две разклонения, зависещи от променливата „наличие на ясно конкурентно предимство“ (Фиг. 4). Тази променлива е най-важна за предсказване на успеха на компаниите от анализирания набор от данни. Стартиращите компании, които имат ясно **конкурентно предимство** (възел 2), са по-успешни от тези, които нямат (възел 1). Потвърждава се тезата, че наличието на ясно конкурентно предимство е ключов фактор за успеха на стартиращите компании [4].



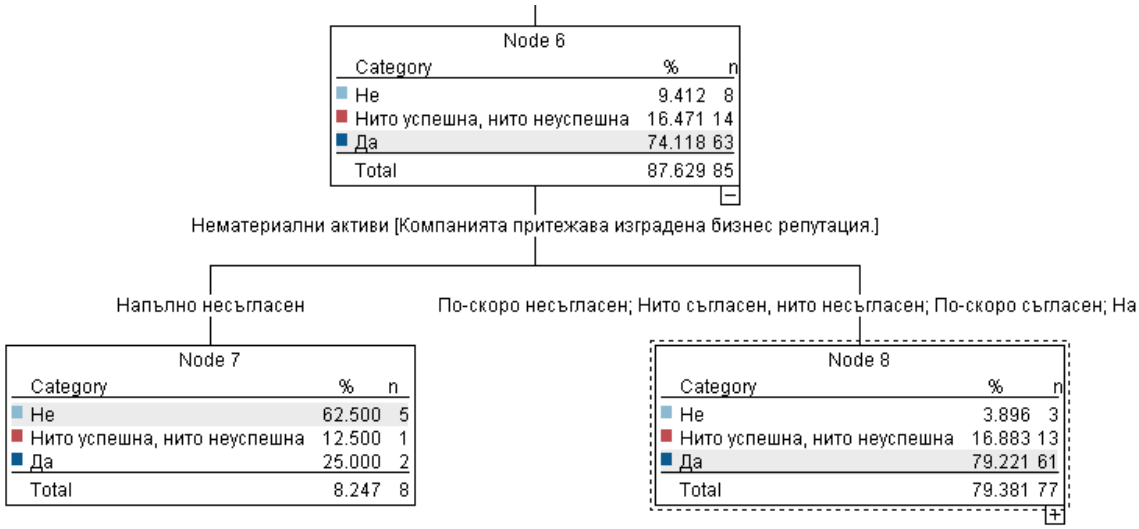
Фиг. 4: Класификационно дърво – първо ниво: Наличие на ясно конкурентно предимство

Възел 2 има две разклонения в зависимост от следващия фактор за успех на компанията „**околната среда е ключов фактор за успеха**“ (Фиг. 5). Тези компании от анализирания набор от данни, които считат околната среда за ключов фактор за успеха (възел 3), са по-малко успешни от останалите (възел 6). Възел 3 има малък брой елементи от набора от данни и по тази причина няма да анализирам в дълбочина неговите наследници.



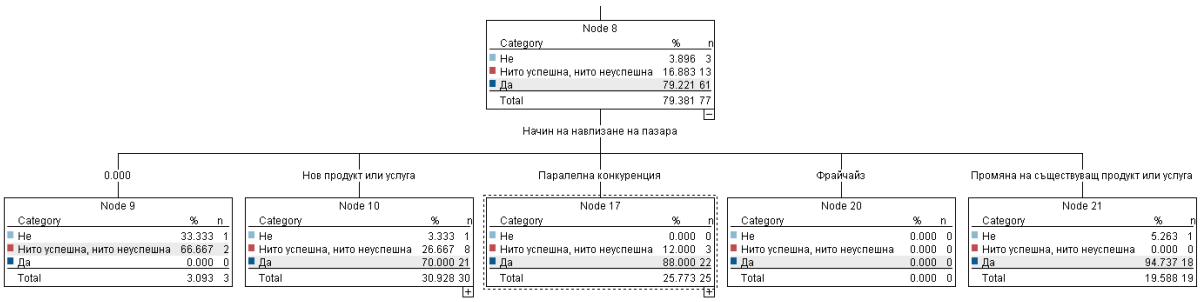
Фиг. 5: Класификационно дърво – клон 2: външната среда е ключов фактор за успеха на компанията

Възел 6 има два разклонения в зависимост от следващия фактор за успех на компанията „**наличие на изградена бизнес репутация**“ (Фиг. 6). Тези компании от анализирания набор от данни, които имат изградена бизнес репутация (възел 8), са по-успешни от останалите (възел 7).



Фиг. 6: Класификационно дърво – клон 2.2: нематериални активи – изградена бизнес репутация на компанията

Възел 8 има множество разклонения в зависимост от следващия фактор за успех на компанията „**начин на навлизане на пазара на стартиращата компания**“ (Фиг. 7). Промяната на съществуващ продукт или услуга е най-успешният начин на навлизане на пазара за компаниите от анализирания набор от данни. Фирмите, които развиват нов продукт или услуга, както и тези, които навлизат на пазара чрез паралелна конкуренция най-често са успешни.



Фиг. 7: Класификационно дърво – клон 2.2.2: начин на навлизане на стартиращата компания на пазара

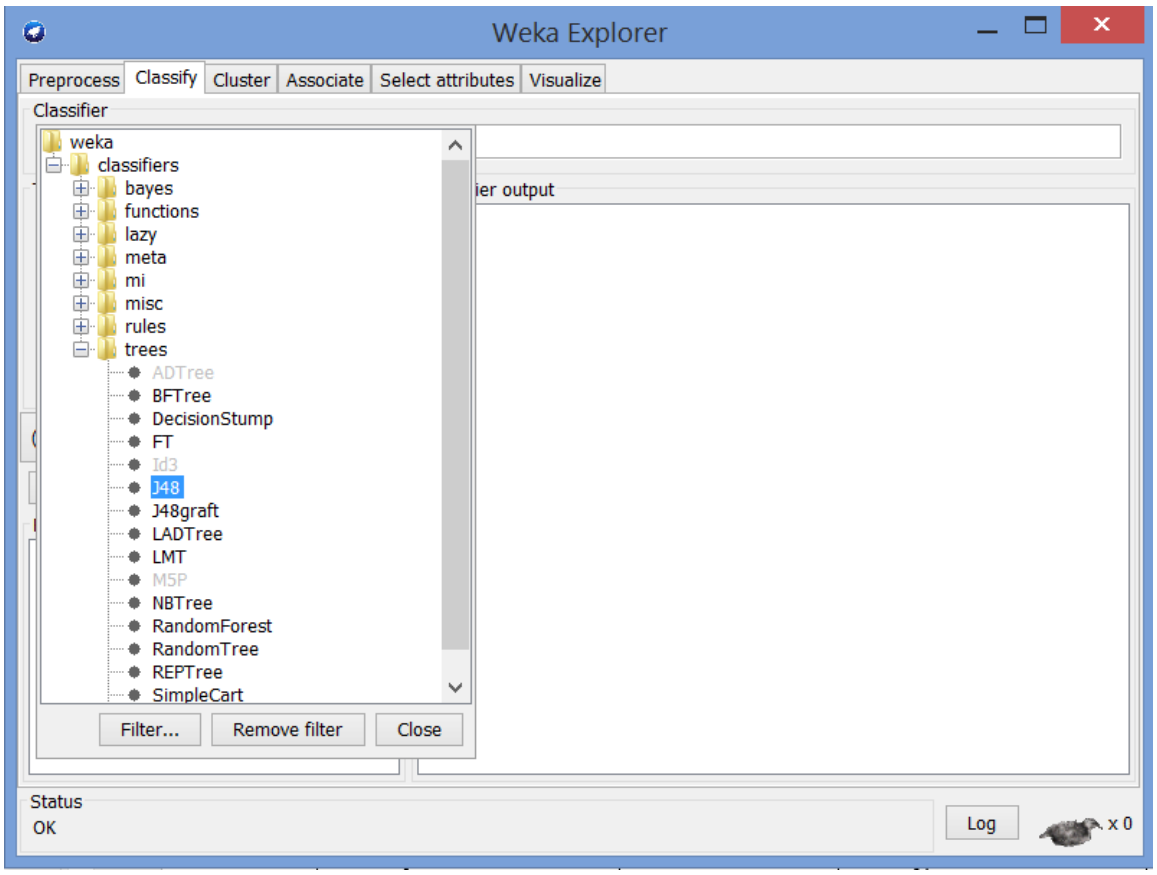
По-нататъшен анализ на структурата на дървото показва разположението по възли, взаимовръзките и относителната важност на останалите фактори за успех на компанията.

3. Извличане на знания от данни и създаване на модел за предсказване на успеха с Weka

С помощта на софтуерния продукт за извличане на знания от данни Weka [5], са създадени модели за предсказване на успеха, които извършват класификация на компаниите - предвиждат в коя категория попадат те: „успешни“, „нито успешни, нито неуспешни“ или „неуспешни“.

Постановка на експеримента

Weka предлага множество алгоритми за създаване на модели за класификация (Фиг. 8), които могат да бъдат избрани и стартирани от графичния потребителски интерфейс.



Фиг. 8: Избор на алгоритми за класификация в Weka

След избор на алгоритъм, се избира цел за модела – променливата, чиято стойност се предсказва, в случая – успеха на компанията. След настройка и изпълнение на алгоритъма, се получава текстова хронология на данните, в която се съдържа информация за модела и неговото синтезиране. Част от получените данни съдържат и прогнозата за точност на модела. Тези действия се повтарят до генериране на желаните модели, които след това се анализират и сравняват.

Сравнение на получените модели за предсказване на успеха

Изпробвани бяха различни алгоритми за класификация, като всички бяха изпълнени с и без прилагане на кръстосана валидация. Резултатите са показани в Таблица 1, където моделите са подредени в низходящ ред според точността при прилагане на кръстосана валидация.

Таблица 1: Сравнение на класификационни модели, получени с Weka

Алгоритъм за класификация	**Тип на модел**	**Точност (без кръстосана валидация)**	**Точност (с кръстосана валидация)**
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%

В таблицата са представени модели, които имат разнороден принцип на работа (запазена е терминологията, използвана в Weka): класификационно дърво (Tree), индукция на правила (Rules), байесов класификатор (Bayes).

От получените класификационни модели, най-висока точност при прилагане на кръстосана валидация има моделът, получен с алгоритъма J48 – 66,67%. Тази точност е получена чрез използване на различни данни за обучение и тестване на модела, и е сравнително реалистична. Точността на модела е като цяло относително ниска, както и по-ниска от точността, постигната, чрез прилагане на софтуерния продукт IBM SPSS Modeler.

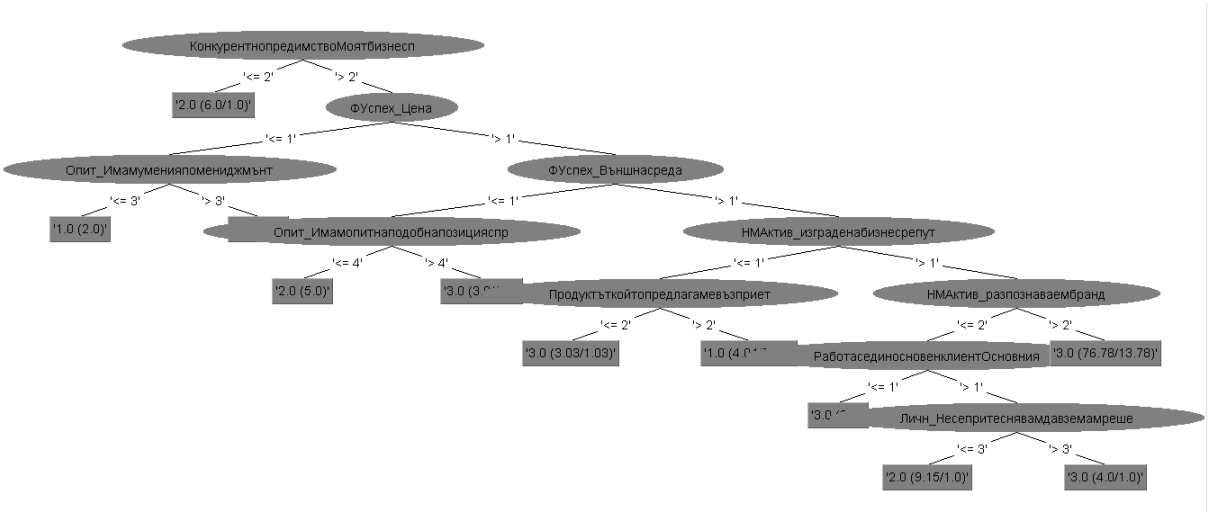
Ще разгледам модела, получен чрез J48 в по-голяма дълбочина с цел идентифициране на факторите, довели до успех на компанията и тяхната

относителна важност. В текстовото представяне на дървото факторите са описани със съкратени имена. Най-важният фактор за успеха на компанията е наличието на ясно конкурентно предимство. Този резултат е консистентен и с резултатите, получени чрез алгоритъма C5.0 в IBM SPSS Modeler. Останалите фактори в модела, генериран чрез Weka и J48 са:

- Конкурентно предимство: Моят бизнес притежава ясно конкурентно предимство.
- Смятам, че следните фактори са ключови за успеха за моя бизнес: Цена
- Смятам, че следните фактори са ключови за успеха за моя бизнес: Външна среда
- Умения и опит: Имам умения по мениджмънт.
- Нематериални активи Компанията притежава изградена бизнес репутация.
- Умения и опит: Имам опит на подобна позиция (спрямо заеманата в момента).
- Нематериални активи: Компанията притежава разпознаваем бранд.
- Продуктът, който предлагам, е възприет от пазара в следната степен.
- Основният източник, използван за първоначално финансиране на фирмата е: Индивидуален инвеститор (бизнес ангел)
- Личност и ценности: Не се притеснявам да вземам решения без да разполагам с необходимата информация.

Част от изброените фактори за успех на компанията да подобни на получените с помощта на IBM SPSS Modeler и алгоритъма C5.0: конкурентно предимство, външната среда като ключов фактор за успеха, изградена бизнес репутация, разпознаваем бранд. Останалите фактори са специфични за модела.

Weka не предлага подробна визуализация на класификационното дърво, за разлика от IBM SPSS Modeler. Има възможност единствено за схематична визуализация (Фиг. 9), която показва възлите на дървото, връзките между тях и съкратено наименование на променливите.



Фиг. 9: Схематична визуализация на класификационното дърво, генерирано с алгоритъм J48 в Weka

4. Заключение

На база на направените анализи, двата най-точни модела за предсказване на успеха на стартиращи компании (при прилагане на метода кръстосана валидация) са:

- класификационно дърво, генерирано с помощта на продукта IBM SPSS Modeler и алгоритъма C5.0 с точност 83,76%,
- класификационно дърво, генерирано с помощта на продукта Weka и J48 с точност 66,76%.

От съпоставката на точността на моделите ясно се вижда, че моделът, получен чрез IBM SPSS Modeler (комерсиален продукт на IBM) е със значително по-голяма точност, от този, получен чрез Weka (безплатен продукт). IBM SPSS Modeler се оказа по-добрият инструмент за решаване на конкретната задача. Като тип, и двата модела представляват класификационно дърво и част от съдържащите се в тях фактори са аналогични. Използваните алгоритми C5.0 и J48 са подобни и се базират на алгоритъма C4.5. Но докато J48 е просто Java имплементация на C4.5, то C5.0 притежава подобрена логика за по-балансирано разделяне на данните, което вероятно е причина за по-добрия резултат.

Предложените модели са имплементирани в бета версията на софтуерния продукт за предсказване на успеха на стартиращи компании I3SP

(Information System for Start-ups Success Prediction). Бъдещите планове включват развитие на софтуерния продукт, събиране на повече данни за стартиращи компании – от България и света и подобрение на моделите за предсказване на успеха.

Изследването е осъществено с помощта на Европейския социален фонд чрез Оперативна програма „Развитие на човешките ресурси”, договор № BG051PO001-3.3.06 - 0052 (2012-2014). Изследването е осъществявено с помощта на IBS Bulgaria, IBM Premier Business Partner.

Използвана литература

1. EuropeanCommision. (2014). HORIZON 2020, The New EU Framework Programme for Research and Innovation 2014-2020.
2. Yankov, B. (2013) A Model for Predicting the Success of New Ventures, Vth International Scientific Conference e-Governance, ISSN 1313-8774, (стр. 128-135)
3. IBM Corporation. (2012). IBM SPSS Modeler 15 User’s Guide. IBM Corporation.
4. Ruskov, P. H. (2012). Online Investigation of SMEs Competitive Advantage. MEB 2012, 10th International Conference on Management, Enterprise and Benchmarking, (стр. 143-159).
5. Hall, M. F. (Volume 11, Issue 1 2009 r.). The WEKA Data Mining Software: An Update. SIGKDD Explorations.