ŽILINSKÁ UNIVERZITA V ŽILINE

ELEKTROTECHNICKÁ FAKULTA

DIPLOMOVÁ PRÁCA

BC, ROMAN MICHALÍK

Techniky dolovania dát pomocou Rapid Miner

Vedúci práce: prof. Ing. Aleš Janota, PhD.

Evidenčné číslo: 28260220182022

Žilina, 2018

ŽILINSKÁ UNIVERZITA V ŽILINE

ELEKTROTECHNICKÁ FAKULTA

DIPLOMOVÁ PRÁCA

BC, ROMAN MICHALÍK

Techniky dolovania dát pomocou Rapid Miner

Študijný program: Aplikovaná telematika
Študijný odbor: 5.2.14 Automatizácia
Školiace pracovisko: Žilinská univerzita v Žiline, Elektrotechnická fakulta,
Katedra riadiacich a informačných systémov
Vedúci diplomovej práce: prof. Ing. Aleš Janota, PhD.

Žilina, 2018

ŽILINSKÁ UNIVERZITAV ŽILINE
Elektrotechnická fakulta
Katedra riadiacich a informačných systémov

Akademický rok 2017/2018 Evidenčné číslo: 28260220182022

ZADANIE DIPLOMOVEJ PRÁCE

Meno diplomanta: Bc. Roman MICHALÍK

Študijný odbor: Automatizácia

Téma diplomovej práce: Techniky dolovania dát pomocou Rapid Miner

Pokyny pre vypracovanie diplomovej práce:

1. Štúdium metód dolovania dát a možností softvérového nástroja.
2. Vypracovanie vzorových metodických postupov pre jednotlivé metódy v zvolenom SW prostredí (vstupné požiadavky, postup, čiastkové operácie, formát výstupu, vlastnosti).
3. Praktická aplikácia metód na vybrané dátové súbory.

Vedúci diplomovej práce: prof. Ing. Aleš Janota, PhD.

Dátum odovzdania diplomovej práce: 6. 6. 2018

Žilina, 26. 10. 2017

prof. Ing. Juraj Spalek, PhD.
vedúci katedry

Poďakovanie

Týmto by som sa chcel poďakovať svojmu vedúcemu diplomovej práce prof. Ing. Alešovi Janotovi, PhD. za odbornú spoluprácu, poskytnutie množstva užitočných rád, pripomienok, materiálov a v neposlednom rade za ochotu. Ďalej sa chcem poďakovať svojim najbližším za podporu a vytvorenie vhodných študijných podmienok počas celého štúdia.

Abstrakt

Hlavným cieľom tejto diplomovej práce bolo oboznámiť sa so zvoleným softvérom RapidMiner, vypracovať vzorové metodické postupy pre jednotlivé metódy dolovania dát a uľahčiť výber operátorov ďalším používateľom. Obsahom práce je vytvorenie zvolených prediktívnych modelov, ich vstupné požiadavky, postup modelovania procesu, vlastnosti metód, čiastkové operácie, ktoré sú potrebné k dosiahnutiu správnej činnosti operátov a praktická aplikácia metód na vybrané dátové súbory. Z výsledkov modelov vyplýva, že metódy sú rozdielne v úspešnosti predikcie pri použití rôznych operátorov a majú veľa možností využitia na rôznych vstupných dátach. Táto práca predstavuje vybrané metódy a opisuje postup ako vytvoriť rôzne procesy s dôrazom na vstupné požiadavky a čiastkové operácie, ktoré sú podmienkou správneho fungovania modelu. Pre uľahčenie výberu operátora používateľovi, bola vytvorená stránka s databázou operátorov, ktorá odporučí výsledný operátor používateľovi na základe jeho vstupných dátach.

Kľúčové slová: RapidMiner, dolovanie dát, predikcia, model.

Abstract

The main goal of this thesis was to familiarize with selected RapidMiner software, to develop sample methodological procedures for individual data mining methods and make it easier for other users to select operators. The subject of the thesis is the creation of selected models, their input requirements, process modelling procedures, the properties of the methods, the partial operations necessary to achieve the correct operation of the operators and the practical application of the methods on the selected data files. The results of the models show that the methods are different in the success of the predictions when using different operators and have many possibilities to use on different input data. This thesis represents selected methods and describes the process for creating various processes with emphasis on input requirements and partial operations that are a condition for proper model operation. To facilitate operator selection for the user, an operator database and website has been created that recommends the resulting operator to the user based on his input data.

Keywords: RapidMiner, data mining, prediction, model.

Obsah

Zoznam obrázkov

Zoznam tabuliek

Zoznam skratiek

Skratka	Anglický význam	Slovenský význam
EMC	Expectation maximization clustering	Zhlukovanie maximalizácie očakávania
GLM	Generalized linear model	Generalizovaný lineárny model
LDA	Linear discriminant analysis	Lineárna diskriminačná analýza
PSO	Particle swarm optimalization	Optimalizácia časticovým rojom
RDA	Regularized discriminant analysis	Regulačná diskriminačná analýza
RVM	Relevance vector machine	Algoritmus relevantného vektora
SVC	Support vector clustering	Zhlukovanie podporných vektorov
SVM	Support vector machine	Algoritmus podporných vektorov
QDA	Quadrattic discriminant analysis	Kvadratická diskriminačná analýza

1 ÚVOD

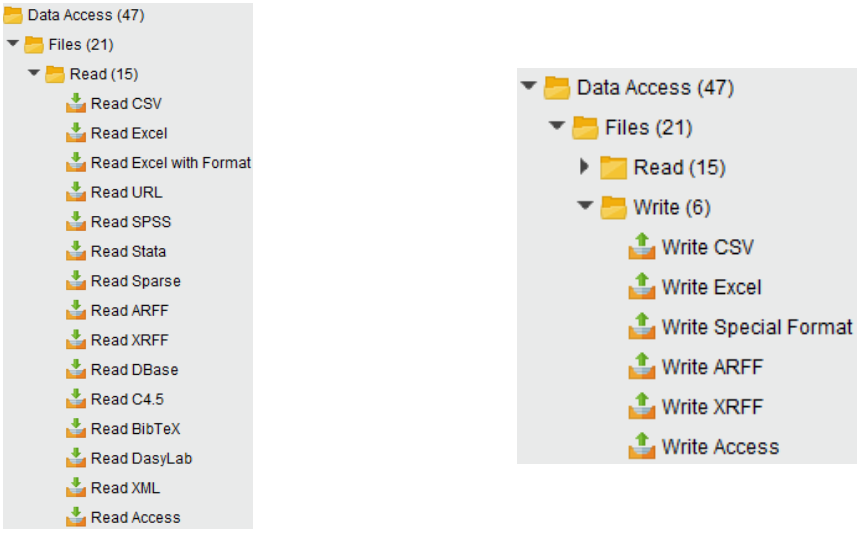
Dolovanie dát, resp. známejší anglický pojem Data mining je výpočtový proces objavovania vzorov vo veľkých množinách údajov. Je to dôležitý proces, pri ktorom sa používajú inteligentné metódy na extrakciu dátových vzorov. Celkovým cieľom dolovania dát je získať informácie z množiny údajov a transformovať ich do zrozumiteľnej štruktúry na ďalšie použitie. Je možné sa s ním stretnúť v bankovníctve, telekomunikáciách, bio-farmaceutickom priemysle, bezpečnostných technológiách a v iných odvetviach, v ktorých je potrebné efektívne spracovať a dolovať dáta na neskoršie použitie pri procese rozhodovania.

V tejto práci sa venujeme softvéru RapidMiner Studio, s ktorým predstavíme techniky dolovania dát, ktoré tento softvér poskytuje. Pred samotným dolovaním dát, je potrebné najskôr konkrétne dáta zozbierať, predspracovať, transformovať a prípadne odfiltrovať na ďalšie použitie v procese. Softvér RapidMiner Studio poskytuje veľké množstvo operátorov na dolovanie dát. Preto sme vytvorili stránku, ktorá používateľovi pomáha s výberom týchto operátorov.

2 RAPIDMINER STUDIO

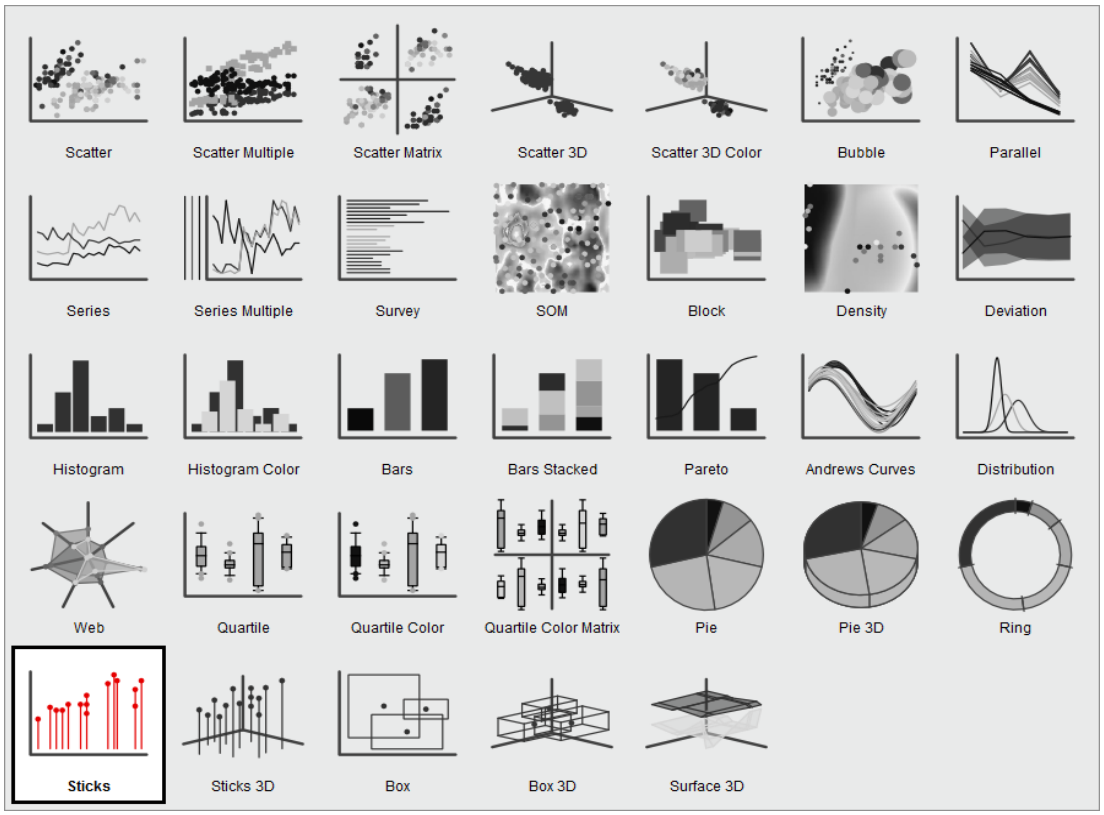
RapidMiner Studio je výkonný vizuálny dizajnér pre rýchle vytváranie prediktívnych analytických postupov, vizualizáciu, dolovanie a ďalšie úlohy. Tento nástroj obsahuje stovky algoritmov na prípravu dát a strojové učenie. Umožňuje pracovať s rôznymi druhmi dát a aj ich vytvárať. Všetky možné formáty vstupu a formáty ukladania výstupu sú na obr. 2.1.

RapidMiner Studio patrí do top 33 softvérov na data mining, kde sa umiestnil ako prvý. Za ním nasledujú: KNIME, Dataiku DSS, SAS Enterprise Miner, IBM SPSS Modeler a iné [13].



Obr. 2.1 Formáty vstupu (vľavo) a formáty uloženia výstupu (vpravo)

Výsledky procesov je možné zobraziť viacerými spôsobmi. RapidMiner Studio umožňuje zobrazenie výsledkov vo forme tabuľky, štatistík a grafov. Na obr. 2.2 sú uvedené rôzne štýly grafov, ktoré program poskytuje.



Obr. 2.2 Štýly grafov

Na predvedenie metód je použitá databáza Titanicu uložená v programe RapidMiner Studio. Množina dát obsahuje nasledujúce atribúty:

- passenger class – trieda vstupeniek (prvá, druhá, tretia),
- name – meno pasažiera,
- sex – pohlavie pasažiera,
- age – vek pasažiera,
- no of siblings or spouses on board – počet súrodencov + manžel/manželka,
- no of parents or children on board – počet rodičov a detí pasažiera,
- ticket number – označenie lístka,
- passenger fare – cestovné,
- cabin – označenie kajuty,
- port of embarkation – prístav nalodenia pasažiera (C = Cherbourg, Q = Queenstown, S = Southampton),
- life boat – záchranný čln,
- survived – či pasažier prežil alebo nie (predikovaný atribút).

Databáza obsahuje celkom 1309 záznamov.

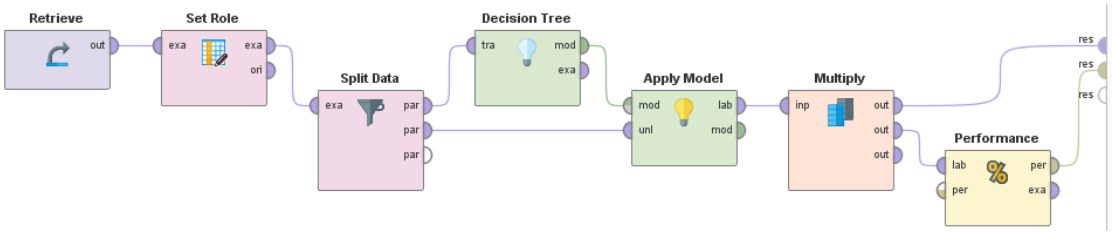
3 ROZHODOVACÍ STROM

Rozhodovací strom je kolekcia uzlov, ktorá má za cieľ vytvoriť rozhodnutie o pridružených hodnotách k triede. Každý uzol predstavuje pravidlo rozdelenia pre jeden konkrétny atribút. Toto pravidlo oddeľuje hodnoty patriace do rôznych tried optimálnym spôsobom pre kritérium vybraného parametra. Budovanie nových uzlov sa opakuje až do splnenia kritérií zastavenia. Predpoveď pre vybraný atribút sa určuje v závislosti od väčšiny príkladov, ktoré dosiahli počas generácie tento uzol.

Rozhodovací strom môže spracovať súbory obsahujúce číselné aj nominálne atribúty. Vybraný atribút, ktorý chceme predpovedať môže byť nominálny aj číselný.

Model rozhodovacieho stromu môže byť aplikovaný na nové príklady pomocou operátora *Apply Model*. Každý príklad nasleduje vetvy stromu v súlade s pravidlom rozdelenia, kým nedosiahne posledný uzol.

Na obr. 3.1 je znázornený proces predikcie pomocou rozhodovacieho stromu a výpočet presnosti tejto predikcie na dátach o pasažieroch Titanicu. Na obr. 3.2 je zobrazený popis rozhodovacieho stromu pre tento prípad.



Obr. 3.1 Proces predikcie

```
Life Boat = 1: Yes {Yes=3, No=0}
Life Boat = 10: Yes {Yes=16, No=0}
Life Boat = 11: Yes {Yes=12, No=0}
Life Boat = 12
|   Sex = Female: Yes {Yes=10, No=0}
|   Sex = Male: No {Yes=0, No=1}
Life Boat = 13: Yes {Yes=21, No=0}
Life Boat = 13 15: Yes {Yes=1, No=0}
Life Boat = 13 15 B: Yes {Yes=1, No=0}
Life Boat = 14: Yes {Yes=13, No=1}
Life Boat = 15: Yes {Yes=19, No=0}
Life Boat = 15 16: Yes {Yes=1, No=0}
Life Boat = 16: Yes {Yes=9, No=0}
Life Boat = 2: Yes {Yes=8, No=0}
Life Boat = 3: Yes {Yes=14, No=0}
Life Boat = 4: Yes {Yes=14, No=0}
Life Boat = 5: Yes {Yes=12, No=0}
Life Boat = 5 7: Yes {Yes=1, No=0}
Life Boat = 5 9: Yes {Yes=1, No=0}
Life Boat = 6: Yes {Yes=11, No=0}
Life Boat = 7: Yes {Yes=11, No=0}
Life Boat = 8: Yes {Yes=14, No=0}
Life Boat = 8 10: Yes {Yes=1, No=0}
Life Boat = 9: Yes {Yes=14, No=0}
Life Boat = ?: No {Yes=9, No=397}
Life Boat = A
|   No of Parents or Children on Board > 0.500: Yes {Yes=2, No=0}
|   No of Parents or Children on Board ≤ 0.500
|   |   Age = ?
|   |   |   Passenger Class = First: Yes {Yes=1, No=0}
|   |   |   Passenger Class = Third: No {Yes=0, No=1}
|   |   Age > 27.500: No {Yes=0, No=3}
|   |   Age ≤ 27.500: Yes {Yes=1, No=0}
Life Boat = B
|   Passenger Fare > 8.875: Yes {Yes=3, No=0}
|   Passenger Fare ≤ 8.875: No {Yes=0, No=1}
Life Boat = C: Yes {Yes=18, No=0}
Life Boat = C D: Yes {Yes=1, No=0}
Life Boat = D: Yes {Yes=8, No=1}
```

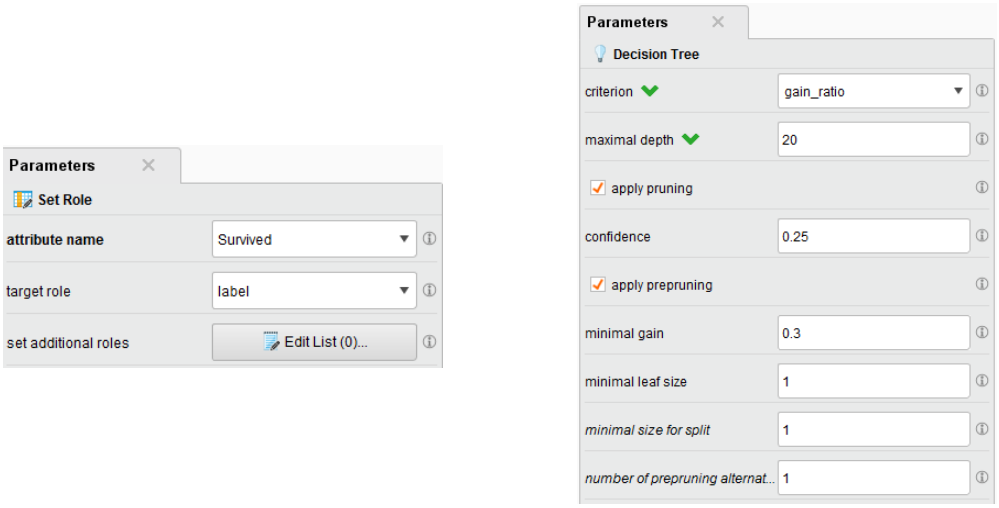
Obr. 3.2 Rozhodovací strom

Pre predpoveď vybraného atribútu pomocou rozhodovacieho stromu, je potrebné nastaviť rolu atribútu na *label*. Toto nastavenie vykonáva operátor *Set Role*. Operátor *Split Data* rozdelí vstupné dáta, v tomto prípade pol na pol. Následne operátor *Apply Model* aplikuje model rozhodovacieho stromu na druhú časť dát a predikuje zvolený atribút. Na výstupe procesu sa zobrazia dáta s pridanými stĺpcami predikovaného atribútu a vypočítaný výkon predikcie. V tabuľke 3.1 sú stručné popisy operátorov použitých v tomto procese.

Tabuľka 3.1 Popisy operátorov

Názov operátora	**Popis**	**Umiestnenie bloku v menu**
Retrieve	Tento operátor načítava uložené informácie do procesu.	Data Access
Set Role	Operátor, ktorý mení role jedného alebo viacerých atribútov.	Blending - Attributes – Names & Roles
Split Data	Operátor, ktorý rozdelí dáta podľa špecifikovaných pomerov.	Blending – Examples - Sampling
Decission Tree	Generuje rozhodovací strom.	Modeling – Predictive - Trees
Apply Model	Tento operátor aplikuje pridelený model na vložené dáta.	Scoring
Multiply	Operátor, ktorý vytvára kópie vstupných objektov.	Utility
Performance	Operátor, ktorý hodnotí výkonnosť.	Validation - Performance

Na obr. 3.3 vľavo je zobrazené bližšie nastavenie parametrov operátora *Set Role*. Atribút, ktorý je predpovedaný je *Survived*, preto je potrebné ho označiť ako *label*. Vpravo na tom istom obrázku sú parametre pre rozhodovací strom.



Obr. 3.3 Nastavenie parametrov operátora Set Role (vľavo) a parametre rozhodovacieho stromu (vpravo)

Nastavením správnych parametrov rozhodovacieho stromu sa zlepšuje presnosť predikcie. V tabuľke 3.2 sú uvedené popisy parametrov rozhodovacieho stromu. Kritéria rozdelenia, ktoré možno vybrať a ich popis sú v tabuľke 3.3.

Tabuľka 3.2 Popis parametrov rozhodovacieho stromu

Názov parametru	**Popis**
criterion	Výber kritéria rozdelenia.
maximal depth	Obmedzenie maximálnej hĺbky stromu.
confidence	Určenie úrovne spoľahlivosti.
minimal gain	Minimálny zisk z uzla pred rozdelením.
minimal leaf size	Minimálny počet príkladov v uzle.
minimal size for split	Minimálna veľkosť parametrov pre rozdelenie.
number of prepruning alternatives	Počet dopredne prerezávaných alternatív.

Tabuľka 3.3 Popis kritérií

Kritérium	**Popis**
gain_ratio	Vyberá sa atribút, ktorého hodnota „pomerného informačného zisku“ je maximálna (informačný zisk podelený entropiou delenia).
information_gain	Vyberá sa atribút, ktorého hodnota informačného zisku je maximálna (t. j. atribút s min. entropiou).
gini_index	Meranie nerovnosti medzi distribúciou vlastností *label* atribútu.
accuracy	Atribút je vybraný na rozdelenie, čo maximalizuje presnosť celého stromu.
least_square	Atribút minimalizuje štvorcovú vzdialenosť medzi priemerom hodnôt vzhľadom na skutočnú hodnotu. Používa sa iba v prípade, že je *label* číselný.

Príklad výslednej predikcie je znázornený na obr. 3.4. Zelenou farbou je označený skutočný atribút *Survived* a výsledná predikcia. Žltou farbou sú označené predikcie s príslušnou dôverou.

Row No.	Survived	prediction(Survived)	confidence(Yes)	confidence(No)	Passenger ...	Name	Sex	Age
200	Yes	No	0.022	0.978	Second	Doling, Mrs. J...	Female	34
201	No	No	0.022	0.978	Second	Eitemiller, Mr....	Male	23
202	No	No	0.022	0.978	Second	Enander, Mr. I...	Male	21
203	No	No	0.022	0.978	Second	Fahlstrom, Mr...	Male	18
204	No	No	0.022	0.978	Second	Faunthorpe, ...	Male	40
205	Yes	Yes	1	0	Second	Faunthorpe, ...	Female	29
206	No	No	0.022	0.978	Second	Fillbrook, Mr. ...	Male	18
207	No	No	0.022	0.978	Second	Fox, Mr. Stanl...	Male	36
208	No	No	0.022	0.978	Second	Funk, Miss. A...	Female	38
209	No	No	0.022	0.978	Second	Gale, Mr. Sha...	Male	34
210	Yes	Yes	1	0	Second	Garside, Mis...	Female	34
211	No	No	0.022	0.978	Second	Gavey, Mr. La...	Male	26
212	No	No	0.022	0.978	Second	Giles, Mr. Fre...	Male	21
213	No	No	0.022	0.978	Second	Hale, Mr. Reg...	Male	30
214	Yes	Yes	1	0	Second	Hamalainen, ...	Male	0.667
215	Yes	Yes	1	0	Second	Hamalainen, ...	Female	24
216	No	No	0.022	0.978	Second	Harper, Rev. ...	Male	28
217	No	No	0.022	0.978	Second	Harris, Mr. W...	Male	30
218	Yes	Yes	0.929	0.071	Second	Hart, Miss. Ev...	Female	7
219	No	No	0.022	0.978	Second	Hart, Mr. Benj...	Male	43
220	Yes	Yes	1	0	Second	Herman, Mis...	Female	24
221	No	No	0.022	0.978	Second	Herman, Mr. ...	Male	49
222	No	No	0.022	0.978	Second	Hickman, Mr. ...	Male	32
223	No	No	0.022	0.978	Second	Hocking, Mr. ...	Male	23
224	No	No	0.022	0.978	Second	Hold, Mr. Ste...	Male	44
225	No	No	0.022	0.978	Second	Jacobsohn, ...	Male	42
226	Yes	Yes	1	0	Second	Jacobsohn, ...	Female	24

Obr. 3.4 Výsledná predikcia

Výsledná presnosť predikcie je znázornená na obr. 3.5. Pri predikovaní, že pasažier prežije, sa model pomýlil len raz a pri opačnom prípade 20 krát. Celková presnosť je 96.79%.

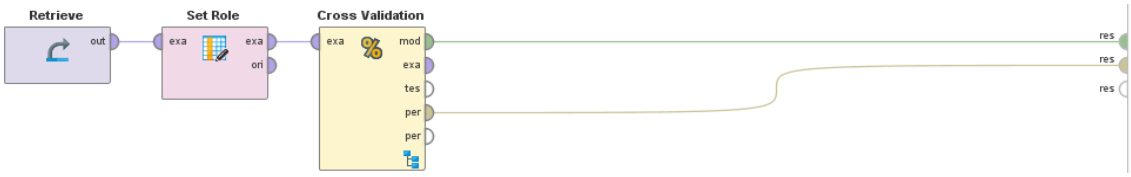
accuracy: 96.79%

	true Yes	true No	class precision
pred. Yes	230	1	99.57%
pred. No	20	403	95.27%
class recall	92.00%	99.75%	

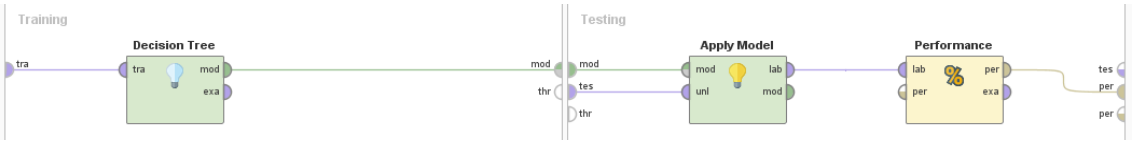
Obr. 3.5 Presnosť predikcie

V programe RapidMiner Studio je viacero operátorov, ktoré sa dajú použiť na overenie presnosti predikcie. Jedným z nich je aj operátor *Cross Validation* – krížová validácia. Má dva podprocesy - tréningový a testovací podproces. Tréningový podproces sa používa na tréning modelu. Model sa potom použije v testovacom podprocese. Výkonnosť modelu sa meria počas testovacej fázy. Množina dát je rozdelená do *k*-podmnožín rovnakej veľkosti. Z *k*-podmnožín zostáva zachovaná jedna podmnožina ako súbor testovacích dát. Zostávajúce podmnožiny *k*-1 sa používajú ako súbor tréningových údajov. Proces krížovej validácie sa potom opakuje *k*-krát, pričom každá z kmeňových

podmnožín sa použije presne raz ako testovacie údaje. Hodnoty z *k* interakcií sú spriemerované na vytvorenie jedného odhadu. Proces s použitím operátora *Cross Validation* je znázornený na obr. 3.6 a podprocesy operátora na obr. 3.7.



Obr. 3.6 Proces s použitím operátora krížovej validácie



Obr. 3.7 Podprocesy operátora krížovej validácie

Výsledná presnosť tohto modelu je znázornená na obr. 3.8. Presnosť modelu je 97,02% +/- 1,5%.

accuracy: 97.02% +/- 1.50% (mikro: 97.02%)

	true Yes	true No	class precision
pred. Yes	469	8	98.32%
pred. No	31	801	96.27%
class recall	93.80%	99.01%	

Obr. 3.8 Presnosť modelu s krížovou validáciou

RapidMiner poskytuje viacero operátorov stromov. Rozdiely medzi ostatnými stromami sú v tabuľke 3.4.

Tabuľka 3.4 Rozdiel ostatných stromov oproti rozhodovaciemu stromu

Operátor	**Rozdiel oproti rozhodovaciemu stromu**
Random Forest	Generuje súbor určeného počtu náhodných stromov – náhodný les. Výsledný model je hlasovacím modelom všetkých stromov.
Gradient Boosted Trees	Súbor stromových modelov so zosilneným gradientom. Poskytuje výstupný port s váhami atribútov vzhľadom na *label* atribút.
CHAID	Generuje strom rovnako ako operátor rozhodovacieho stromu s výnimkou, že používa kritérium chi-squared. Môže byť použitý len na príkladoch s nominálnymi údajmi.
ID3	Iterative Dichotomiser 3 vytvára rozhodovací strom z pevného súboru príkladov. Ak sa testuje malá vzorka, dáta môžu byť nadhodnotené, alebo príliš špecifické a strom nevyhodnotí iné príklady správne. Tento operátor nedokáže spracovať numerické atribúty a chýbajúce dáta.
Decision Stump	Vytvorí strom iba s jedným rozdelením.
Decision Tree (Multiway)	Vytvára viacradový rozhodovací strom. Obsahuje podproces, v ktorom musí byť operátor generujúci stromový model.
Decision Tree (Weight-Based)	Generuje rozhodovací strom, založený na ľubovoľnom teste relevantnosti atribútu. Obsahuje podproces, v ktorom musí byť operátor generujúci váhy atribútov. Môže byť použitý len na príkladoch s nominálnymi údajmi. Nespracuje chýbajúce údaje.
Random Tree	Pracuje ako operátor rozhodovacieho stromu s výnimkou, že vyberie náhodnú podmnožinu atribútov ešte predtým, ako sa aplikuje. Veľkosť podmnožiny je určená parametrom pomeru podmnožiny.

4 BAYESOV NAIVNÝ KLASIFIKÁTOR

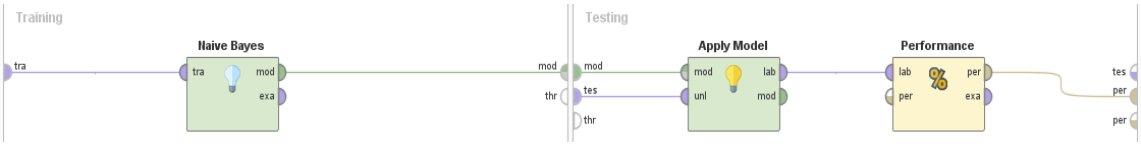
Bayesov naivný klasifikátor je jednoduchý pravdepodobnostný klasifikátor, založený na uplatňovaní Bayesovho teorému so silnými (naivnými) predpokladmi nezávislosti. Jednoducho klasifikátor predpokladá, že prítomnosť (alebo neprítomnosť) atribútu nesúvisí s prítomnosťou (alebo neprítomnosťou) inej funkcie.

Výhodou klasifikátora je, že vyžaduje len malé množstvo tréningových údajov na odhad prostriedkov a odchýlok premenných, ktoré sú potrebné na klasifikáciu. Pretože sa predpokladajú nezávislé premenné, treba určiť len odchýlky premenných pre odhadovaný atribút - *label* a nie pre celú kovariačnú maticu.

Operátor *Naive Bayes* má len jeden parameter a to *laplace correction*. Tento parameter indikuje, či sa má použiť Laplaceova korekcia, aby sa zabránilo vysokému vplyvu nulových pravdepodobností.

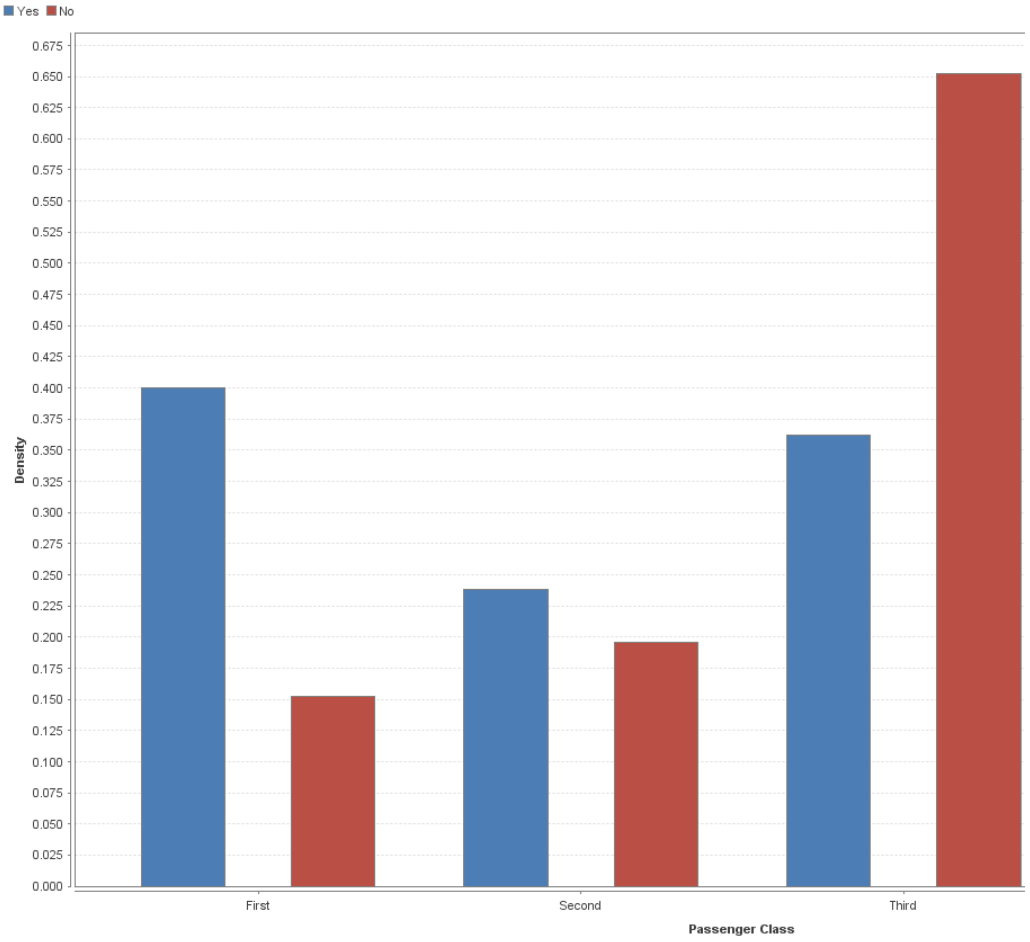
Tento operátor môže spracovať súbory obsahujúce číselné aj nominálne atribúty. Vybraný atribút, ktorý chceme predpovedať musí byť nominálny.

Na ukážku je použitá rovnaká databáza pasažierov Titanicu ako pri rozhodovacích stromoch. Proces je totožný s obr. 3.6, len je použitý operátor *Naive Bayes* v tréningovej časti validácie. Podprocesy krížovej validácie sú znázornené na obr. 4.1.



Obr. 4.1 Podprocesy krížovej validácie s použitím operátora Naive Bayes

Po spustení procesu vznikne model rozdelenia pre atribút *survived* a výsledná presnosť predikcie. Výsledné rozdelenie pre triedu, že pasažier prežije je 0,382 a pre opačný prípad 0,618. Rozdelenie pre každý atribút je možné zobraziť aj graficky. Na obr. 4.2 je znázornené rozdelenie podľa atribútu triedy vstupeniek pasažierov. Modrá farba znázorňuje pasažierov, ktorí prežili a červená, ktorí neprežili. Možné hodnoty atribútu triedy vstupeniek sú: prvá, druhá a tretia trieda. Výsledná presnosť je na obr. 4.3.



Obr. 4.2 Rozdelenie podľa triedy vstupeniek pasažierov

accuracy: 90.07% +/- 2.54% (mikro: 90.07%)

	true Yes	true No	class precision
pred. Yes	459	89	83.76%
pred. No	41	720	94.61%
class recall	91.80%	89.00%	

Obr. 4.3 Presnosť predikcie pri použití operátora Naive Bayes

Rozdiel oproti operátoru *Naive Bayes (Kernel)* je taký, že *Naive Bayes (Kernel)* generuje model pomocou odhadovaných hustôt jadra a nedokáže spracovať chýbajúce údaje. Odhad hustôt jadra sa dá modifikovať pomocou parametrov operátora, ako napríklad režimom odhadu, alebo šírkou pásma.

5 NEURÓNOVÉ SIETE

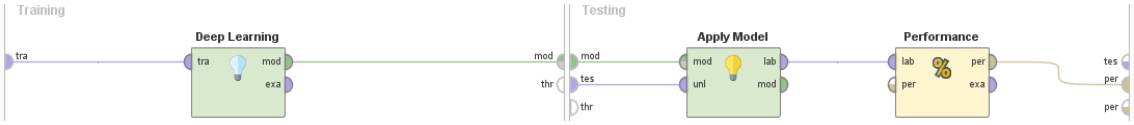
Program RapidMiner Studio poskytuje štyri operátory neurónovej siete pre prediktívne modelovanie. Ako prvý operátor v tejto skupine je *Deep Learning*.

Hlboké učenie je založené na viacvrstvovej doprednej umelej neurónovej sieti, ktorá je trénovaná so stochastickým gradientom klesania s využitím spätného šírenia. Sieť môže obsahovať veľké množstvo skrytých vrstiev pozostávajúcich z neurónov s rôznymi aktivačnými funkciami. Parametre operátora *Deep Learning* sú znázornené v tabuľke 5.1.

Tabuľka 5.1 Parametre operátora Deep Learning

Parameter	**Popis**
activation	Výber aktivačnej funkcie neurónov v skrytých vrstvách. Napríklad hyperbolický tangens.
hidden layer sizes	Počet a veľkosť skrytých vrstiev.
reproducible (uses 1 thread)	Zapne vynútené reprodukovanie na malých dátach s použitím iba jedného vlákna.
epochs	Počet epoch (t. j. koľkokrát sa množina údajov má opakovať).
compute variable importances	Určuje, či sa majú vypočítať významy vstupných funkcií. Implementovaná metóda zvažuje váhy spájajúce vstupné prvky s prvými dvoma skrytými vrstvami.
train samples per iteration	Počet riadkov, ktoré sa majú spracovať na jedno opakovanie. Špeciálne hodnoty sú 0 pre jednu epochu na iteráciu, -1 pre spracovanie max. množstva údajov a -2 zapne automatický režim.
adaptive rate	Možnosť použitia algoritmu adaptívneho učenia.
epsilon	Parameter adaptívneho učenia podobný rýchlosti učenia.
rho	Parameter adaptívneho učenia podobný hybnosti.
standardize	Povolenie štandardizácie dát.
L1	Normalizačná metóda, ktorá obmedzuje absolútnu hodnotu váh, znižuje zložitosť a zabraňuje nadmernému namáhaniu.
L2	Metóda regularizácie, ktorá obmedzuje súčet štvorcových váh.
max w2	Maximálna hodnota na súčte štvorcových prichádzajúcich váh do jedného neurónu. Špeciálna hodnota 0 znamená nekonečno.
loss function	Výber funkcie straty, ktorá má model minimalizovať.
distribution function	Výber distribučnej funkcie pre tréningové dáta.
early stopping	Ak je povolené skoré zastavenie, musí byť špecifikované.
missing values handling	Spracovanie chýbajúcich hodnôt. Má dve možnosti: preskočiť alebo pridať strednú hodnotu.
max runtime seconds	Max. povolená doba behu tréningu modelu v sekundách.
expert parameters	Možnosť pridania expertných parametrov.

S použitím rovnakej databázy pasažierov Titanicu ako pri predchádzajúcich prípadoch je proces len upravený s operátorom *Deep Learning* v tréningovej časti validácie. Podprocesy krížovej validácie sú znázornené na obr. 5.1. Výsledná presnosť 96.34% je zobrazená na obr. 5.2.



Obr. 5.1 Podprocesy krížovej validácie s použitím operátora Deep Learning

accuracy: 96.34% +/- 1.95% (mikro: 96.33%)

	true Yes	true No	class precision
pred. Yes	479	27	94.66%
pred. No	21	782	97.38%
class recall	95.80%	96.66%	

Obr. 5.2 Presnosť predikcie pri použití operátora Deep Learning

Druhý operátor v skupine neurónových sieti je *Neural Net*. Tento operátor dokáže spracovať iba číselné atribúty (okrem atribútu *label,* ktorý môže byť číselný aj nominálny) a taktiež nesmú žiadne hodnoty chýbať. V tomto operátore sa ako aktivačná funkcia používa obvyklá sigmoidná funkcia. Hlavnými parametrami tohto operátora sú: tréningové cykly používané pri učení neurónovej siete, meno a veľkosť skrytých vrstiev, rýchlosť učenia a momentum váh.

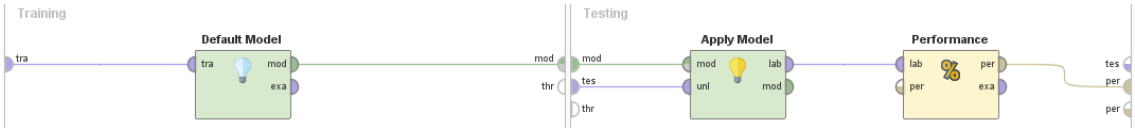
Ďalší operátor je *AutoMLP*. Je to jednoduchý, efektívny algoritmus pre rýchle učenie aj nastavenie veľkosti neurónových sietí počas tréningu. Hlavné parametre sú: počet maximálnych tréningových cyklov používaných na tréning neurónovej siete, počet generácií a MLP. Takisto ako operátor *Neural Net* dokáže spracovať iba číselné atribúty (okrem atribútu *label*, ktorý nesmie byť číselný) bez chýbajúcich hodnôt.

Posledný operátor je *Perceptron* – najjednoduchší, jednovrstvový druh neurónovej siete. Spracúva ako aj *AutoMLP* iba číselné atribúty (okrem *labelu*) bez chýbajúcich hodnôt. *Perceptron* má len dva parametre a to počet kôl a rýchlosť učenia.

6 LENIVÉ OPERÁTORY

Do kategórie lenivých operátorov patria *Default Model* a *k-NN*. Označujú sa ako lenivé, pretože sú najjednoduchšie. Prvý operátor *Default Model* generuje model, ktorý predpovedá predvolenú hodnotu atribútu *label* vo všetkých príkladoch. Tento operátor má iba jeden hlavný parameter a to výber metódy. Pri číselnom atribúte *label* môže byť predvolená hodnota stredná, priemerná alebo zadaná konštantná hodnota. Pri nominálnych hodnotách sa môže použiť režim atribútu *label* alebo výberu atribútu.

Pri zapojení operátora *Default Model* do tréningovej časti operátora krížovej validácie ako je na obr. 6.1 je použitá metóda režimu *label*. Použitím tejto metódy model použije na predikciu najčastejšie sa vyskytujúcu hodnotu atribútu *label*. V tomto prípade to je hodnota *NO* – čiže pasažier neprežije. Výsledná presnosť 61.8% je znázornená na obr. 6.2.



Obr. 6.1 Podprocesy krížovej validácie s použitím operátora Default Model

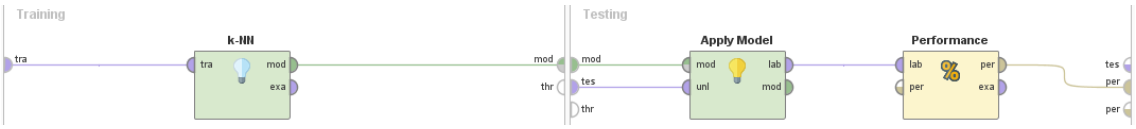
accuracy: 61.80% +/- 0.28% (mikro: 61.80%)

	true Yes	true No	class precision
pred. Yes	0	0	0.00%
pred. No	500	809	61.80%
class recall	0.00%	100.00%	

Obr. 6.2 Presnosť predikcie pri použití operátora Default Model

Druhý operátor *k-NN* je založený na učení s porovnávaním testovacích a tréningových dát, ktoré sú podobné. Tréningové dáta sú opísané *n* atribútmi. Každý záznam predstavuje bod v *n*-dimenzionálnom priestore. Týmto spôsobom sú uložené všetky tréningové záznamy. Ak je daný neznámy záznam, potom algoritmus *k*-najbližšieho suseda vyhľadá vzorový priestor pre *k*-záznamy tréningu, ktoré sú najbližšie k neznámemu záznamu. Meranie najbližších susedov je definované pomocou vybranej metódy v parametroch operátora, ako je napríklad Euklidovská vzdialenosť.

Podproces krížovej validácie s použitím *k-NN* operátora je znázornený na obr. 6.3. Pri zvolenej hodnote $k = 1$ model dosiahol presnosť 62.42%. Predikcie sú uvedené na obr. 6.4.



Obr. 6.3 Podprocesy krížovej validácie s použitím operátora k-NN

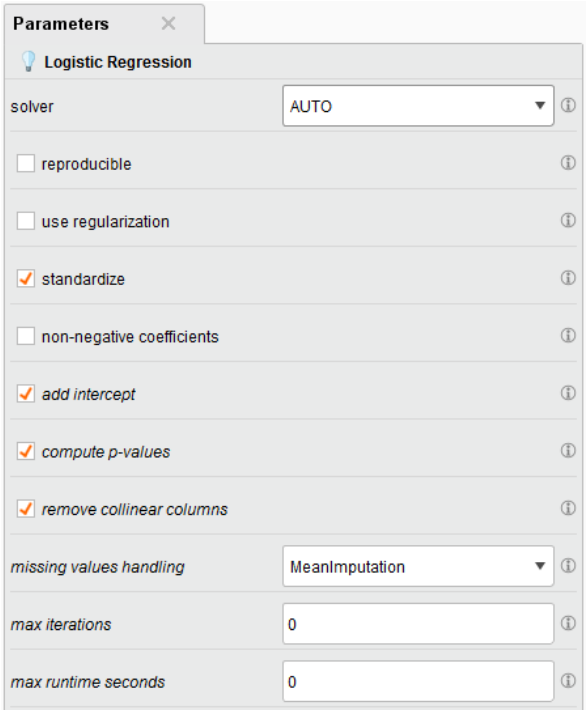
accuracy: 62.42% +/- 4.22% (mikro: 62.41%)

	true Yes	true No	class precision
pred. Yes	213	205	50.96%
pred. No	287	604	67.79%
class recall	42.60%	74.66%	

Obr. 6.4 Presnosť predikcie pri použití operátora k-NN

7 LOGISTICKÁ REGRESIA

Logistická regresia sa používa na opis dát a na vysvetlenie vzťahu medzi jednou závislou binárnou premennou a jednou alebo viacerými premennými nominálnej alebo číselnej hodnoty. RapidMiner Studio poskytuje tri operátory logistickej regresie. Operátor *Logistic Regression* poskytuje jednoduchú logistickú regresiu pre zvolené najdôležitejšie parametre. Na obr. 7.1 sú znázornené tieto parametre a v tabuľke 7.1 ich popisy. Tento operátor dokáže spracovať nominálne aj číselné hodnoty, ale atribút *label* musí byť nominálny.



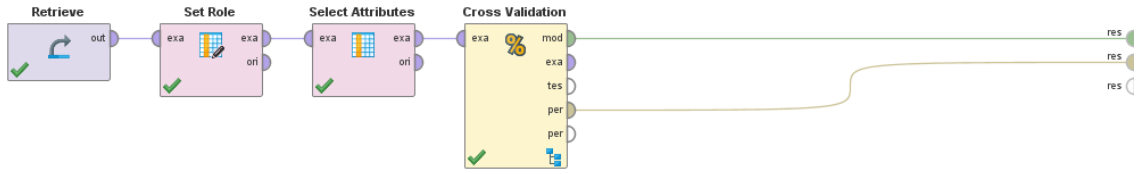
Obr. 7.1 Parametre operátora Logistic Regression

Tabuľka 7.1 Opis parametrov operátora Logistic Regression

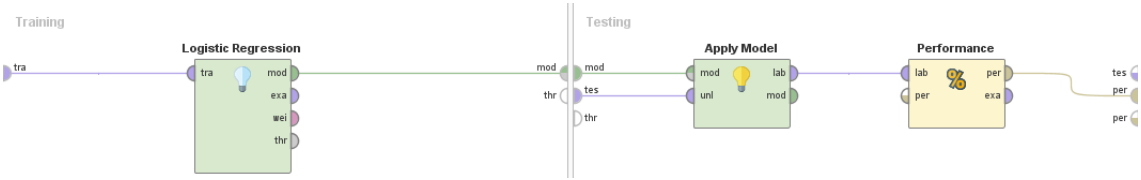
Parameter	**Opis**
solver	Výber metódy riešenia.
reproducible	Umožňuje paralelné riešenie modelu.
use regularization	Použitie regularizácie hodnôt.
standardize	Štandardizuje číselné stĺpce proti nulovým odchýlkam.
non-negative coefficients	Obmedzenie negatívnych koeficientov.
add intercept	Zahrnutie konštánt v modeli.

Parameter	Opis
compute p-values	Výpočet hodnôt p.
remove collinear columns	Odstránenie závislých stĺpcov.
missing values handling	Spôsob spracovania chýbajúcich hodnôt. Môžu sa preskočiť alebo uvažovať priemerné hodnoty.
max iterations	Max. počet opakovaní. Nula znamená bez obmedzenia.
max runtime seconds	Max. povolená doba behu v sekundách.

V procese s operátorom *Logistic Regression* je zapojený aj operátor *Select Atrributes* (umiestnenie: *Blending-Attributes-Selection*), ktorý vyberá atribúty pre ďalšie spracovanie. Kvôli najefektívnejšiemu modelovaniu sú vybraté z databázy Titanicu atribúty: *Age*, *Life Boat*, *Passenger Fare* a *Survived*. Proces je znázornený na obr. 7.2, podprocesy krížovej validácie sú na obr. 7.3 a výsledná presnosť 94.96% na obr. 7.4.



Obr. 7.2 Proces s výberom atribútov



Obr. 7.3 Podprocesy krížovej validácie s použitím operátora Logistic Regression

accuracy: 94.96% +/- 1.24% (mikro: 94.96%)

	true Yes	true No	class precision
pred. Yes	443	9	98.01%
pred. No	57	800	93.35%
class recall	88.60%	98.89%	

Obr. 7.4 Presnosť predikcie pri použití operátora Logistic Regression

Druhý operátor *Logistic Regression (SVM)* dokáže spracovať iba číselné hodnoty (okrem atribútu *label*, ktorý musí byť nominálny) a nedokáže pracovať s chýbajúcimi hodnotami. Líši sa výberom jadra. Operátor *Logistic Regression (Evolutionary)* má rovnaké nároky na dáta ako *Logistic Regression (SVM),* ale neponúka na výstupe váhy atribútov. Líši sa výberom jadra a počtom generácií.

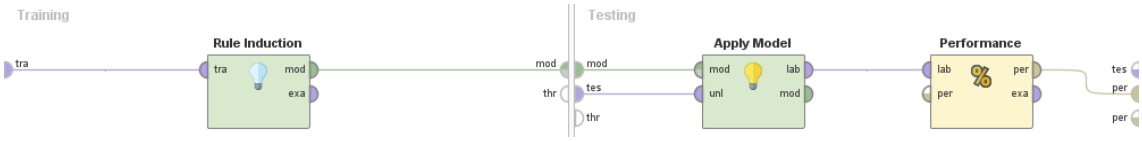
8 PRAVIDLOVÉ OPERÁTORY

Pravidlové operátory využívajú induktívne učenie sa. Učiaci systém hľadá také pravidlá, ktoré najlepšie klasifikujú tréningové príklady. RapidMiner Studio poskytuje päť operátorov v tejto skupine. Prvý operátor *Rule Induction* sa naučí súbor pravidiel s ohľadom na vstupné dáta. Algoritmus začína s najmenej prevládajúcou triedou, pričom postupne rastie a orezáva pravidlá, až kým nezostanú žiadne pozitívne príklady, alebo veľkosť chyby nie je väčšia ako 50%. V rastúcej fáze sa pre každé pravidlo pridávajú podmienky, až kým nie je 100% presné. Procedúra skúša každú možnú hodnotu každého atribútu a vyberie podmienku s najvyšším ziskom informácií. V tabuľke 8.1 sú znázornené parametre tohto operátora. Tento operátor nedokáže spracovať číselný *label*.

Tabuľka 8.1 Opis parametrov operátora Rule Induction

Parameter	**Opis**
criterion	Určuje kritérium pre výber atribútov a numerické rozdelenie. Môže byť zvolený informačný zisk alebo presnosť.
sample ratio	Určenie pomeru tréningových dát pre rast a orezávanie.
pureness	Určenie minimálneho pomeru hlavnej triedy v podmnožine, aby sa podmnožina považovala za čistú.
minimal prune benefit	Určenie minimálnej výšky prospechu, ktorá musí byť prekročená nad neorezaným prospechom, aby sa orezávalo.
use local random seed	Možnosť použiť náhodné číslo.

Podproces krížovej validácie s použitím operátora *Rule Induction* je znázornený na obr. 8.1 a výsledná presnosť 61,8% na obr. 8.2.



Obr. 8.1 Podprocesy krížovej validácie s použitím operátora Rule Induction

accuracy: 61.80% +/- 0.28% (mikro: 61.80%)

	true Yes	true No	class precision
pred. Yes	0	0	0.00%
pred. No	500	809	61.80%
class recall	0.00%	100.00%	

Obr. 8.2 Výsledná presnosť pri použití operátora Rule Induction

Druhý operátor *Single Rule Induction* vytvára jedno najlepšie pravidlo konjunkcie. Nedokáže spracovať číselné atribúty (ani číselný *label*) a chýbajúce dáta. Pri použití na databáze Titanicu zvolil pravidlo: *Sex = Female* → *Survived = Yes*. Čiže predpokladá, že ak je pasažier žena, tak prežije. Výsledná presnosť 78% je znázornená na obr. 8.3.

accuracy: 78.00% +/- 4.19% (mikro: 78.00%)

	true Yes	true No	class precision
pred. Yes	339	127	72.75%
pred. No	161	682	80.90%
class recall	67.80%	84.30%	

Obr. 8.3 Výsledná presnosť pri použití operátora Single Rule Induction

Ďalší operátor v tejto skupine *Single Rule Induction (Single Attribute)* sa zameriava na jeden atribút a určuje najlepšie podmienky rozdelenia pre minimalizáciu tréningovej chyby. Výsledkom je jediné pravidlo podobne, ako pri predošlom operátore. Nedokáže spracovať chýbajúce údaje, *label* musí byť nominálny a nemá žiadne nastaviteľné parametre.

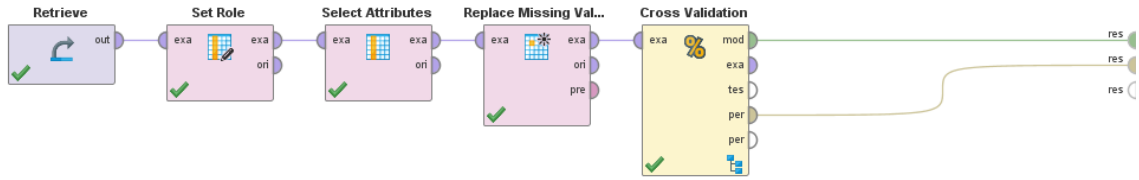
Operátor *Subgroup Discovery* vyhľadáva podskupiny. Cieľom vyhľadávania podskupín je nájsť pravidlá popisujúce podmnožiny populácie, ktoré sú dostatočne veľké a štatisticky nezvyčajné. Tento operátor nedokáže spracovať číselné atribúty (ani číselný *label*) a chýbajúce dáta. V parametroch operátora sa dá vybrať z dvoch režimov a to režim minimálnej užitočnosti alebo režim *k*-najlepších pravidiel.

Posledný operátor v tejto skupine *Tree to Rules* určuje súbor pravidiel z daného modelu rozhodovacieho stromu, ktorý je použitý v jeho podprocese. Podproces preto musí obsahovať operátor, ktorý generuje stromový model. Podľa typu použitého stromového operátora sa prenášajú obmedzenia použitia v procese. Napríklad použitím operátora *CHAID* v podprocese sa prenesie jeho vlastnosť spracovať iba nominálne atribúty.

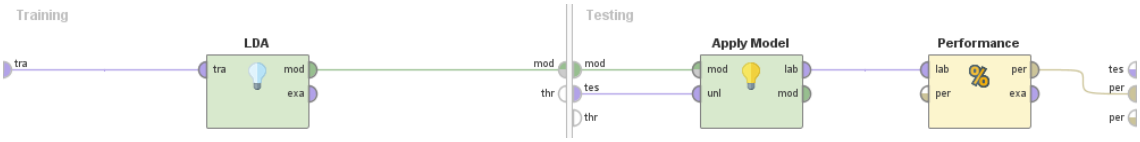
9 DISKRIMINAČNÁ ANALÝZA

Diskriminačná analýza je klasifikačná metóda, ktorá meria dôležitosť faktorov, určujúcich patričnosť do kategórie. RapidMiner Studio poskytuje tri operátory v tejto skupine. Prvý operátor *Linear Discriminant Analysis* sa pokúša nájsť lineárnu kombináciu funkcií, ktorá najlepšie oddelí dve alebo viac tried príkladov. Výsledná kombinácia sa potom použije ako lineárny klasifikátor. Dokáže spracovať iba číselné atribúty okrem *labelu*, ktorý musí byť nominálny. Nedokáže spracovať chýbajúce údaje.

Tento operátor (LDA) neposkytuje žiadne nastaviteľné parametre. Pri aplikovaní na databázu Titanicu je potrebné vybrať číselné atribúty, zvoliť ako *label* - *survived* a nahradiť chýbajúce údaje. Kompletný proces je znázornený na obr. 9.1, podproces validácie na obr. 9.2 a výsledná presnosť 66,54% na obr. 9.3.



Obr. 9.1 Kompletný proces pre príklad s operátorom LDA



Obr. 9.2 Podprocesy krížovej validácie s použitím operátora LDA

accuracy: 66.54% +/- 1.96% (mikro: 66.54%)

	true Yes	true No	class precision
pred. Yes	98	36	73.13%
pred. No	402	773	65.79%
class recall	19.60%	95.55%	

Obr. 9.3 Výsledná presnosť pri použití operátora LDA

Druhý operátor *Quadratic Discriminant Analysis* vykonáva kvadratickú diskriminačnú analýzu (QDA) pre nominálny *label* a číselné atribúty. Nedokáže spracovať chýbajúce údaje. Kvadratická diskriminačná analýza úzko súvisí s lineárnou diskriminačnou analýzou, kde sa predpokladá, že merania z každej triedy sú normálne rozdelené. Avšak v QDA neexistuje predpoklad, že kovariancia každej z tried je identická. Tento operátor taktiež neposkytuje žiadne nastaviteľné parametre.

Posledný operátor v tejto skupine *Regularized Discriminant Analysis* vykonáva regulačnú diskriminačnú analýzu (RDA) pre nominálny *label* a číselné atribúty. RDA vznikla kombináciou LDA a QDA, kde sa optimalizuje kombinácia rozptylových matíc. Rovnako ako ostatné diskriminačné analýzy nedokáže spracovať chýbajúce údaje. Poskytuje jeden nastaviteľný parameter a to parameter *alpha*, ktorý určuje silu regulácie.

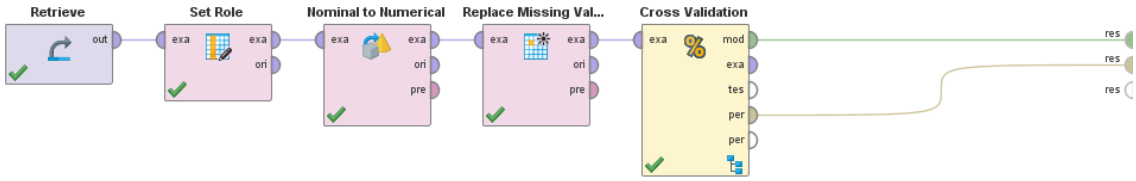
10 PODPORNÉ VEKTORY (SVM)

Táto metóda môže byť použitá pre regresiu aj klasifikáciu. Cieľom SVM je nájsť nadrovinu, ktorá optimálne rozdeľuje priestor príznakov (atribútov) na podpriestory obsahujúce tréningové dáta príslušné odlišným triedam. Optimálna nadrovina je taká, ktorá má hodnotu minima vzdialeností bodov od roviny čo najväčšiu, tzn. maximálny odstup reprezentovaný čo najširším oddeľujúcim pruhom bez bodov. Opísať nadrovinu potom možno iba bodmi ležiacimi na okraji tohto pruhu, ktorých je zvyčajne málo. Tieto body sa nazývajú podporné vektory. Ak je možné objekty s rôznymi príznakmi od seba oddeliť, ide o separovateľnú úlohu. Ak ich nie je možné oddeliť, je to neseparovateľná úloha. Neseparovateľné úlohy je možné previesť na separovateľné pomocou tzv. jadrových transformácií. Ide o transformáciu dát do priestoru inej dimenzie. V novom priestore je potom možné aplikovať optimalizačný algoritmus pre nájdenie rozdeľujúcej nadroviny.

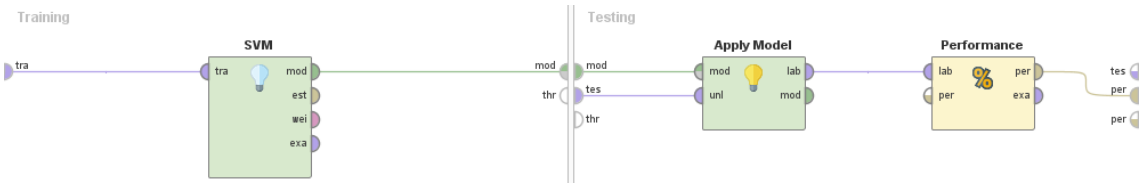
RapidMiner Studio poskytuje sedem operátorov v tejto skupine. Všetky operátory dokážu pracovať iba s číselnými atribútmi a bez chýbajúcich dát, pričom *label* môže byť iba binominálny a u niektorých aj číselný. Prvý operátor *Support Vector Machine* podporuje nasledovné typy jadrovej transformácie:

- bodové (*dot*): jadro je definované ako $k(x,y)=x*y$,
- radiálne (*radial*): jadro je definované ako $exp(-\gamma ||x-y||^2)$,
- polynomiálne (*polynomial*): $k(x,y)=(x*y+1)^d$, kde d je stupeň polynómu,
- neurónové (*neural*): $tanh(\alpha\ x*y + b)$, kde b je zastavovacia konštanta; $\alpha=1/N$, kde N je dimenzia dát,
- anova (*anova*): definované mocninou d výrazu $exp\ (-\gamma\ (x-y))$, kde d je stupeň polynómu,
- epanechnikov (*epanechnikov*): funkcia $¾(1-u2)$ pre $-1 < u < 1$ a 0, ak je u mimo tohto rozsahu,
- gaussovské združenie (*gaussian_combination*): gaussovské jadro s nastaviteľnými parametrami σ_1, σ_2, σ_3,
- multikvadratické (*multiquadratic*): definované ako $\sqrt{||x-y||^2 + c^2}$, kde c je konštanta zložitosti, $0 \leq c < +\infty$.

Príklad procesu aplikovaného na databázu Titanicu je znázornený na obr. 10.1. V procese je nastavený ako *label* atribút *survived* (dokáže spracovať aj číselný *label*), nominálne hodnoty sú prevedené na numerické a chýbajúce údaje doplnené priemernou hodnotou. Podproces krížovej validácie je na obr. 10.2 a výsledná presnosť 80,75% na obr. 10.3. Rozdelenie záznamov operátorom *SVM* je znázornené na obr. 10.4, kde modrá farba znamená, že pasažier prežil a červená znamená, že nie.



Obr. 10.1 Proces s aplikáciou SVM

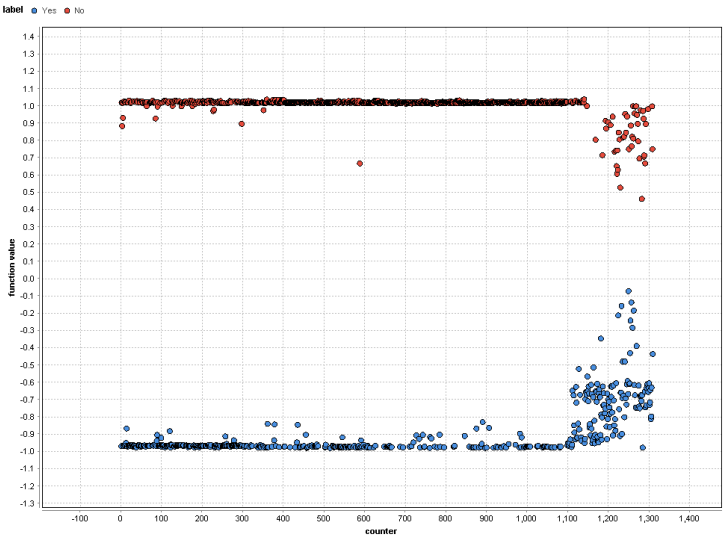


Obr. 10.2 Podproces krížovej validácie s operátorom SVM

accuracy: 80.75% +/- 2.57% (mikro: 80.75%)

	true Yes	true No	class precision
pred. Yes	252	4	98.44%
pred. No	248	805	76.45%
class recall	50.40%	99.51%	

Obr. 10.3 Výsledná presnosť modelu s operátorom SVM



Obr. 10.4 Rozdelenie záznamov operátorom SVM

Druhý operátor *Support Vector Machine (LibSVM)* používa knižnicu pre SVM od Chih-Chung Chang a Chih-Jen Lin. Tento operátor podporuje typy C-SVC a nu-SVC SVM pre klasifikačné úlohy, ako aj typy epsilon-SVR a nu-SVR SVM pre regresné úlohy. Jedno-triedový typ SVM poskytuje možnosť učiť sa z jedinej triedy príkladov a neskôr testovať či sa nové príklady zhodujú s tými známymi. Nespracuje číselný *label*.

Ďalší operátor *Support Vector Machine (Linear)* používa Java implementáciu mySVM od Stephana Rüpinga. Je obmedzený na bodové (lineárne) jadro, ale prináša výkonný model, ktorý obsahuje len lineárny koeficient pre rýchlejšiu aplikáciu modelu. Spracuje aj číselný *label*.

Operátor *Support Vector Machine (Evolutionary)* používa evolučnú stratégiu pre optimalizáciu. Tento operátor je implementácia SVM pomocou evolučného algoritmu na vyriešenie dvojitej optimalizácie SVM. Je tiež schopný učiť sa s rovnakými jadrami ako operátor *Support Vector Machine* a spracuje číselný *label*.

Support Vector Machine (PSO) je operátor, ktorý využíva optimalizáciu časticovým rojom (Particle Swarm Optimization). PSO je evolučná výpočtová metóda, ktorá modeluje proces podľa spoločenského správania kŕdľa vtákov. PSO algoritmy využívajú častice pohybujúce sa v *n*-dimenzionálnom priestore na hľadanie riešení pre problém s optimalizáciou funkcií *n*-premennej. Tento operátor je tiež schopný učiť sa s rovnakými jadrami ako operátor *Support Vector Machine*, ale nespracuje číselný *label*.

Fast Large Margin je predposledný operátor v tejto skupine. Používa rýchle učenie založené na schéme učenia lineárnych podporných vektorov, navrhnuté od R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang a C.J. Lin. Hoci je výsledok podobný výsledkom klasických SVM, tento lineárny klasifikátor je schopný pracovať na množine údajov s miliónmi príkladov a atribútov, ale nespracuje číselný *label*.

Posledný operátor v tejto skupine *Hyper Hyper* používa minimálnu implementáciu SVM. Model je postavený iba s jedným pozitívnym a jedným negatívnym príkladom. Ide o najjednoduchší operátor v tejto skupine. Ponúka len jeden nastaviteľný parameter a to použitie náhodného čísla. Ako *label* dokáže spracovať iba binominálny atribút.

11 FUNKCIE

RapidMiner Studio poskytuje v skupine funkcií osem operátorov, ktoré môžu byť použité na klasifikáciu alebo regresiu.

Prvý operátor v tejto skupine sa nazýva *Generalized Linear Model*. Generalizované lineárne modely (GLM) sú rozšírením tradičných lineárnych modelov tým, že maximalizujú logickú pravdepodobnosť. Výpočet modelu je extrémne rýchly a veľmi dobre sa hodí pre modely s obmedzeným počtom prediktorov s nenulovými koeficientmi. Dokáže spracovať číselné aj nominálne atribúty (čo platí aj pre *label*) a takisto aj chýbajúce dáta. Hlavný parameter tohto operátora je *family*. Určuje výber metódy vzhľadom na typ atribútu *label* a má nasledovné možnosti:

- *AUTO*: automatický výber,
- *gaussian*: pre numerické dáta,
- *binomial*: pre nominálne dáta s dvoma úrovňami,
- *multinomial*: pre polynominálne dáta s viac ako dvoma úrovňami,
- *poisson*: pre nezáporné numerické dáta,
- *gamma*: pre spojité a kladné numerické dáta,
- *tweedie*: pre spojité a nezáporné numerické dáta.

Pri použití tohto operátora na dátach z Titanicu bol nastavený *label* na atribút *survived* a operátor GLM umiestnený v tréningovej časti krížovej validácie. Výsledná presnosť 94,96% je znázornená na obr. 11.1.

accuracy: 94.96% +/- 1.24% (mikro: 94.96%)

	true Yes	true No	class precision
pred. Yes	443	9	98.01%
pred. No	57	800	93.35%
class recall	88.60%	98.89%	

Obr. 11.1 Presnosť modelu s použitím operátora GLM

Ďalší operátor *Linear Regression* sa používa na numerickú predikciu. Regresia je štatistický výpočet, ktorý sa pokúša určiť silu vzťahu medzi jednou závislou premennou (t. j. atribút *label*) a sériou ďalších meniacich sa premenných, známych ako nezávislé

premenné (bežné atribúty). Rovnako ako pri klasifikácii sa používa na predpovedanie atribútu *label*, ale regresia sa skôr používa na predpovedanie kontinuálnej hodnoty (číselnej). Operátor *Linear Regression* sa pokúša modelovať tento vzťah medzi atribútmi tak, že vytvorí lineárnu rovnicu k pozorovaným údajom. Nedokáže spracovať nominálne atribúty a chýbajúce dáta. Dokáže teda spracovať numerické atribúty, pričom *label* môže byť numerický alebo binominálny. Pri potrebe použitia nominálnych atribútov je možné aplikovať operátor *Nominal to Numerical*, ktorý transformuje dáta z nominálnych na numerické.

Operátor *Polnynomial Regression* používa polynomickú regresiu, čo je forma lineárnej regresie, v ktorej je vzťah medzi nezávislou premennou *x* a závislou premennou *y* modelovaný ako *n*-rádový polynóm. V RapidMiner Studio je *y* atribút *label* a *x* je súbor bežných atribútov, ktoré sa používajú na predpoveď *y*. Tento operátor nedokáže spracovať nominálne atribúty (ani nominálny *label*) a chýbajúce dáta.

Vector Linear Regression je operátor, ktorý vykonáva vektorovú lineárnu regresiu. Vytvára vektor atribútu *label* zo všetkých bežných atribútov. Tento operátor nedokáže spracovať nominálne atribúty a chýbajúce dáta. Dokáže pracovať len s číselnými atribútmi a číselným atribútom *label*.

Operátor *Local Polynomial Regression* poskytuje funkciu na vykonanie lokálnej regresie. To znamená, že ak sa požaduje hodnota *labelu* pre bod v dátovom priestore, tak sa vyhľadá miestne okolie tohto bodu. Pre toto vyhľadávanie sa používa meranie vzdialenosti špecifikované v parametri numerického merania. Po určení okolia sa jeho údaje používajú na vytvorenie vhodného polynómu (špecifikovaného stupňa) pomocou váženej optimalizácie najmenších štvorcov. Výpočty sú pomalšie ako pri SVM, lineárnej regresii alebo naivnému Bayesovi. Výpočtová rýchlosť závisí hlavne od počtu záznamov a atribútov. Tento operátor dokáže spracovať iba číselné atribúty bez chýbajúcich dát.

Ďalší operátor *Seemingly Unrelated Regression* vykonáva regresiu na viacerých súboroch s údajmi s rôznymi atribútmi *label* a berie do úvahy koreláciu zvyškov. Regresia môže byť vykonaná na rôznych atribútoch, ale všetky dodané príklady musia mať rovnaký počet záznamov. Hlavná množina obsahujúca spojenie všetkých atribútov vo všetkých podmnožinách sa musí napojiť na prvý vstupný port operátora. Na všetky ostatné porty môžu byť pripojené podmnožiny, pričom musí byť pripojená aspoň jedna podmnožina. Tento operátor dokáže spracovať iba číselné atribúty bez chýbajúcich dát.

Predposledný operátor v tejto skupine *Gaussian Regression* je implementáciou Gaussovho procesu, ktorý je pravdepodobnostnou metódou klasifikácie aj regresie. Gaussovský proces je stochastický proces, ktorého realizácie pozostávajú z náhodných hodnôt spojených s každým bodom v rozsahu časov (alebo priestoru) tak, že každá takáto náhodná premenná má normálne rozloženie. Tento operátor takisto ako predošlí dokáže spracovať iba číselné atribúty bez chýbajúcich dát.

Posledný operátor v tejto skupine je *Relevance Vector Machine*. Používa sa pri klasifikácii aj regresii využitím Bayesovej inferencie. RVM má podobnú funkčnú formu ako SVM, ale podporuje pravdepodobnostnú klasifikáciu. Tento operátor dokáže spracovať iba číselné atribúty bez chýbajúcich dát. Atribút *label* môže byť aj binominálny.

12 ZHLUKOVANIE

Zhlukovanie je podobné klasifikácii. Obe techniky rozdeľujú objekty do skupín alebo tried. Rozdielom pri zhlukovaní je, že triedy nie sú preddefinované. Metódy zhlukovania sa líšia aj podľa použitého typu metódy učenia. Klasifikačné metódy používajú učenie s učiteľom, zatiaľ čo zhlukovanie učenie bez učiteľa.

RapidMiner Studio 8.1 poskytuje vo voľnej verzii 13 operátorov pre zhlukovanie. Na uvedených príkladoch je použitá rovnaká databáza ako pri predikciách v predošlých kapitolách.

12.1 Operátory k-Means

Operátor *k-Means* vykonáva zhlukovanie pomocou algoritmu k-means. Zoskupuje dokopy záznamy, ktoré sú navzájom podobné. Keďže nie je potrebný žiaden atribút *label*, zhlukovanie sa môže použiť na neoznačených údajoch a je to algoritmus strojového učenia bez učiteľa.

Zhluk je určený polohou centra v *n*-dimenzionálnom priestore všetkých záznamov. Algoritmus začína bodmi, ktoré sa považujú za centroid potenciálnych zhlukov. Tieto začiatočné body môžu byť polohou náhodného záznamu (bod môže byť aj imaginárny), alebo sú určené heuristicky, ak je to zvolené v parametroch operátora.

Algoritmus priraďuje každý záznam do najbližšieho zhluku, ktorý pozostáva zo záznamov, medzi ktorými je podobnosť. Táto podobnosť je založená na meraní vzdialenosti medzi nimi. Potom sú centroidy zhlukov prepočítané spriemerovaním všetkých záznamov jedného zhluku. Predchádzajúce kroky sa opakujú, kým sa centroidy už nepohybujú, alebo sa nedosiahnu maximálne optimalizačné kroky.

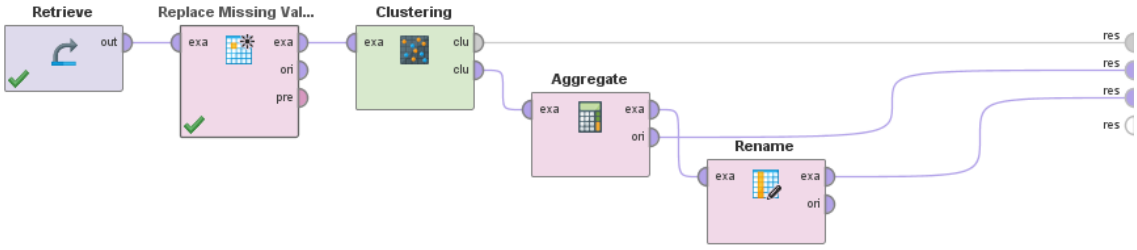
Počet vytvorených zhlukov patrí medzi hlavné parametre operátora, ktoré sú znázornené v tabuľke 12.1.

Tabuľka 12.1 Hlavné parametre operátora k-Means

Parameter	**Popis**
add_cluster_attribute	Pridá atribút *cluster* a *id*.
add_as_lalel	Atribút *cluster* premenuje na label a zmení jeho rolu na *label*.
k	Počet zhlukov, ktoré sa majú vytvoriť.

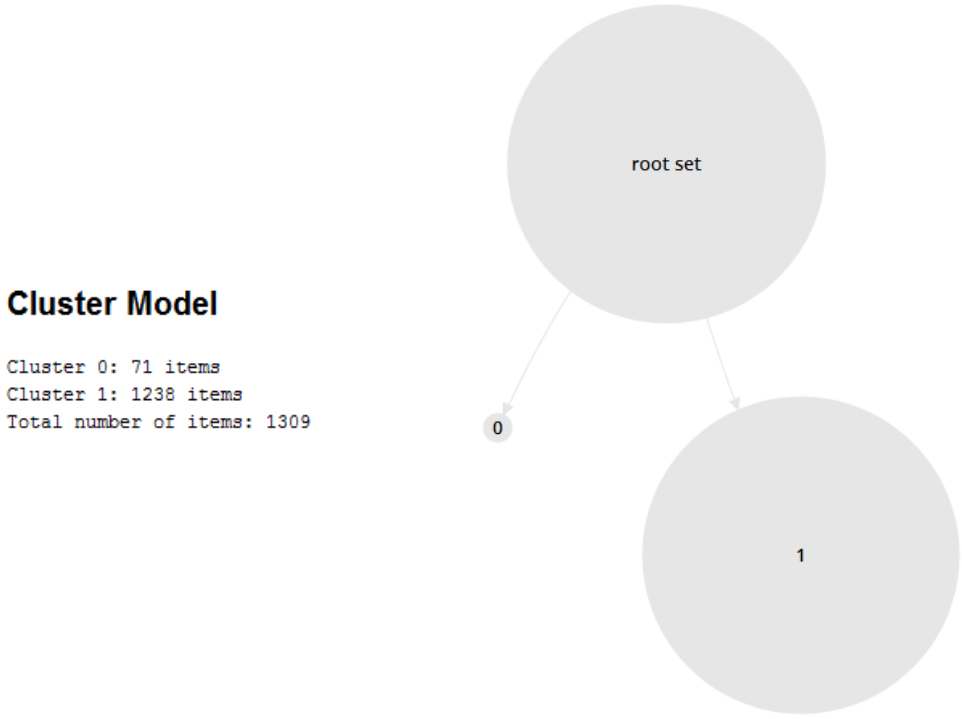
Parameter	Popis
max_runs	Maximálny počet, koľko krát sa má spustiť algoritmus s náhodným začiatkom.
measure_types	Výber typu výpočtu, ktorý sa má použiť. Možnosti sú: výpočet pre nominálne aj numerické atribúty, iba pre nominálne, iba numerické a BregmannDivergences.

Celkový proces zhlukovania s použitím operátora *k-Means* je znázornený na obr. 12.1. Parameter *k* bol zvolený 2.



Obr. 12.1 Proces zhlukovania s použitím operátora k-Means

V procese je použitý operátor *Replace Missing Values*, kvôli doplneniu chýbajúcich dát, keďže operátor *k-Means* nedokáže spracovať záznamy s chýbajúcimi dátami. Na prvom výstupe je model zhlukovania, ktorý je znázornený na obr. 12.2.



Obr. 12.2 Výsledný model zhlukovania s použitím operátora k-Means (vľavo číselne, vpravo graficky)

Originálne dáta po zhlukovaní vo forme tabuľky sú na druhom výstupe procesu a znázornená je ich časť na obr. 12.3.

Row No.	id	cluster	Passenger ...	Name	Sex	Age	No of Sibling...	No of Parent...	Ticket Number
1	1	cluster_0	First	Allen, Miss. E...	Female	29	0	0	24160
2	2	cluster_0	First	Allison, Mast...	Male	0.917	1	2	113781
3	3	cluster_0	First	Allison, Miss. ...	Female	2	1	2	113781
4	4	cluster_0	First	Allison, Mr. H...	Male	30	1	2	113781
5	5	cluster_0	First	Allison, Mrs. ...	Female	25	1	2	113781
6	6	cluster_1	First	Anderson, Mr....	Male	48	0	0	19952
7	7	cluster_1	First	Andrews, Mis...	Female	63	1	0	13502
8	8	cluster_1	First	Andrews, Mr. ...	Male	39	0	0	112050
9	9	cluster_1	First	Appleton, Mrs...	Female	53	2	0	11769
10	10	cluster_1	First	Artagaveytia, ...	Male	71	0	0	PC 17609

Obr. 12.3 Dáta po zhlukovaní s operátorom k-Means

Pre jednoduchšiu vizualizáciu dát sú dáta na treťom výstupe spracované operátorom *Aggregate* a *Rename*. Operátor *Aggregate* spočíta *id* v zhlukoch a zoskupí dáta podľa zvolených atribútov: *cluster* a *Survived*. Operátor *Rename* len dodatočne premenuje názov vytvoreného atribútu spočítaných *id*. Vo výslednej tabuľke je znázornené zastúpenie pasažierov, ktorý prežili alebo neprežili v daných zhlukoch aj s ich počtom. Výber atribútu *Survived* je náhodný a zvolený len pre znázornenie počtu záznamov. Takisto je možné znázorniť aj napríklad zastúpenie pohlavia v daných zhlukoch. Výsledné spracované dáta sú na obr. 12.4.

Row No.	Survived	cluster	count
1	No	cluster_0	20
2	No	cluster_1	789
3	Yes	cluster_0	51
4	Yes	cluster_1	449

Row No.	Sex	cluster	count
1	Female	cluster_0	45
2	Female	cluster_1	421
3	Male	cluster_0	26
4	Male	cluster_1	817

Obr. 12.4 Výsledné spracované dáta pri použití operátora k-Means (vľavo podľa prežitia, vpravo podľa pohlavia pasažiera)

Podobne operátor *k-Means (Kernel)* vykonáva zhlukovanie, ale s pridaným parametrom jadra. Výber typu jadra je možný z týchto možností:

- bodové (*dot*): jadro je definované ako $k(x,y)=x*y$,
- radiálne (*radial*): jadro je definované ako $exp(-\gamma ||x-y||^2)$,

- polynomiálne (*polynomial*): $k(x,y)=(x*y+1)^d$, kde *d* je stupeň polynómu,
- neurónové (*neural*): *tanh(α x*y + b)*, kde *b* je zastavovacia konštanta; *α=1/N*, kde *N* je dimenzia dát,
- anova (*anova*): definované mocninou *d* výrazu *exp (-γ (x-y))*, kde *d* je stupeň polynómu,
- epanechnikov (*epanechnikov*): funkcia *¾(1-u2)* pre -1 < *u* < 1 a 0, ak je *u* mimo tohto rozsahu,
- gaussovské združenie (*gaussian_combination*): gaussovské jadro s nastaviteľnými parametrami σ_1, σ_2, σ_3,
- multikvadratické (*multiquadratic*): definované ako $\sqrt{||x-y||^2+c^2}$, kde *c* je konštanta zložitosti, $0 \leq c < +\infty$.

Taktiež tento operátor nedokáže spracovať chýbajúce dáta a môže pracovať iba s číselnými hodnotami.

Rýchlejším ekvivalentom operátora *k-Means* je operátor *k-Means (fast)*. Tento operátor dokáže spracovať nominálne aj číselné hodnoty, ale nedokáže pracovať s chýbajúcimi dátami. Na rozdiel od štandardnej implementácie *k-Means* je v mnohých prípadoch rýchlejší, hlavne pri dátach s mnohými atribútmi a vysokou hodnotou požadovaných zhlukov *k*. K vyššej rýchlosti potrebuje viac pamäte.

Ďalší operátor *X-Means* je operátor, ktorý určuje správny počet centroidov na základe heuristiky. Začína s minimálnou skupinou centroidov a potom sa iteračne rozrastá. Dokáže spracovať nominálne aj číselné hodnoty, ale nie chýbajúce dáta. V parametroch operátora je možné vybrať medzi algoritmami k-Means a k-Means (fast).

Operátor *k-Medoids* vykonáva zhlukovanie podobne ako operátor *k-Means*, ale stred zhluku bude vždy jeden z bodov zhluku. Taktiež dokáže spracovať nominálne aj číselne hodnoty, ale nie chýbajúce dáta.

12.2 Operátor DBSCAN

Tento operátor vykonáva zhlukovanie pre „Density-Based Spatial Clustering of Applications with Noise“ – priestorové zhlukovanie založené na hustote aplikácií so šumom. Je to algoritmus založený na hustote, pretože vytvára zhluky z odhadovanej distribúcie hustoty zodpovedajúcich uzlov.

Definícia zhluku je založená na pojme dosiahnutia hustoty. Bod q je priamo hustotou dosiahnuteľný od bodu p, ak nie je vzdialenejší ako daná vzdialenosť *epsilon*. *DBSCAN* vyžaduje dva parametre: *epsilon* a minimálny počet bodov potrebných na vytvorenie zhluku. Algoritmus začína ľubovoľným východiskovým bodom. V tomto bode je načrtnuté okolie *epsilon* a ak obsahuje dostatok bodov, zhluk je spustený. V opačnom prípade je bod označený ako šum. Tento bod by sa neskôr mohol nachádzať v dostatočne veľkom prostredí *epsilon* iného bodu a teda byť súčasťou zhluku. Ak sa zistí, že bod je súčasťou zhluku, tak sa pridajú všetky body. Tento proces pokračuje až do úplného nájdenia zhluku. Následne sa načíta a spracuje nový bod, čo vedie k objaveniu ďalšieho zhluku alebo šumu.

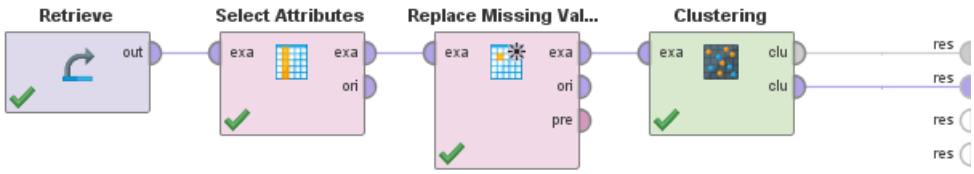
Operátor *DBSCAN* dokáže spracovať nominálne aj číselné údaje, ale nedokáže pracovať s chýbajúcimi dátami.

12.3 Operátor maximalizácie očakávania

V programe RapidMiner Studio je to operátor *Expectation Maximization Clustering* (*EMC*). Rozširuje základný prístup zhlukovania v niektorých dôležitých smeroch.

Technika maximalizácie očakávania je podobná technike *k-Means*. Rozširuje základný prístup priraďovania bodov do pevných počtov zhlukov *k*. Priraďovanie bodov do zhlukov v technike *k-Means* je s cieľom maximalizovať rozdiely, pričom technika maximalizácie očakávania vypočítava pravdepodobnosť členstva v zhlukoch. Cieľom algoritmu zhlukovania je potom maximalizovať celkovú pravdepodobnosť údajov vzhľadom na konečné zhluky. Inými slovami, každý záznam patrí ku každému zhluku s určitou pravdepodobnosťou.

Operátor *EMC* dokáže spracovať iba číselné údaje bez chýbajúcich dát. Proces zhlukovania s použitím tohto operátora je znázornený na obr. 12.5. Po načítaní dát sa vyberú iba číselné dáta a doplnia chýbajúce.

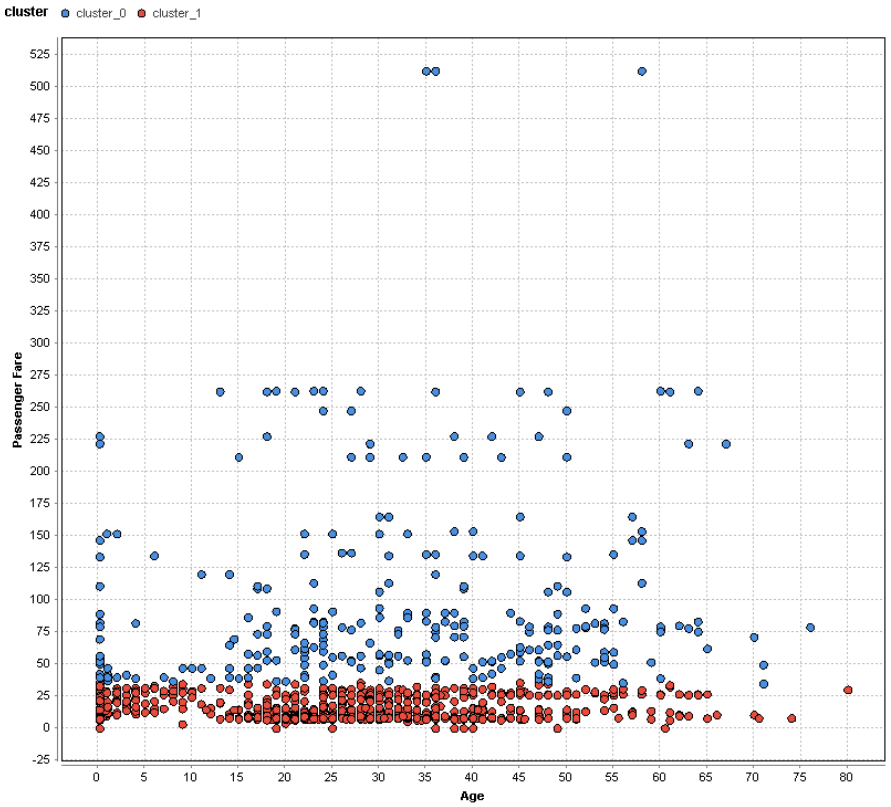


Obr. 12.5 Proces zhlukovania s použitím operátora EMC

Na obr. 12.6 sú znázornené pravdepodobnosti záznamov k jednotlivým zhlukom. Počet zhlukov *k* bol zvolený 2. Grafické zobrazenie zhlukovania s použitím operátora *EMC* je na obr. 12.7. Celkový počet záznamov bol 1039, pričom v zhluku 0 bolo 293 záznamov a v zhluku 1: 1016 záznamov.

id	cluster	cluster_0_probability	cluster_1_probability	Age	Passenger Fare
1	cluster_0	1	0	29	211.338
2	cluster_0	1	0	0.917	151.550
3	cluster_0	1	0	2	151.550
4	cluster_0	1	0	30	151.550
5	cluster_0	1	0	25	151.550
6	cluster_1	0.101	0.899	48	26.550
7	cluster_0	1.000	0.000	63	77.958
8	cluster_1	0.120	0.880	39	0
9	cluster_0	0.999	0.001	53	51.479
10	cluster_0	0.998	0.002	71	49.504

Obr. 12.6 Dáta po zhlukovaní s operátorom EMC



Obr. 12.7 Grafické zobrazenie zhlukovania s použitím operátora EMC

12.4 Operátor SVC

Operátor *Support Vector Clustering* (*SVC*) je implementáciou zhlukovania založeného na podporných vektoroch. Dátové body sú mapované z dátové priestoru do inej dimenzie. V tomto priestore sa hľadá najmenšia oblasť, ktorá obklopuje obraz údajov. Táto oblasť je mapovaná späť do dátového priestoru, kde vytvára súbor kontúr, ktoré obklopujú dátové body. Tieto kontúry sú interpretované ako hranice zhluku. Body obklopené jednotlivými kontúrami sú spojené s tým istým zhlukom. Zmenou parametrov operátora sa mení počet zhlukov. Hlavnými parametrami operátora *SVC* sú: minimálny počet bodov v zhluku a výber jadra. Možnosti výberu typu jadra sú nasledovné:

- bodové (*dot*): jadro je definované ako $k(x,y)=x*y$,
- radiálne (*radial*): jadro je definované ako $exp(-\gamma ||x-y||^2)$,
- polynomiálne (*polynomial*): $k(x,y)=(x*y+1)^d$, kde d je stupeň polynómu,
- neurónové (*neural*): $tanh(\alpha\ x*y + b)$, kde b je zastavovacia konštanta; $\alpha=1/N$, kde N je dimenzia dát.

Operátor *SVC* dokáže spracovať iba číselné údaje bez chýbajúcich dát. V prípade, ak záznamy nepatria do žiadneho zhluku, tak ich operátor považuje za šum. Zhluk so všetkým šumom je následne nazvaný *noise*. Ukážka dát po zhlukovaní s operátorom *SVC* je na obr. 12.8.

id	cluster	Age	Passenger Fare
485	noise	34	10.500
486	noise	36	12.875
487	cluster_2	24	10.500
488	noise	61	12.350
489	noise	50	26
490	noise	42	26
491	noise	57	10.500
492	cluster_4	0.167	15.046
493	noise	1	37.004
494	noise	31	37.004
495	noise	24	37.004
496	noise	0.167	15.579
497	cluster_1	30	13
498	noise	40	16
499	noise	32	13.500
500	cluster_1	30	13
501	noise	46	26
502	noise	13	19.500
503	noise	41	19.500
504	cluster_2	19	10.500
505	noise	39	13

Obr. 12.8 Dáta po zhlukovaní s operátorom SVC

12.5 Operátor náhodného zhlukovania

Operátor *Random Clustering* vykonáva náhodné zhlukovanie daných záznamov. Tento algoritmus nezaručuje, že všetky zhluky nebudú prázdne. Náhodne priraďuje záznamy do zhlukov. Pre správne zhlukovanie sa odporúča použiť iný algoritmus, ktorý implementuje zhlukovanie napríklad operátorom *k-Means*.

Tento operátor má štyri základné parametre:

- add_cluster_attribute – pridá atribút *cluster* a *id*,
- add_as_label – zmení rolu atribútu *cluster* na *label*,
- remove_unlabeled – neoznačené záznamy vymaže,
- number_of_clusters – počet požadovaných zhlukov, ktoré sa majú vytvoriť.

Proces s použitím operátora *Random Clustering* je znázornený na obr. 12.9. Keďže dokáže spracovať číselné aj nominálne hodnoty, tak nie je nutné v procese zapojiť operátor na výber atribútov. Taktiež dokáže spracovať aj chýbajúce dáta.

Obr. 12.9 Proces zhlukovania s použitím operátora Random Clustering

12.6 Operátor Agglomerative Clustering

Tento operátor vykonáva hierarchické zhlukovanie zdola-nahor. Podporuje tri rôzne stratégie:

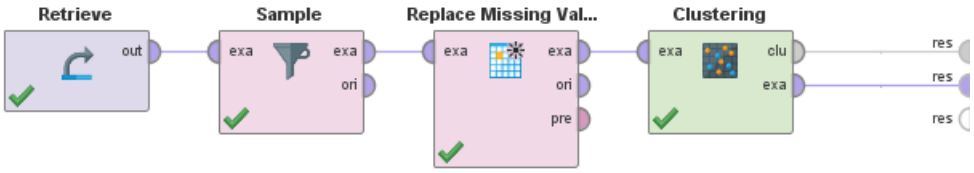
- single link – spája v každom kroku dva zhluky, ktorých najbližšie členy majú najmenšiu vzdialenosť (alebo dva zhluky s najmenšou minimálnou vzdialenosťou párov),
- complete link – spája zhluky, ktorých zlúčenie má najmenší priemer (alebo dva zhluky s najmenšou maximálnou vzdialenosťou párov),
- average link – kompromis medzi citlivosťou *complete link* stratégie na odľahlé hodnoty a tendenciou vytvárať dlhé reťazce *single link* stratégie.

Výsledkom tohto operátora je model hierarchického zhlukovania, ktorý poskytuje informáciu o vzdialenosti v dendrograme. V dendrograme os *y* označuje vzdialenosť, s ktorou sa zhluky spájajú, pričom objekty sú umiestnené pozdĺž osi *x*.

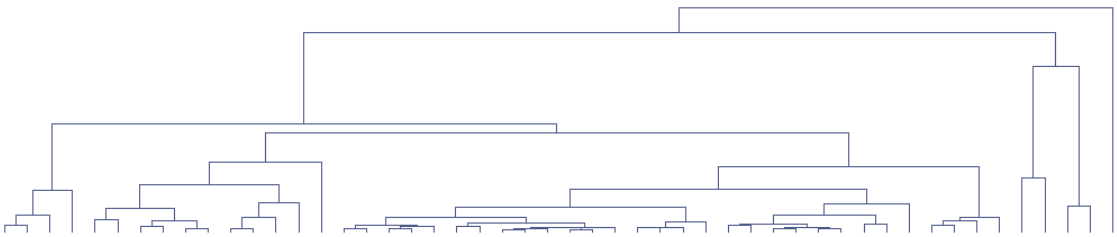
Hierarchické zhlukovanie sa rozdeľuje do dvoch typov:

- Aglomeratívne: Ide o prístup zdola-nahor, kde sú jednotlivé zhluky iteratívne spájané do väčších celkov. Každý objekt patrí do vlastného zhluku a dva najbližšie vytvoria nový zhluk.
- Divízne: Prístup je opačný (zhora-nadol) ako pri aglomeratívnom zhlukovaní. Na začiatku sú všetky objekty zaradené do jedného zhluku. Zhluk sa postupne delí, až kým nezostanú jednoprvkové zhluky.

Operátor *Agglomerative Clustering* dokáže spracovať nominálne aj číselné údaje, ale nie chýbajúce dáta. Pre lepšiu čitateľnosť výsledného dendrogramu je použitý operátor na výber menšieho počtu záznamov. Celkový proces je znázornený na obr. 12.10 a výsledný dendrogram na obr. 12.11.



Obr. 12.10 Proces zhlukovania s použitím operátora Agglomerative Clustering



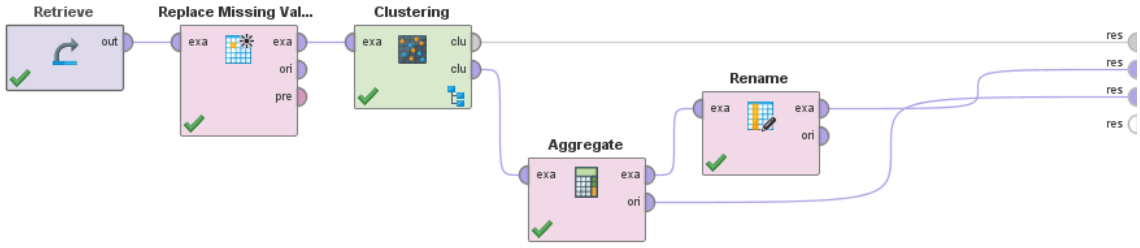
Obr. 12.11 Dendrogram pre zhlukovanie s použitím operátora Agglomerative Clustering

12.7 Operátor Top Down Clustering

Tento operátor vykonáva zhlukovanie zhora-nadol tým, že rekurzívne uplatňuje vnútornú schému zhlukovania. Má podproces, v ktorom musí byť operátor zhlukovania, ako napríklad *k-Means*. Výsledkom tohto operátora je model hierarchického zhlukovania. Keďže má podproces, tak dokáže spracovať iba taký formát dát, ktorý podporuje spracovať použitý operátor v ňom. Avšak nedokáže pracovať s chýbajúcimi dátami.

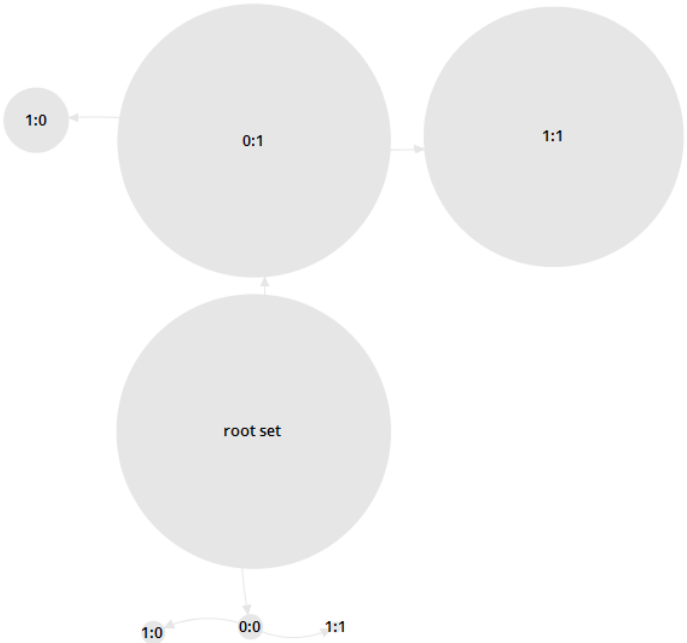
Tento operátor poskytuje dva základné parametre. Jeden na vytvorenie atribútu *cluster*, ako pri predošlých operátoroch a druhý na určenie maximálnej hĺbky zhlukovania.

Celkový proces zhlukovania s použitím operátora *Top Down Clustering* a jeho podproces je znázornený na obr. 12.12 a obr. 12.13. Výsledný model hierarchického zhlukovania je na obr. 12.14.



Obr. 12.12 Proces zhlukovania s použitím operátora Top Down Clustering

Obr. 12.13 Podproces operátora Top Down Clustering



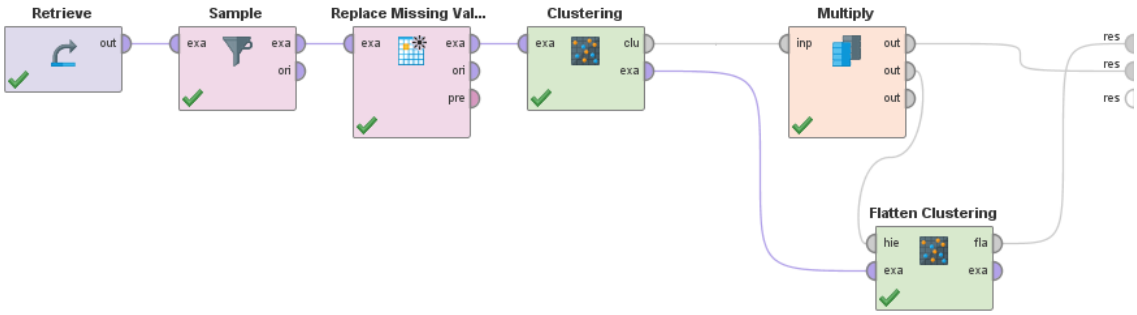
Obr. 12.14 Výsledný model hierarchického zhlukovania s použitím operátora Top Down Clustering

12.8 Operátor Flatten Clustering

Tento operátor vytvára model plochého zhlukovania, t. j. nehierarchický model z hierarchického modelu. Rozširuje uzly hierarchického modelu, až kým sa nedosiahne požadovaný počet zhlukov, určený v parametroch operátora.

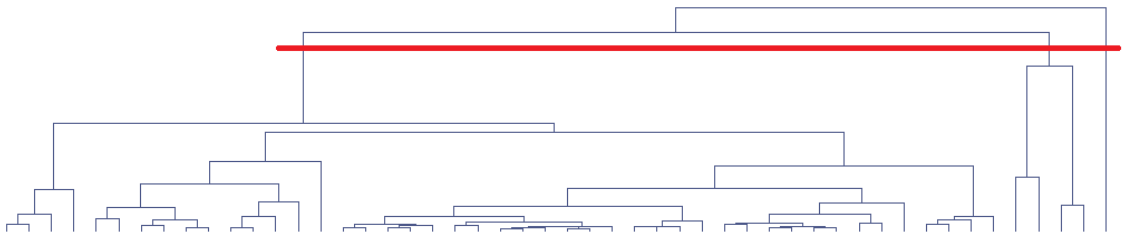
Keďže potrebuje na vstupe hierarchický model, môže byť zapojený iba za operátor *Agglomerative Clustering* alebo *Top Down Clustering*. Z čoho vyplýva, že dokáže spracovať číselné aj nominálne hodnoty, ale nie chýbajúce dáta.

Celkový proces s použitím operátora *Flatten Clustering* je na obr. 12.15. Operátor je zapojený do procesu z hierarchického zhlukovania z obr. 12.10.



Obr. 12.15 Proces plochého zhlukovania s použitím operátora Flatten Clustering

Pri zvolenom počte zhlukov 3, operátor z hierarchického modelu na obr. 12.16 vytvorí model plochého zhlukovania, ktorý je na obr. 12.17.



Obr. 12.16 Rozdelenie dendrogramu na tri zhluky operátorom Flatten Clustering

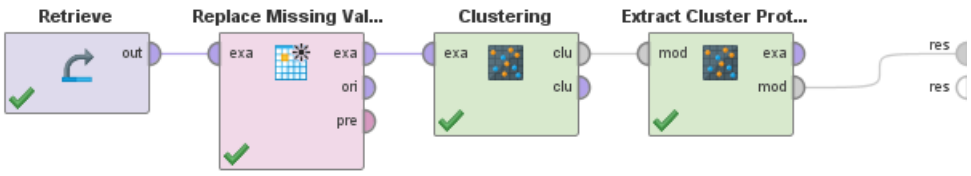
```
Cluster 0: 4 items
Cluster 1: 1 items
Cluster 2: 45 items
Total number of items: 50
```

Obr. 12.17 Výsledný plochý model zhlukovania operátora Flatten Clustering

12.9 Operátor Extract Cluster Prototypes

Tento operátor generuje prototypy modelu zhlukovania. Aplikuje sa po operátoroch zhlukovania, ktoré vytvárajú centroidný model. V programe RapidMiner Studio sú to operátory: *k-Means*, *k-Means (fast)*, *k-Medoids* a *X-Means*. Operátor *Extract Cluster Prototypes* extrahuje prototypy, ktoré uloží do formy tabuľky a modelu pre ďalšie možné použitie. Dokáže spracovať nominálne aj číselné údaje, ale nie chýbajúce dáta. Keďže iba extrahuje prototypy z predošlého operátora, tak nemá žiadne parametre.

Príklad zapojenia operátora *Extract Cluster Prototypes* s operátorom *k-Means* je na obr. 12.18. Výsledný model prototypov je znázornený na obr. 12.19.



Obr. 12.18 Proces extrahovania prototypov operátorom Extract Cluster Prototypes

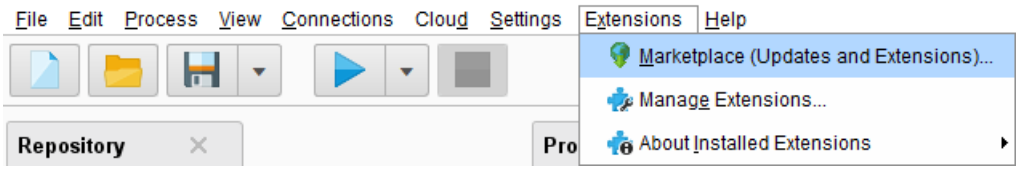
Attribute	cluster_0	cluster_1
Passenger Class	1	2.369
Name	152.930	683.013
Sex	1.366	1.660
Age	35.309	29.570
No of Siblings or Spouses on Board	0.676	0.489
No of Parents or Children on Board	0.915	0.355
Ticket Number	40.014	455.338
Passenger Fare	208.996	23.219
Cabin	73.070	69.894
Port of Embarkation	1.521	1.387
Life Boat	9.099	15.298
Survived	1.282	1.637

Obr. 12.19 Model extrahovaných prototypov operátorom Extract Cluster Prototypes

13 ROZŠÍRENIA V RAPIDMINER STUDIO

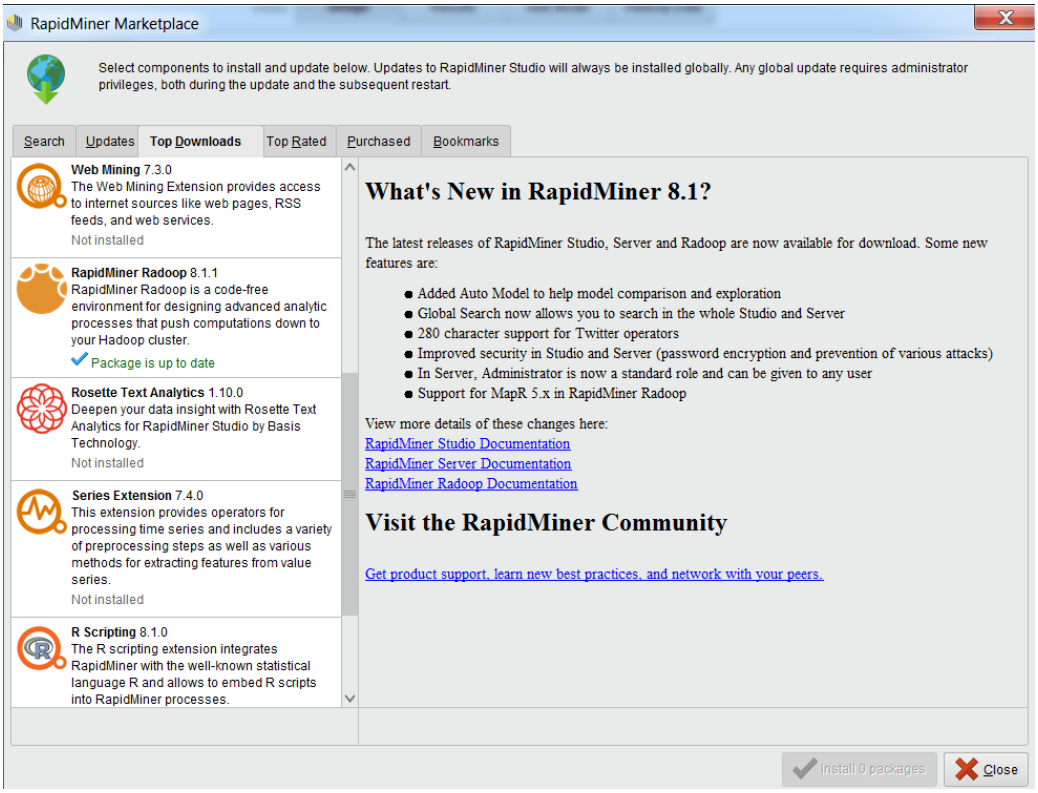
RapidMiner Studio poskytuje pre používateľov rozšírenia, ktoré si môže voľne stiahnuť. Rozšírenia zahŕňajú operátory špecifické pre určitú oblasť, ako napríklad textové spracovanie, dolovanie web stránok, jazyk R, Python, Radoop atď.

Rozšírenia sú dostupné v záložke *Extensions – Marketplace (Updates and Extensions)*. Umiestnenie záložky rozšírenia je zobrazené na obr. 13.1.



Obr. 13.1 Záložka rozšírenia

Po otvorení okna *Marketplace* sa zobrazia možnosti hľadania rozšírenia. Možné je vyhľadať konkrétne rozšírenie, dostupné aktualizácie pre už stiahnuté rozšírenia, alebo zobraziť najviac sťahované rozšírenia. Príklad zobrazenia okna *Marketplace* je na obr. 13.2. Výberom rozšírenia a potvrdením inštalácie sa rozšírenie stiahne a nainštaluje.



Obr. 13.2 Okno Marketplace

13.1 RapidMiner Radoop

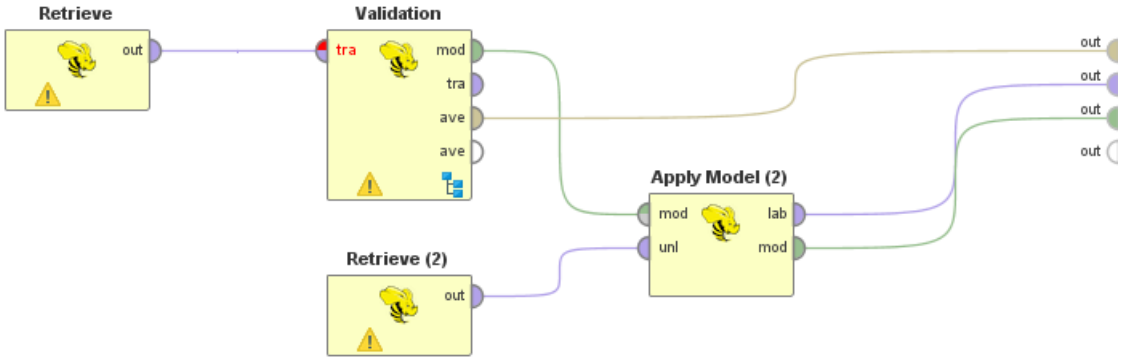
RapidMiner Radoop je rozšírenie, ktoré poskytuje ľahko použiteľné grafické rozhranie na analýzu veľkých dát na *Hadoop* klastroch. *Radoop* vyžaduje, aby *Hadoop* klaster bol prístupný z klienta *RapidMiner Studio*.

Po inštalácii rozšírenia *Radoop* bude pridaných 71 operátorov a panel *Hadoop Data*, v ktorom je možné zobraziť dostupné dáta.

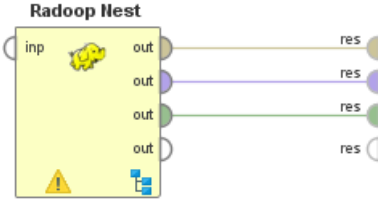
Operátory poskytujú rôzne funkcie na prácu s dátami, pričom využívajú rôzne technológie, ako napríklad *Hive*, *Apache Spark*, *Mahout* na prípravu a modelovanie dát.

Pre prácu s dátami je potrebné spojenie s *Hadoop* klastrom, ktoré je možné nastaviť v záložke *Connections – Manage Radoop Connections*.

Vytvorenie procesu začína použitím operátora *Radoop Nest*. Tento operátor zabezpečuje, že všetko prebieha na *Hadoop* klastri a nie na lokálnom počítači. Preto je nutné nastaviť spojenie v parametri operátora *connection*, ktoré bolo predtým vytvorené. Následne ďalšie operátory, ktoré poskytujú rôzne funkcie sa umiestňujú do podprocesu operátora *Radoop Nest*. Príklad zapojenia v podprocese môže vyzerať ako je na obr. 13.3. Z hlavného procesu následne vedú tri výstupy ako je znázornené na obr. 13.4.



Obr. 13.3 Podproces operátora Radoop Nest

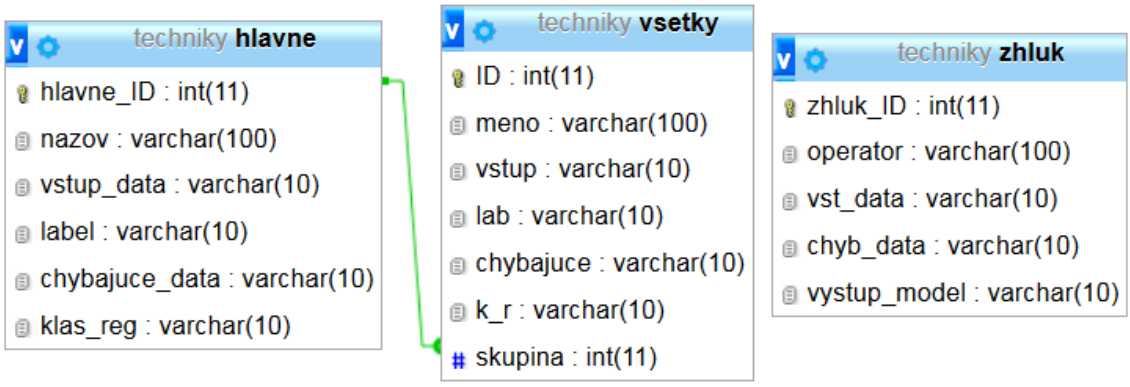


Obr. 13.4 Hlavný proces s operátorom Radoop Nest

14 VÝBER TECHNIKY

RapidMiner Studio poskytuje veľké množstvo operátorov. Preto sme vytvorili stránku, ktorá používateľovi pomôže s výberom.

Prvým krokom bolo vytvorenie databázy. Databáza sa skladá z troch tabuliek. Prvá tabuľka obsahuje skupiny na predikciu, druhá tabuľka jednotlivé operátory na predikciu a tretia tabuľka operátory pre zhlukovanie. Štruktúra databázy a jej vzťahy sú znázornené na obr. 14.1.



Obr. 14.1 Štruktúra databázy

Prvá tabuľka sa skladá z:

- *hlavne_ID* – je to číslo príslušnej skupiny - primárny kľúč,
- *nazov* – názov príslušnej skupiny,
- *vstup_data* – formát vstupných dát, ktoré skupina podporuje,
- *label* – typ predpovedaného atribútu, ktorý skupina podporuje,
- *chybajuce_data* – či skupina dokáže spracovať chýbajúce dáta alebo nie,
- *klas_reg* – typ výstupného modelu.

Štruktúra tejto tabuľky je na obr. 14.2. Jednotlivé záznamy majú priradené čísla, ktorých význam je v tabuľke 14.1. Skupina s viacerými číslami v atribúte poskytuje možnosť pracovania s viacerými typmi dát. Ak používateľovi nechýbajú dáta, môže byť vybraná skupina, ktorá chýbajúce dáta spracuje aj nespracuje, keďže žiadne dáta nechýbajú.

Tabuľka 14.1 Význam čísiel v tabuľke skupín

Atribút	**Význam**
vstup_data	1 – číselné dáta, 2 – nominálne dáta, 3 – kombinácia
label	1 – číselný, 2 – polynominálny, 3 - binominálny
chybajuce_data	1 – spracuje, 2 - nespracuje
klas_reg	1 – klasifikácia, 2 – regresia, 3 - pravidlá

hlavne_ID	nazov	vstup_data	label	chybajuce_data	klas_reg
1	Lazy	123	123	12	12
2	Naive Bayes	123	23	12	1
3	Neural Nets	123	123	12	12
4	Logistic Regression	123	3	12	2
5	Trees	123	123	12	12
6	Rules	123	23	12	3
7	SVM	1	123	2	12
8	Functions	123	123	12	12
9	Discriminant Analysis	1	23	2	1

Obr. 14.2 Štruktúra tabuľky skupín predikcie

Druhá tabuľka databázy, v ktorej sú všetky operátory na predikciu sa skladá z:

- *ID* – číslo operátora – primárny kľúč,
- *meno* – názov operátora,
- *vstup* - formát vstupných dát, ktoré operátor podporuje,
- *lab* - typ predpovedaného atribútu, ktorý operátor podporuje,
- *chybajuce* - či operátor dokáže spracovať chýbajúce dáta alebo nie,
- *k_r* - typ výstupného modelu,
- *skupina* – príslušná skupina do ktorej patrí operátor – cudzí kľúč.

Časť tabuľky je znázornená na obr. 14.3. Význam čísiel v záznamoch je rovnaký ako pre tabuľku skupín v tabuľke 14.1.

ID	meno	vstup	lab	chybajuce	k_r	skupina
1	Default Model	123	123	12	12	1
2	k-NN	123	123	12	12	1
3	Naive Bayes	123	23	12	1	2
4	Naive Bayes (Kernel)	123	23	2	1	2
5	Deep Learning	123	123	12	12	3
6	Neural Net	1	123	2	12	3
7	AutoMLP	1	23	2	1	3
8	Perceptron	1	3	2	1	3

Obr. 14.3 Časť tabuľky operátorov predikcie

Tretia tabuľka databázy obsahujúca operátory zhlukovania sa skladá z:

- *zhluk_ID* - číslo operátora – primárny kľúč,
- *operator* - názov operátora,
- *vst_data* - formát vstupných dát, ktoré operátor podporuje,
- *chyb_data* - či operátor dokáže spracovať chýbajúce dáta alebo nie,
- *vystup_model* - typ výstupného modelu.

Časť tejto tabuľky je na obr. 14.4. Jednotlivé záznamy majú priradené čísla, ktorých význam je v tabuľke 14.2.

zhluk_ID	operator	vst_data	chyb_data	vystup_model
1	k-Means	123	2	2
2	k-Means (Kernel)	1	2	1
3	k-Means (fast)	123	2	2
4	X-Means	123	2	2
5	k-Medoids	123	2	2

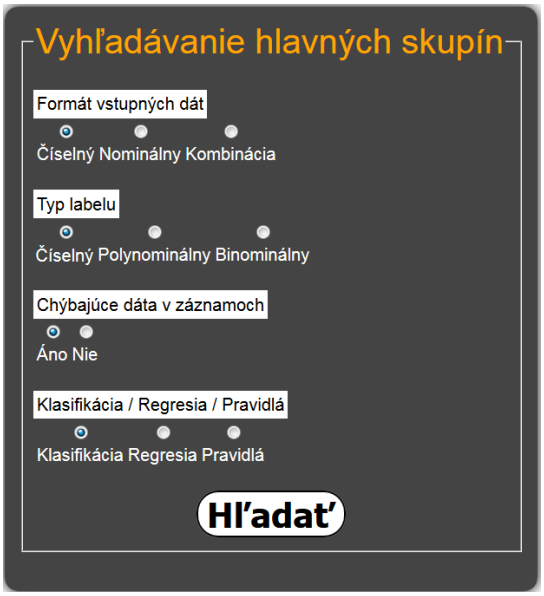
Obr. 14.4 Časť tabuľky operátorov zhlukovania

Tabuľka 14.2 Význam čísiel v tabuľke operátorov zhlukovania

Atribút	**Význam**
vst_data	1 – číselné dáta, 2 – nominálne dáta, 3 – kombinácia
chyb_data	1 – spracuje, 2 - nespracuje
vystup_model	1 – cluster, 2 – centroid, 3 – heirarchický, 4 - prototyp

Druhým krokom po vytvorení databázy bolo vytvorenie hlavnej stránky. Stránka je zložená z dvoch častí. V prvej časti je vyhľadávanie pre operátory predikcie a v druhej pre operátory zhlukovania. Zdrojový kód stránky bol písaný v programe *Visual Studio Code*. Pre prácu s databázou a stránkou sa použil program *XAMPP*, ktorý dokáže vytvoriť server na lokálnom počítači.

Prvá časť stránky na vyhľadávanie operátorov predikcie sa skladá z viacerých vstupných polí. Používateľ si môže vybrať s akými dátami disponuje, aký model potrebuje a spustiť vyhľadávanie. Táto časť stránky je znázornená na obr. 14.5. Rovnako funguje druhá časť stránky, ktorá je na obr. 14.6.



Obr. 14.5 Prvá časť stránky na vyhľadávania skupín predikcie

Obr. 14.6 Druhá časť stránky na vyhľadávanie operátorov zhlukovania

Ďalším krokom bolo vytvorenie funkcií, ktoré spracujú požiadavky od používateľov a zobrazia požadovaný výsledok. Pri vytváraní týchto stránok a funkcií bolo použité:

- *HTML* – pre prezentáciu stránky a vytvorenie vstupného formuláru,
- *PHP* – pre spracovanie zadaných hodnôt od používateľa,
- *SQL* – pre prácu s databázou,
- *JavaScript* – pre dynamickú prácu so stránkou,
- *CSS* – pre upravenie štýlu stránky.

Pri vybratí možností používateľom sa príslušné čísla odošlú na ďalšiu stránku, kde sa spracujú a porovnajú s hodnotami v databáze. Pre vyhľadávanie operátorov predikcie si musí používateľ vybrať okrem vstupných parametrov aj výslednú skupinu, z ktorej sa následne vyhľadajú operátory podľa zvolených parametrov. Skupiny operátorov sa taktiež vyhľadávajú podľa zvolených parametrov. Tieto skupiny sú zobrazené po kliknutí na tlačidlo *Hľadať* na ďalšej stránke. Používateľ kliknutím na vybranú skupinu spustí vyhľadávanie operátorov iba vo vybranej skupine so zvolenými parametrami, ktoré zadal na hlavnej stránke. Vyhľadané operátory sa zobrazia na ďalšej stránke. Po kliknutí na názov operátora si môže používateľ stiahnuť príslušný proces s týmto operátorom. Príklad vyhľadávania je na obr. 14.7. Vstupné hodnoty boli zvolené: kombinácia formátu

vstupných dát, polynominálny atribút predikcie, chýbajúce dáta v záznamoch, model klasifikácie a následne vybraná skupina *Naive Bayes*. Výsledný operátor z tejto skupiny je iba jeden a to *Naive Bayes* operátor. Pri zvolení modelu pravidiel je používateľ upozornený, že neexistujú operátory, ktoré podporujú číselný *label*.

Obr. 14.7 Príklad vyhľadania operátora predikcie

Pre vyhľadanie operátorov zhlukovania si používateľ nevyberá skupiny. Priamo po spustení vyhľadávania sa zobrazia príslušné operátory. Príklad vyhľadania operátora zhlukovania je na obr. 14.8. Zvolené parametre sú: kombinácia formátu vstupných dát, chýbajúce dáta v záznamoch, výstupný model *Cluster*. Výsledný operátor je *Random Clustering*. Pri zvolení možnosti, že chýbajú dáta v záznamoch je dovolené používateľovi vybrať len *Cluster* model. To je z toho dôvodu, že ostatné operátory nedokážu spracovať chýbajúce dáta bez dodatočného predspracovania. Potom by vyhľadanie skončilo bez výsledku.

Obr. 14.8 Príklad vyhľadania operátora zhlukovania

Stránka je optimalizovaná pre zobrazenie aj na mobilných zariadeniach. Štýl stránky je overený validačnou službou *W3C – World Wide Web Consortium* a spĺňa ich štandardy. Taktiež je overený zdrojový kód a spĺňa štandardy *HTML5*.

15 ZÁVER

V tejto diplomovej práci sme sa zamerali na predstavenie vybraných metód dolovania dát, ktoré poskytuje softvér RapidMiner. Vypracovali sme vzorové metodické postupy pre jednotlivé metódy dolovania dát. Táto práca obsahuje aj vstupné požiadavky, postup modelovania, vlastnosti a čiastkové operácie metód, ktoré je potrebné dodržať pri tvorbe modelu, aby bolo možné ho realizovať. Metódy sú aplikované na vybrané dáta o pasažieroch Titanicu pre vizualizáciu modelovania, validácie a vyhodnotenie výstupných dát. Pri modelovaní rôznych metód nastávajú isté okolnosti, ktoré zabraňujú vytvoreniu modelu a to sú napríklad: typ vstupných dát, typ predikovaného atribútu, chýbajúce dáta a požadovaný výstupný formát. Aby mohol model správne fungovať, je potrebné vstupné dáta odfiltrovať, predspracovať a transformovať do požadovaného tvaru, ktorý konkrétny operátor vyžaduje pre svoju činnosť. Ďalej sme v tejto práci okrem operátorov predikcie predstavili aj operátory na zhlukovanie a možnosti rozšírení v softvéri RapidMiner Studio. Pre uľahčenie výberu operátora používateľovi bola vytvorená stránka, na ktorej si môže pomocou vyhľadávania v databáze operátorov vybrať správny operátor pre požadované vstupné dáta. Po vyhľadaní požadovaného operátora si môže používateľ stiahnuť proces, v ktorom je tento operátor použitý. Vypracovanie tejto diplomovej práce súvisí s riešením projektu KEGA 014ŽU-4/2018 Rozšírenie obsahu študijného odboru o aktuálne požiadavky praxe v oblasti metód umelej inteligencie a IT.

16 ZOZNAM POUŽITEJ LITERATÚRY

[1] **Kotu, Vijay a Deshpande, Bala.** *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner.* ISBN: 978-0-12-801460-8.

[2] **Hofmann, Markus a Klinkenberg, Ralf.** *RapidMiner: Data Mining Use Cases and Business Analytics Applications.* ISBN: 978-1-4822-0549-7.

[3] **Spangler, Scott.** *Accelerating discovery: Mining unstructured information for hypothesis generation.* ISBN: 978-1-4822-3914-0.

[4] **Sammut, Claude a Webb, Geoffrey I.** *Encyclopedia of Machine Learning and Data Mining: Second Edition.* ISBN: 978-1-4899-7685-7.

[5] **Olson, David L. a Wu, Desheng.** *Predictive Data Mining Models.* ISBN: 978-981-10-2542-6.

[6] **Chisholm, Andrew M.** *Exploring Data with RapidMiner.* ISBN: 978-1-78216-933-8.

[7] **Abe, Shigeo.** *Support Vector Machines for Pattern Classification.* ISBN: 978-1-85233-929-6.

[8] **Fielding, Alan H.** *Cluster and Classification Techniques for the Biosciences.* ISBN: 978-0-511-26119-0.

[9] **McCue, Collen.** *Data Mining and Predictive Analysis: Inteligence gathering and crime analysis.* ISBN: 978-0-7506-7796-7.

[10] **Izenman, Alan Julian.** *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* ISBN: 978-0-387-78188-4.

[11] **Tso, Brandt a Mather, Paul M.** *Classification methods for remotely sensed data: Second edition.* ISBN: 978-1-4200-9072-7.

[12] **Olson, David L.** *Descriptive Data Mining.* ISBN: 978-981-10-3339-1.

[13] **Predictive analytics today.** Top 33 Data mining software. [Online] [Dátum: 3. December 2017.] https://www.predictiveanalyticstoday.com/top-data-mining-software.

ČESTNÉ VYHLÁSENIE

Vyhlasujem, že som zadanú diplomovú prácu vypracoval samostatne, pod odborným vedením vedúceho diplomovej práce, ktorým bol prof. Ing. Aleš Janota, PhD. a používal som len literatúru uvedenú v práci.

Súhlasím so zverejnením práce a jej výsledkov.

5. 6. 2018, Žilina

_____

podpis

ŽILINSKÁ UNIVERZITA V ŽILINE

ELEKTROTECHNICKÁ FAKULTA

DIPLOMOVÁ PRÁCA

BC, ROMAN MICHALÍK

TECHNIKY DOLOVANIA DÁT POMOCOU RAPID MINER

PRÍLOHOVÁ ČASŤ

Žilina, 2018

Zoznam príloh

Príloha A | DVD

DVD obsahuje zdrojový kód stránky na výber operátora, databázu operátorov, všetky procesy s danými operátormi a textovú časť diplomovej práce.