



Модел за прогнозиране на смъртността, причинена от сърдечна недостатъчност

Светлин Маринов

Model for Predicting Heart Failure Mortality

Svetlin Marinov

Abstract: *The aim of this paper is to create a model for predicting heart failure mortality using the decision tree method, and to apply it to a specific set of information using the RapidMiner data analysis platform. Due to the rapid pace of medical development, there is an urgent need to provide recommendations derived from filtering the entire set of medical data. It becomes difficult for patients to make the right decision due to the huge variety of treatments and ways to improve their lifestyle. This necessitates the emergence of predictive methods that aim to offer the patient the most accurate choice according to specific complaints instead of wandering among multiple random choices with varying degrees of significance in reality. Event prediction methods extract the most vital information from a huge data set, “reading” the patient’s problems and suggesting the most appropriate treatment for him/her. Most patients are late in taking the necessary care of their own health, and they realize the importance of prevention at a later stage. If every person could predict what might happen to him/her later in life, he/she would be much more cautious and make more timely efforts for his/her health. The need for decision-making in clinical practice often has important long-term consequences [1].*

Keywords: *tree; medicine; method; patient.*

1. Въведение

Сърдечно-съдовите заболявания (ССЗ) са най-често срещаната причина за смърт в световен мащаб, като отнемат около 17,9 милиона живота всяка година, което представлява 31% от всички смъртни случаи в света.

Сърдечната недостатъчност е често срещано събитие, причинено от ССЗ и този набор от данни съдържа 12 характеристики, които могат да се използват за прогнозиране на смъртност от сърдечна недостатъчност.

Повечето сърдечно-съдови заболявания могат да бъдат предотвратени чрез справяне с поведенчески рискови фактори като употреба на тютюн, нездравословна диета и затлъстяване, физическа неактивност и прекомерна употреба на алкохол, като се използват стратегии за цялото население.

Хората със сърдечно-съдови заболявания или тези, които са с висок сърдечно-съдов риск (поради наличието на един или повече рискови фактори като хипертония, диабет, хиперлипидемия или вече установено заболяване) се нуждаят от ранно откриване и управление, при което моделът за машинно обучение може да бъде от голяма помощ. [6]

Сърдечно-съдовите заболявания са водеща причина за заболяемост и смъртност в световен мащаб. Проблемът е особено актуален за нашата страна, като данните сочат, че в България 2 от 3-ма българи умират от сърдечно-съдови заболявания (миокарден инфаркт, мозъчен инсулт, сърдечна недостатъчност). Водещо е значението на атеросклерозата като основна причина за тяхната поява, развитие и настъпващи усложнения. [3]

С напредъка на медицината, голямото количество данни изисква механизми за ефективно и бързо актуализиране на информацията. Една от техниките, използвани за справяне с този проблем, се нарича дърво на решенията. Идеята за използването на методи идва от нагласата, че хората често не получават най-точните и най-бързите препоръки от медицинските лица. Методите за предсказване на събитие изследват техники за съпоставяне на хора със сходни проблеми и отправяне на препоръки на тази основа. Методите за предсказване на събитие имат нужда от участие на пациенти и лекари, които могат да диагностицират хора със сходни проблеми. При всеки здравословен проблем хората се допитват до медицинско лице за точна диагноза. Препоръчителните системи, използващи методи за предсказване на събитие, са софтуерни инструменти и техники, помагачи на пациента да прецени какво може да подобри и към какъв специалист да се обърне. Тези методи улесняват вземането на решения, като разчитат на наличната вече информация.

2. Изложение

2.1 Преглед на подобни разработки

Ще разгледаме разработките на Мей Шоуман, Тим Търнър и Роб Стокър от Факултета по инженерство и информационни технологии, университет на Нов Южен Уелс в Австралийската академия на отбранителните сили и на Мехди Иляс и Имран Шаик от Института по управленски науки Алана. В първата се прави разработка на използването на дърво на решенията според възраст, пол, вид болка в гърдите, кръвно налягане в покой, серумен холестерол, кръвна захар на гладно, максимален пулс и други, въз основа, на които се поставя диагноза – здрав или пациент, който е обект на възможни сърдечни заболявания. В набора от данни има 76 необработени атрибута. Въпреки това, всички от публикуваните експерименти се отнасят само до 13 от тях. Следователно, за да позволим сравнение с литературата, ние ограничено тестване до същите тези атрибути. Наборът от данни съдържа 303 реда, от които 297 са завършени. Шест реда съдържат липсващи стойности и те са отстранени от експеримента. Според разработката, сърдечните заболявания са водещата причина за смърт в света през последните 10 години. Изследователите са използвали няколко техники за извличане на данни в помощ на здравеопазването, в диагностиката на сърдечни заболявания. Дърво на решенията е една от използваните успешни техники за извличане на данни. [10]

Темата на втората статия е еднаква с първата. Разгледани са възраст, пол, болка в гърдите, анемия, серумния холестерол, максимален пулс и други, които спомагат за диагнозата. Тази система превзе Кливланд данни за обучение от UCI хранилище (т.е. Университета на Калифорния, Ървайн). Базата данни на Кливланд се състои от 303 записа. Таблото за управление е създадено с помощта на Power BI. Ние използваме дървото на решенията за прогнозиране на сърдечни заболявания. Предоставени са данни от Кливланд към модела на дървото на решенията, след което създава дървото на решенията с различен метод като 10-кратно кръстосано валидиране и Процентно разделяне. Но резултатите са еднакви и за двата метода ние използваме процентно разделяне с кръстосано съгване валидиране 20 и нивото на доверие е 0,25%. и ние ще разработим а система, използваща дървото на решенията, тази система взема данни за лица, и да се предвиди дали лицето има сърдечно заболяване или не. Според тях прогнозирането на сърдечно заболяване в ранен стадий може да намали до известна степен коефициента на смъртност. Техният софтуер помага за прогнозиране на сърдечни заболявания в ранен стадий. Днес здравните организации генерират огромно количество данни, но те са силно неорганизирани. Ако тези данни се орга-

низират по подходящ начин, използвайки техника за извличане на данни, която може лесно да прогнозира сърдечни заболявания. [5]

Разработка в [8] сравнява ефективността на методите за машинно обучение, особено Naive Bayes, Decision Tree, Random Forest и ID3 за задача за класификация (т.е. класифициране на самолети в определена категория и в страна на произход) с помощта на инструмента RapidMiner. Тези методи се сравняват въз основа на тяхната точност, проценти на грешки, прецизност и извикване за класифициране на самолети. Резултатите разкриват, че алгоритмите Random Forest и ID3 дават добра точност на класификация поради естеството им, което прогресивно се подобрява, освен от дървото на решенията. Моделът доказва по-добрите резултати и помага на експертите в областта и дори на лицето, свързано с медицинската област, да планират по-добра и ранна диагноза за пациента. Тази система се представя реалистично добре дори без преквалификация.

Има области на науката, публикации, които се развиват динамично във времето. В настоящата разработка не са използвани излишни и неразбираеми данни, които водят до тромавост и отекчителност на речта. Точността е в такава степен, която позволява достъпност и дава възможност да се надградят знания и да се предизвика любознателност по темата. Използвани са съвременни източници с актуални данни и са представени нови открития и достатъчно примери. Подбрани са много изображения и фигури с описание към тях. Включени са конкретни данни за българското население.

2.2 Представяне на изследването

В настоящата статия е изследвано създаването на модел за прогнозиране на смъртността, причинена от сърдечна недостатъчност с помощта на метода дърво на решенията. Изграденият модел е приложен за конкретен набор от данни чрез платформата за анализ на данни RapidMiner и са анализирани получените резултати.

Машинното обучение е важна техника, която помага на компании, организации и отделни лица да подобрят качеството на вземане на решения. В днешния сценарий, особено с появата на науката за данни, може да се види как техниките за машинно обучение се използват за анализ на данни. Има различни техники за машинно обучение за задачи на науката за данни, които могат да бъдат категоризирани, както следва: класификация, прогнозиране, регресия, анализ на асоциациите, групиране, прогнозиране на времеви редове и много други. Тъй като има много различни безплатни инструменти за машинно обучение, изборът на подходяща техника за анализ е от решаващо значение за решаване на проблем. [8]

1.3 Подготовка на данните

Наборът от данни, с които ще работим, са предоставяни свободно от Kaggle - дъщерно дружество на Google LLC, онлайн общност от учени, практикуващи машинно обучение и данни. Наборът съдържа данни за над 300 пациенти с тяхната възраст, пол и предшестващи заболявания. Изграждането на цялостния процес в платформата минава през характерните за добиването на знания от данни стъпки: съдържа подготовка на данни, моделиране, прилагане на модела към тестов набор и стъпки за оценка на производителността. [7]

Импортираме първият набор (`heart_failure_clinical_records_dataset.csv`) с помощта на Import Configuration Wizard. Той съдържа следната информация:

Import Data - Format your columns. ×

Format your columns.

Date format Replace errors with missing values ⓘ

	age <i>integer</i>	anaemia <i>integer</i>	creatinine_... <i>integer</i>	diabetes <i>integer</i>	ejection_fr... <i>integer</i>	high_blood... <i>integer</i>	platelets <i>real</i>	serum_cre... <i>real</i>
1	75	0	582	0	20	1	265000.000	1.900
2	55	0	7861	0	38	0	263358.030	1.100
3	65	0	146	0	20	0	162000.000	1.300
4	50	1	111	0	20	0	210000.000	1.900
5	65	1	160	1	20	0	327000.000	2.700

Фиг.1 Импортиране на heart_failure_clinical_records_dataset.csv в RapidMiner.

1. Възраст (age): Годишите на пациента.
2. Анемия (anemia): Намалване на червените кръвни клетки или хемоглобина.
3. Креатининова фосфокиназа (creatinine phosphokinase): Ниво на ензима СРК в кръвта (mcg/L).
4. Диабет (diabetes): Ако пациентът има диабет.
5. Фракция на изтласкване (ejection_fraction): Процент на кръвта, напускаща сърцето при всяко съкращение (в проценти).
6. Високо кръвно налягане (high blood pressure): Ако пациентът има хипертония (булева).
7. Тромбоцити (platelets): Тромбоцити в кръвта (kiloplatelets/mL).
8. Серумен креатинин (serum creatinine): Ниво на серумния креатинин в кръвта (mg/dl).
9. Серумен натрий (serum sodium): Ниво на серумния натрий в кръвта (mEq/L).
10. Пол (sex): Жена или мъж (булева).

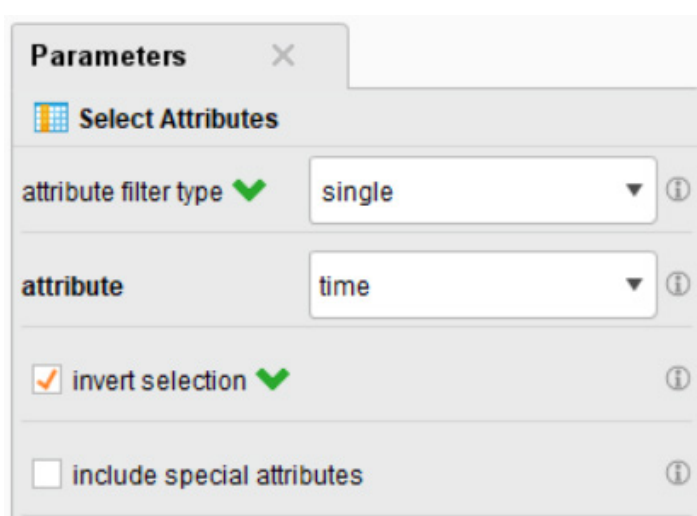
Добавяме данните с оператора Retrieve. Този оператор осъществява достъп до съхранената информация в хранилището и ще я зареди в процеса, който изграждаме.

Изглед на заредените данни в RapidMiner:

Name	Type	Missing	Statistics	Filter (13 / 13 attributes):
age	Integer	0	Min 40, Max 95, Average 60.836	
anaemia	Integer	0	Min 0, Max 1, Average 0.431	
creatinine_phosphokinase	Integer	0	Min 23, Max 7861, Average 581.839	
diabetes	Integer	0	Min 0, Max 1, Average 0.418	
ejection_fraction	Integer	0	Min 14, Max 80, Average 38.084	
high_blood_pressure	Integer	0	Min 0, Max 1, Average 0.351	
platelets	Real	0	Min 25100, Max 850000, Average 263358.029	
serum_creatinine	Real	0	Min 0.500, Max 9.400, Average 1.394	
serum_sodium	Integer	0	Min 113, Max 148, Average 136.625	

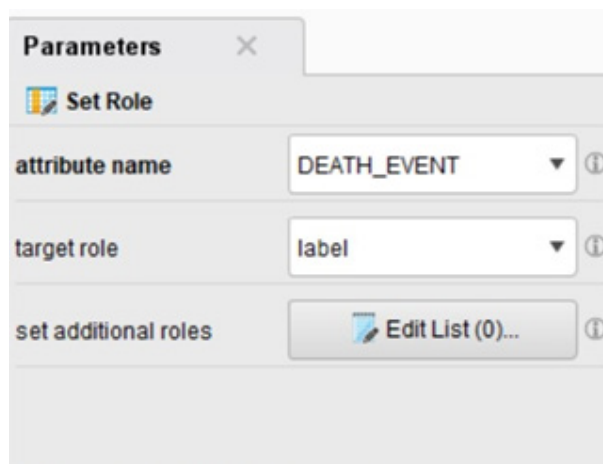
Фиг.2 Изглед на заредените данни към момента в RapidMiner.

Тъй като атрибутът “time” не ни е необходим, следва да го премахнем, за да не се появява в крайните резултати. Използваме оператора Select Attributes, и поставяме отментка на полето “invert selection”, за да премахнем само “time” от селектираните данни:



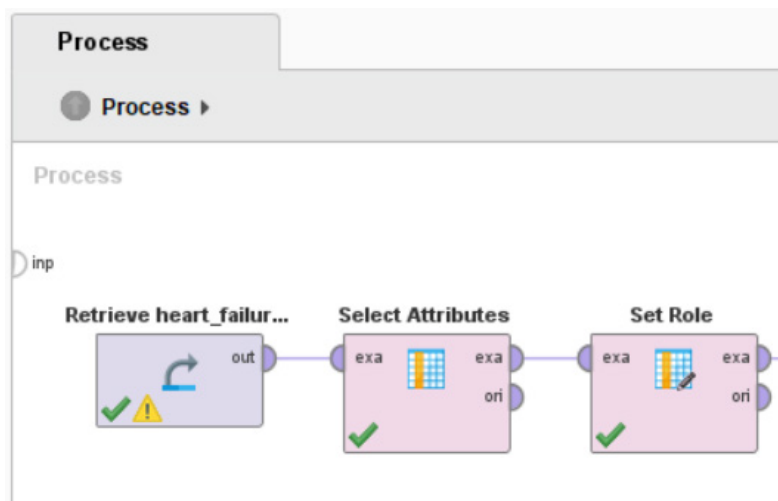
Фиг.3 Визуализация на параметрите на оператора Select Attributes.

Пристъпваме към операторът Set role, който се използва, за да декларира кой атрибут е етикета. Атрибут с ролята label действа като целеви атрибут за изучаване на операторите. Етикетът често се нарича също „целева променлива“ или „клас“.



Фиг.4 Визуализация на параметрите на оператора Set role.

С това, стъпката по подготовка на данните приключва и преминаваме към изграждане на самия модел.



Фиг.5 Приключване на фазата по подготовка на данните.

1.4 Изграждане на модела

В основата на изграждането на модела са операторите за препоръки на платформата, които могат да се свалят безплатно през програмата от Marketplace > Recommender extension, тъй като не са инсталирани по начало. Това разширение на препоръчителната система съдържа оператори, които са подходящи за типични задачи за препоръки: филтриране на информация за конкретния потребител чрез предсказване на най-вероятните оценки за елементите на някои DMR (задача за прогнозиране на рейтинга) или класиране на невидими елементи от потребителя според прогнозните топ-елемента (Top-N Items). В този случай използваме метода Decision tree.

Дървото на решенията е съвкупност от възли, предназначена за създаване на решение за принадлежност на стойностите към даден клас или за оценка на числена целева стойност. Всеки

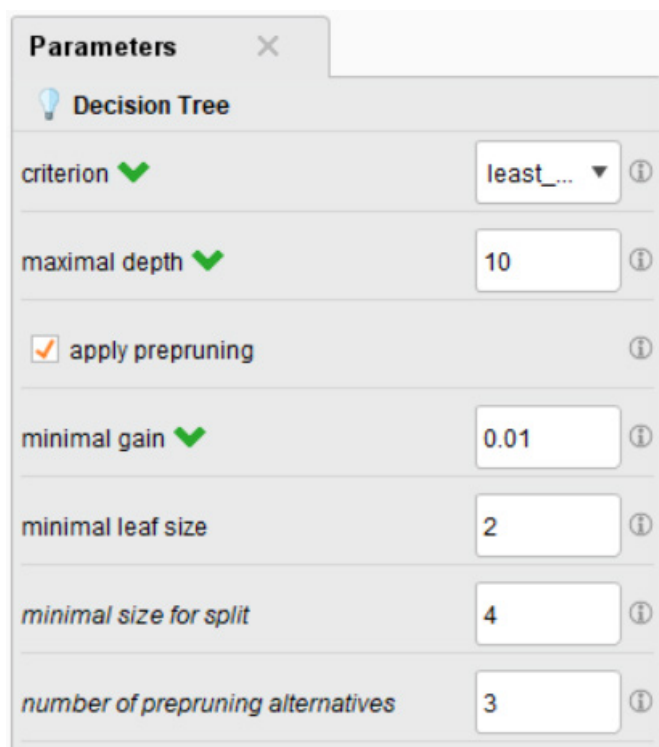
възел представлява правило за разделяне на един конкретен атрибут. За класификация това правило разделя стойностите, принадлежащи към различни класове, а за регресия ги разделя, за да се намали грешката по оптимален начин за избрания критерий на параметъра.

Изграждането на нови възли се повтаря, докато не бъдат изпълнени критериите за спиране. Прогнозата за етикета на класа Атрибут се определя в зависимост от мнозинството от примерите, достигнали до този лист по време на генерирането, докато оценката за числова стойност се получава чрез осредняване на стойностите в листа.

Този оператор може да обработва набори от примери, съдържащи както номинални, така и числови атрибути. Атрибутът на етикета трябва да бъде номинален за класификация и цифров за регресия.

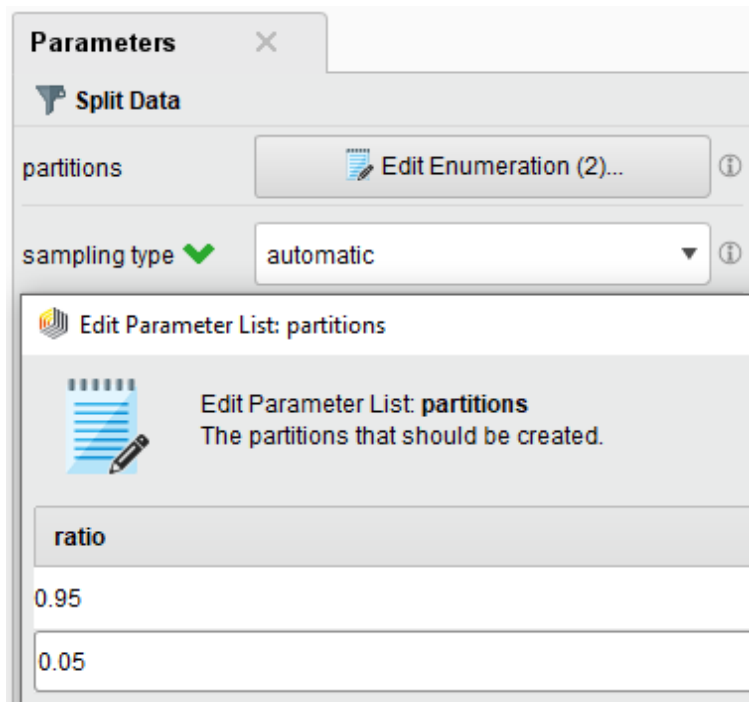
След генерирането моделът на дървото на решенията може да бъде приложен към нови примери с помощта на оператора Apply Model (Прилагане на модела). Всеки пример следва разклоненията на дървото в съответствие с правилото за разделяне, докато се достигне до лист. [11]

Като метод за съвместно филтриране, единственият вход, необходим за оператора за моделиране, е матрицата на рейтингите. Матрицата с рейтинги е под формата на идентификация на потребителя, идентификация на артикул и етикет (рейтинг), които зададохме с оператора Set Role. Всеки друг атрибут от набора от данни не се използва в основния оператор за моделиране. Независимо от формата на входните данни, той трябва да бъде трансформиран в набор от данни с идентификация на потребителя, идентификация на артикула и цифров етикет за оценка. Има няколко параметъра, които трябва да конфигурираме при Decision tree:



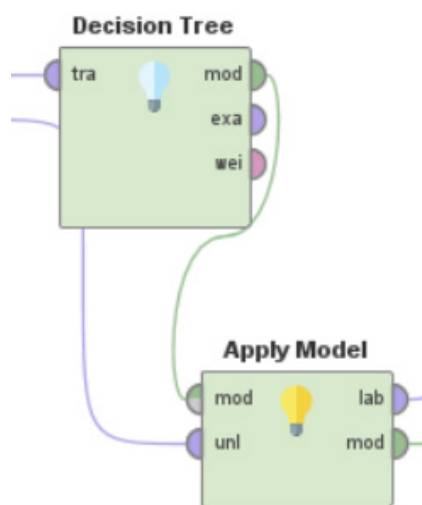
Фиг.6 Визуализация на параметрите на оператора Decision tree.

Избираме за данните за оценките да се разделят на случаен принцип между данните за обучение (70%) и тестовия набор от данни (30%), като се използва операторът за разделяне на данни Split data.

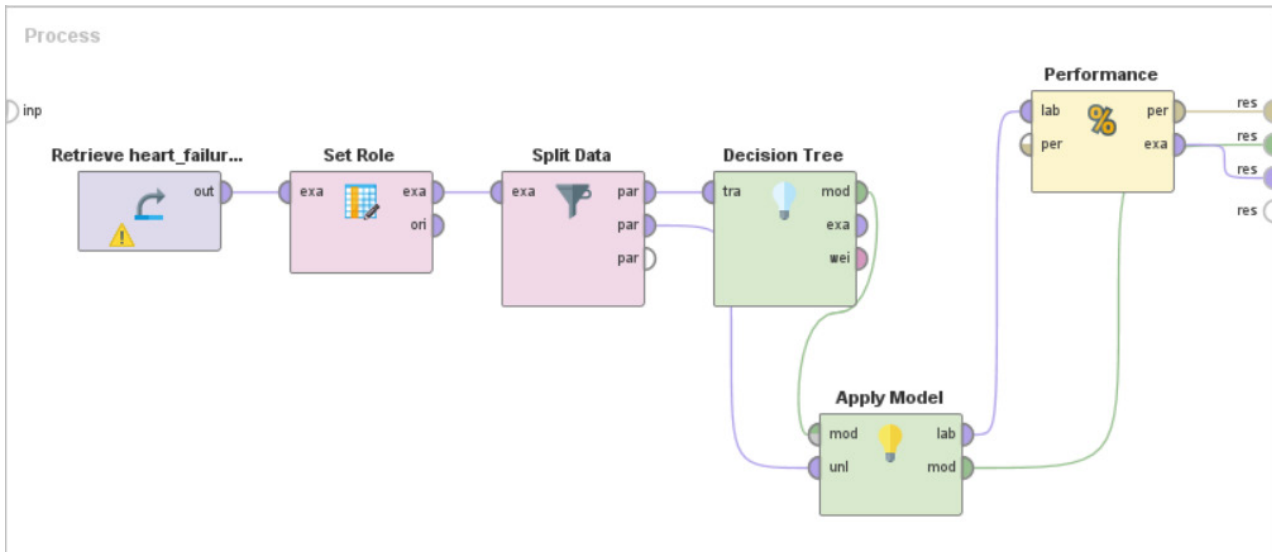


Фиг.7 Визуализация на параметрите на оператора Split data.

За да приложим обучителния модел към тестовия набор от данни, следва да използваме оператора Apply Model (Rating Prediction). Част от оригиналния набор от данни, който е запазен за тестване, се използва като вход за оператора Apply Model, който изчислява прогнозираните оценки за всяка комбинация от потребителски елементи в тестовия набор от данни. Операторът Apply Model връща примерен набор, съдържащ първите n-класирани препоръки за всеки потребител. Приложимият модел се използва и при окончателното внедряване, където може да се предоставят комбинации от потребителски елементи, за които рейтингът трябва да бъде предвиден. С това приключва стъпката по изграждане на модела и неговото прилагане.

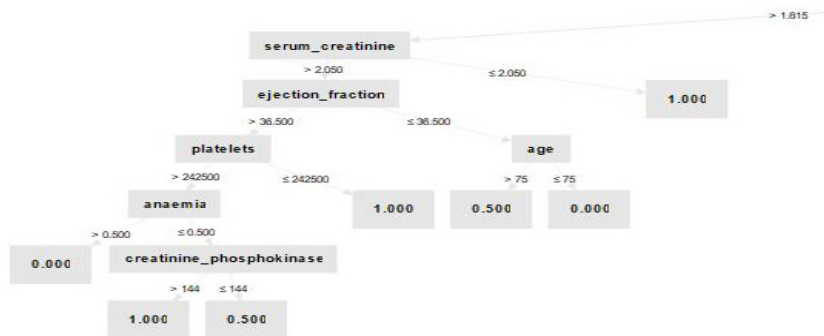
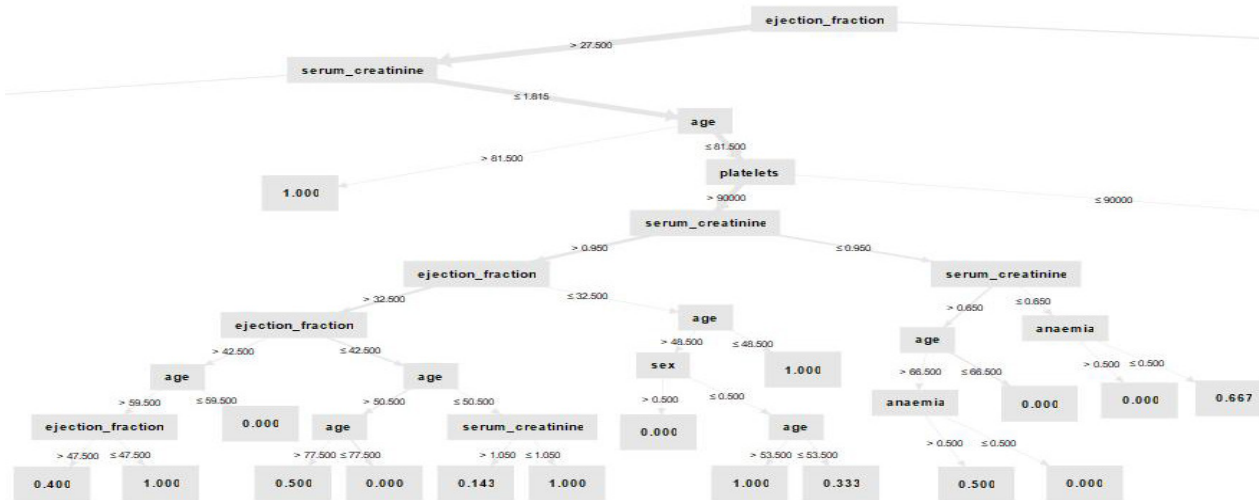


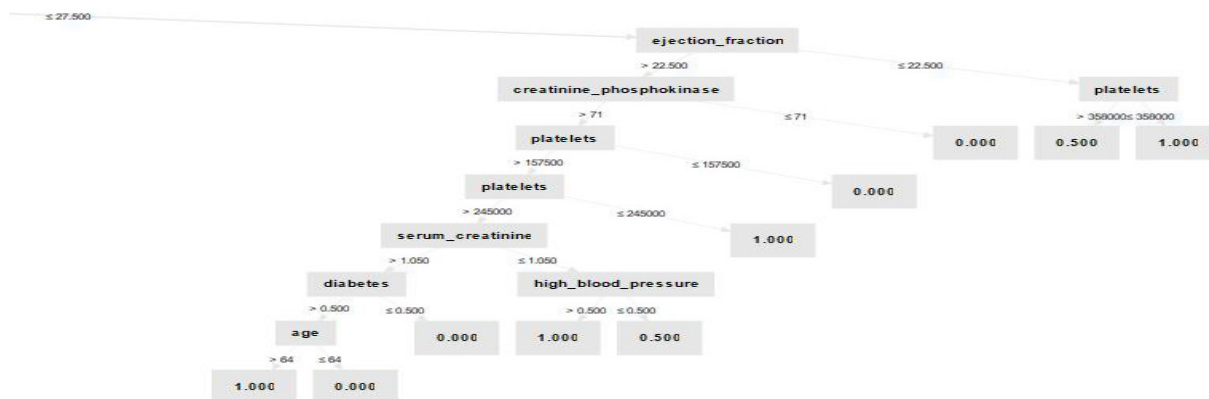
Фиг.8 Приключване на фазата по изграждане на модела за препоръки и неговото прилагане.



Фиг.9 Окончателна визуализация на използваните оператори в RapidMiner.

1.5 Експериментални резултати





Фиг.10,11,12 Резултати от програмата.

Започваме от измерване на обема на кръвта напускаща сърцето при всяко съкращение, което се измерва в проценти. Ако е над 27.5% и серумния креатин е между 1.815(mg/dl) и 2.050(mg/dl) шансът за сърдечен удар е 100%. Ако серумният креатин е над 2.050(mg/dl) и кръвта, напускаща сърцето е под 36.5%, при възраст по-малка или равна на 75 години, шансът за сърдечен удар е 0%, а ако възрастта е над 75 години – шансът е 50%.

При процент на кръвта, напускаща сърцето над 36.5% и съдържание на тромбоцити в кръвта под 242.500(kiloplatelets/mL), шансът за сърдечен удар е 100%. При по-малко от 242.500(kiloplatelets/mL) тромбоцити и анемия със стойност 1, шансът е 0%. При наличие на всички описани по-горе данни, съвместно с наличие на анемия със стойност 0, има два варианта. Първият е, когато нивото на ензима СРК в кръвта е по-малък или равен на 144(mcg/L), шансът се равнява на 50%, но ако е по-голям – шансът се увеличава до 100%.

Връщаме се в началото – при над 27.5% кръв, напускаща сърцето и серумен креатин по-малък или равен на 1.815(mg/dl), при възраст на пациента над 81 години, шансът за получаване на сърдечен удар е 100%. Но ако пациентът е под 81 години и нивото на тромбоцитите в кръвта е над 91 000(kiloplatelets/mL) има 11 варианта на развитие. При ниво на серумния креатин над 0.950(mg/dl) са налични 9 от 11-те варианта:

Първият вариант е, когато процентът на кръвта напускаща сърцето е над 42.5%, а пациентът е под 60 години, шансът за получаване на сърдечен удар е 0%. Но, когато възрастта на пациента е над 60 години и кръвта напускаща сърцето е до 47.5%, шансът се увеличава до 100%. При наличие на кръв, напускаща сърцето над 47.5%, шансът за сърдечен удар се намалява до 40%.

Ако кръвта, напускаща сърцето е над 42.5% и пациентът е под 50 години, ако серумният креатин е под 1.050(mg/dl), шансът е 100%. Ако серумният креатин е над тази стойност, шансът се смалява до 14.3%.

При 42.5% кръв, напускаща сърцето и пациент на възраст между 50 и 70 години, шансът за получаването на сърдечен удар е 0%. Ако възрастта му е над 77 години, шансът се увеличава до 50%.

Връщаме се на случая, в който кръвта напускаща сърцето е под 32.5% и пациент под 48 години, шансът за получаването на сърдечен удар е 100%. При мъжете под 48 години, шансът е 0%, а при жените между 48 и 53 години, шансът е 33%. При жени над 53 години, шансът се увеличава до 100%. Това бяха 11-те възможни варианта, при които серумният креатин беше над 0.950(mg/dl).

Сега ще разгледаме възможните варианти под 0.950(mg/dl), които са 5 на брой:

Ако серумният креатин е между 0.950(mg/dl) и 0.650(mg/dl) и пациентът е под 66 години, шансът се равнява на 0%. Но ако пациентът е над 66 години има два изхода:

Ако не е налична анемия, шансът е 0%. Ако е налична анемия, шансът се увеличава на 50%. Ако серумният креатин е по-малък или равен на 0.650(mg/dl) и е налична анемия, шансът е равен на 0%. Ако не е налична анемия, шансът става 67%.

Връщаме се в началото, ако обема на кръвта напускаща сърцето при всяко съкращение е под 27.5% имаме 10 варианта. Ако кръвта напускаща сърцето е и под 22.5%, гледаме съдържанието на тромбоцити в кръвта е над 358000(kiloplatelets/mL) шансът е 50%, а ако е над тази стойност 100%. Ако процента на кръвта напускаща сърцето при всяко съкращение е между 22.5% и 27.5%, ще разгледаме останалите 8 варианта. Ако креатиновата фосфокиназа е под 71(mcg/L) шансът е 0%. Ако е над 71(mcg/L), гледаме съдържанието на тромбоцитите, ако е по-малко от 157500(kiloplatelets/mL) шансът е 0%, а ако е между 15700(kiloplatelets/mL) и 245000(kiloplatelets/mL), шансът е 100%. Ако съдържанието е над 245000(kiloplatelets/mL), проследяваме серумния креатин, който е по-малък или равен на 1.050(mg/dl) и кръвното налягане е под 0.500 шансът е 50%, а ако е под тази стойност шансът е 100%. Ако серумният креатин е над 1.050(mg/dl) и диабета е по-малък или равен на 0.500, шансът е 0%, а ако е по-голям и пациента е над 64 години шансът е 100%, а ако е под 0%.

PerformanceVector

```
PerformanceVector:  
root_mean_squared_error: 0.540 +/- 0.000  
squared_error: 0.291 +/- 0.432
```

Средноквадратното отклонение (RMSE) или средноквадратната грешка (RMSE) е често използвана мярка за разликите между стойностите (стойности на извадката или популацията), предвидени от модел или оценител, и наблюдаваните стойности. RMSE представлява корен квадратен от втория момент на извадката на разликите между прогнозираните стойности и наблюдаваните стойности или средноквадратичната стойност на тези разлики. Тези отклонения се наричат остатъци, когато изчисленията се извършват върху извадката от данни, която е била използвана за оценка, и се наричат грешки (или грешки при прогнозиране), когато са изчислени извън извадката. RMSE служи за агрегиране на величините на грешките в прогнозите за различни точки от данни в единна мярка за предсказваща сила. RMSE е мярка за точност, за да се сравняват грешките при прогнозиране на различни модели за конкретен набор от данни, а не между набори от данни, тъй като зависи от мащаба. [4]

В статистиката средната квадратична грешка (MSE) или средноквадратното отклонение (MSD) на оценител (на процедура за оценка на ненаблюдавана величина) измерва средната стойност на квадратите на грешките – тоест средната квадратна разлика между прогнозните стойности и действителната стойност. MSE е рискова функция, съответстваща на очакваната стойност на квадратната загуба на грешка. Фактът, че MSE е почти винаги строго положителен (а не нула), се дължи на случайност или защото оценителят не отчита информация, която би могла да даде по-точна оценка. При машинното обучение, по-специално минимизирането на емпиричния риск, MSE може да се отнася до емпиричния риск (средната загуба на наблюдаван набор от данни), като оценка на истинския MSE (истинския риск: средната загуба при действителното разпределение на населението).

MSE е мярка за качеството на оценителя. Тъй като се извлича от квадрата на евклидовото разстояние, тя винаги е положителна стойност, която намалява, когато грешката се доближи до нула.

MSE е вторият момент (относно произхода) на грешката и по този начин включва както дисперсията на оценителя (колко широко разпространени са оценките от една извадка от данни в друга), така и нейното отклонение (колко далеч е средната изчислена стойност от истинската стойност). MSE е дисперсията на оценителя. Подобно на дисперсията, MSE има същите мерни единици като квадрата на количеството, което се оценява. По аналогия със стандартното отклонение, вземането на корен квадратен от MSE дава средноквадратната грешка или средноквадратното отклонение (RMSE или RMSD), което има същите единици като количеството, което се оценява; за безпристрастен оценител, RMSE е корен квадратен от дисперсията, известен като стандартна грешка. [2]

Въз основа на практическо правило може да се каже, че стойностите на RMSE между 0,2 и 0,5 показват, че моделът може относително точно да предскаже данните. В допълнение, коригиран R-квадрат (MSE) повече от 0,75 е много добра стойност за показване на точността. В някои случаи също е приемлив коригиран R-квадрат (MSE) от 0,4 или повече. В тази разработка стойността на RMSE е 0,540, което е приемлива стойност, стойността на MSE трябва да се коригира, защото не трябва да има отрицателна стойност, а положителната е твърде ниска.

3. Заключение

Не е възможно да се направят конкретни заключения от резултатите от този доклад. Тезата да може да създаде ранен скрининг. Диагностична помощ се прави само с атрибути, които пациентът може да извлече без напреднал медицински преглед. Следователно крайният резултат не може да бъде доказан, нито опроверган. Резултатите са неубедителни поради риска от пренастройване, липса на атрибути и твърде малко количество пациенти. Факторът с най-голямо въздействие върху експеримента е липсата на достатъчно данни. Въпреки че могат да се направят конкретни заключения върху моментните резултати и наблюдения, генерираната информация в този доклад насърчава дискусиата и по-изчерпателните изследвания за по-нататъшно разследване на този въпрос. Добивните техники от това проучване трябваше да бъдат приложени към по-голяма база данни на пациенти. При най-малко три пъти повече пациенти точността ще бъде положително повлияна. От това проучване виждаме потенциал за това, домашният скрийнинг да стане общ инструмент и да помогне за намаляване на смъртността поради сърдечни заболявания в бъдеще. [9]

Анализът на решенията е инструмент, който позволява на потребителите да прилагат основана на доказателства медицина, за да вземат информирани и обективни клинични решения, когато са изправени пред сложни ситуации. Дървото на решенията, заедно с извлечените от литературата вероятности и дефинираните стойности на резултатите, се използва за моделиране на даден проблем и помагат да се определи най-добрият курс на действие. Анализът на чувствителността позволява изследване на важни променливи за крайните клинични резултати. След това лицето, което взема решение, може да установи предпочитан метод на лечение и да проучи променливи, които влияят на крайния резултат. Настоящият документ има за цел да направи преглед на анализа на решенията и неговото приложение при вземането на клинични решения.

Една от следващите стъпки за подобряване на процеса е допълване на данните, за да се получи по-точен краен резултат и коефициентите на RMSE и MSE да се подобрят. Този модел може да се използва и върху други данни с подобно медицинско значение. Най – оптималният вариант, който може достигне до повече хора е да се разработи уеб приложение по метода на тази разработка, в което само да се добавят набори от данни от медицински лица, а дори и от други потребители и то да показва резултата заедно с коефициентите на точност. Все пак то няма да дава диагнози и всеки човек трябва да се консултира с лекар, но с него ще може да се дават насоки за превенции.

ЛИТЕРАТУРА

- [1] Piyas A., Schemitsch, E., and Hanson, B. 2008. “What is a Clinical Decision Analysis Study?” // *Indian Journal of Orthopaedics*, vol. 42, no. 2, Apr.–June 2008, pp. 137–139.
- [2] Bermejo, S., and Cabestany, J. 2001. “Oriented Principal Component Analysis for Large Margin Classifiers.” // *Neural Networks*, vol. 14, no. 10, Dec. 2001, pp. 1447–1461.
- [3] Assoc. Prof. Dr. Bozhidar Finkov. 2015. Cardiovascular Diseases (Mar. 2015). Retrieved 20 Mar. 2015 from www.mh.government.bg/.
- [4] Hyndman, R., and Koehler, A. 2006. “Another Look at Measures of Forecast Accuracy.” // *International Journal of Forecasting*, vol. 22, no. 4, Oct.–Dec. 2006, pp. 679–688.

- [5] **Piyas, M. K., and Shaikh, I. S.** 2021. Prediction of Heart Disease Using Decision Tree, MIT Research Lab Technical Report. Faculty, MCA Department Allana Institute of Management Sciences Pune.
- [6] **Dr Zsuzsanna Jakab.** 2021. Cardiovascular Diseases (CVDs), (June 2021). Retrieved from https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.
- [7] **Lardinois, F., Lynley, M., and Mannes, J.** 2017. Google is Acquiring Data Science Community Kaggle (Mar. 2017). Retrieved from <https://techcrunch.com/>.
- [8] **Marzukhi, S. et al.** 2021. RapidMiner and Machine Learning Techniques for Classifying Aircraft Data (June 2021), pp. 1–5.
- [9] **Olsson, A., and Nordlöf, D.** 2015. Early Screening Diagnostic Aid for Heart Disease Using Data Mining, PhD Dissertation. KTH Computer Science and Communication.
- [10] **Shouman, M., Turner, T., and Stocker, R.** 2011. Using Decision Tree for Diagnosing Heart Disease Patients. – In: Proceedings of the 9th Australasian Data Mining Conference (AusDM'11), Dec. 2011, pp. 23–29.
- [11] **Singh, N.** 2022. Decision Tree Algorithm (Dec. 2022). Retrieved 9 Feb. 2022 from www.kdnuggets.com.

ИНФОРМАЦИЯ ЗА АВТОРА

Светлин Маринов – докторант, специалност „Информатика и компютърни науки“, Факултет „Математика и информатика“, Великотърновски университет „Св. св. Кирил и Методий“, e-mail: svetlinmarinov9@gmail.com

ABOUT THE AUTHOR

Svetlin Marinov – PhD student in Informatics, Faculty of Mathematics and Informatics, St. Cyril and St. Methodius University of Veliko Tarnovo, e-mail: svetlinmarinov9@gmail.com