ZZN projekt – 2. časť
Databáze počasí v Austrálii

Natália Ivanisková, xivanin00
Emma Krompaščíková, xkromp00

1 Úvod

V rámci tohto projektu riešime tri dolovacie úlohy pomocou nástroja *RapidMiner* (Altair AI Studio). Cieľom je aplikovať rôzne metódy spracovania a analýzy dát, ktoré by poskytli odpoveď na konkrétne otázky týkajúce sa predikcie počasia v Austrálii. Na základe vstupného datasetu, ktorý obsahuje údaje o počasí z rôznych lokalít a časových období, vykonávame predspracovanie dát, analýzu a modelovanie. Výsledky jednotlivých úloh sú interpretované s ohľadom na pôvodnú úlohu a účel analýzy, pričom sa zameriavame na využitie vhodných dolovacích techník na získanie zmysluplných výstupov.

2 Popis datovej sady

CSV súbor obsahuje dáta o počasí v Austrálii a má 145 460 záznamov. Datová sada je voľne prístupná na internete [1]. Dáta boli zaznamenávané na 49 miestach, avšak pre každé miesto sa časové rozmedzie záznamov líši. Maximálne časové rozpätie záznamov je takmer 10 rokov od 01.11.2007 do 25.06.2017. Každý záznam je popísaný atribútmi typu string alebo float:

- **[string]:** Date, Location, WindGustDir, WindDir9am, WindDir3pm, RainToday a RainTomorrow
- **[float]:** MinTemp, MaxTemp, Rainfall, Evaporation, Sunshine, WindGustSpeed, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Cloud9am, Cloud3pm, Temp9am a Temp3pm

Hodnoty *Evaporation, Sunshine, Cloud9am a Cloud3pm* vo viac ako 38% záznamov chýbajú, a preto tieto atribúty budú vylúčené. V prípade ostatných atribútov chýba menej ako 11% hodnôt a to bude ošetrené podľa potrieb konkrétnej dolovacej úlohy.

Dataset obsahuje záznamy zo 49 rôznych meteostaníc vo všetkých oblastiach Austrálie. Kvôli lepšej generalizácii dát sme rozdelili vzorky do 7 podnebných pásiem na základe geografickej polohy meteostanice a podnebia, v ktorom sa táto stanica nachádza. Týmto

[1] `https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package`

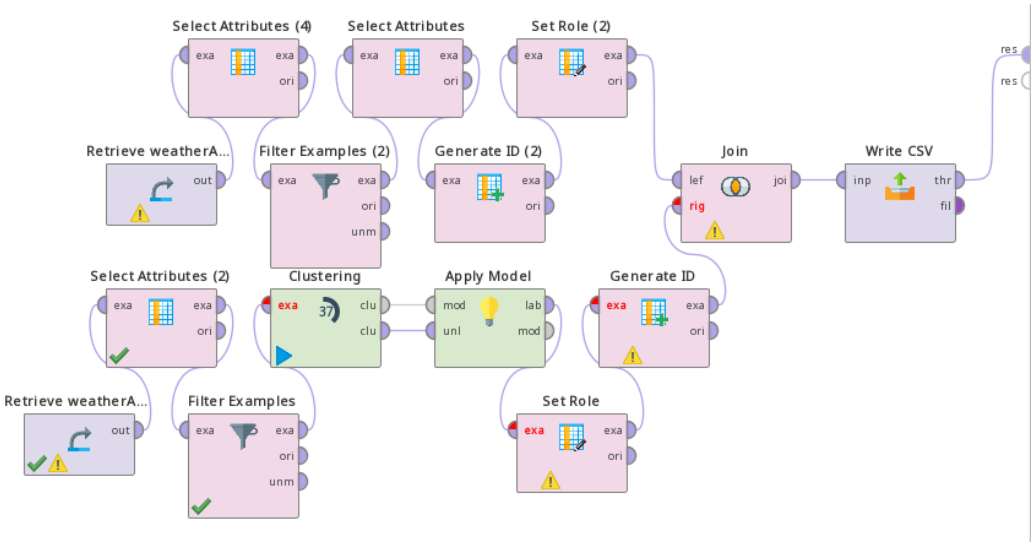
vznikol nový atribút *Climate* a atribút *Location* už nebol potrebný, preto bol z datasetu odstránený. Upravený bol tiež atribút *Date* obsahujúci dátum merania, odkiaľ bola extrahovaná informácia len o mesiaci, v ktorom bol záznam urobený.

3 Dolovacia úloha č. 1 – Clustering

V prvej úlohe sme chceli skúsiť, či sa dáta z datasetu dokážu správne priradiť na klimatickú oblasť, v ktorej sa konkrétne merania počasia vykonávalo. V *RapidMiner* sme data pomocou rôznych operátorov upravili až do nového súboru s dátami (Obr. 1). Konkrétne na pôvodné data boli použité opetátory (dolná vetva):

- **Select Attributes** – Zo všetkých atribútov boli vybrané údaje o vlhkosti, maximálnej teplote, minimálnej teplote, tlaku vzduchu, zrážkach, rýchlosti nárazového vetra a rýchlosti vetra. Atribút *Climate* popisujúci podnebie nebol vybraný.
- **Filter Examples** – V tomto kroku boli odfiltrované všetky riadky, ktoré mali aspoň jednu chýbajúcu hodnotu. Kvôli možnému skresleniu výsledkov nebolo vhodné chýbajúce hodnoty dopĺňať.
- **Clustering** – K-means clustering je metóda neprediktívneho učenia, ktorá rozdeľuje dáta do predom definovaného počtu klastrov, ktoré sú čo najviac homogénne vo vnútri a odlišné medzi sebou. Tento konkrétny operátor má nastavených niekoľko parametrov:
  - **k=7** – Počet klastrov bol nastavený na 7, keďže dataset obsahuje 7 klimatických oblastí, pričom cieľom bolo priradiť každú oblasť do jedného z klastrov. Konkrétne to sú oblasti *Alpine, Desert, Oceanic, Semiarid, Subtropical, Temperate* a *Tropical*. Znamená to, že k-means sa pokúsi rozdeliť dáta na 7 skupín.
  - **add_cluster_attribute=true** – Pridá nový atribút označujúci, do ktorého klastru patrí každý záznam. Tým sa vytvorí nový stĺpec v tabuľke pomenovaný *Cluster*.
  - **max_runs=10** – Urobí maximálne 10 pokusov na optimalizáciu výsledku.

  Výsledok clusteringu je napojený na opetátor **Apply Model**, ktorý aplikuje tento model na nové dáta a priradí ich do príslušných klastrov podľa toho, ako boli určené počas tréningu.
- Pomocou **Set Role** bola novému atribútu *Cluster* zmenená jeho rola na *regular* a pomocou **Generate ID** bolo pre každý riadok pridané jeho identifikačné číslo.



Obr. 1: Ukážka použitých operátorov pre klusterovanie v nástroji *RapidMiner*.

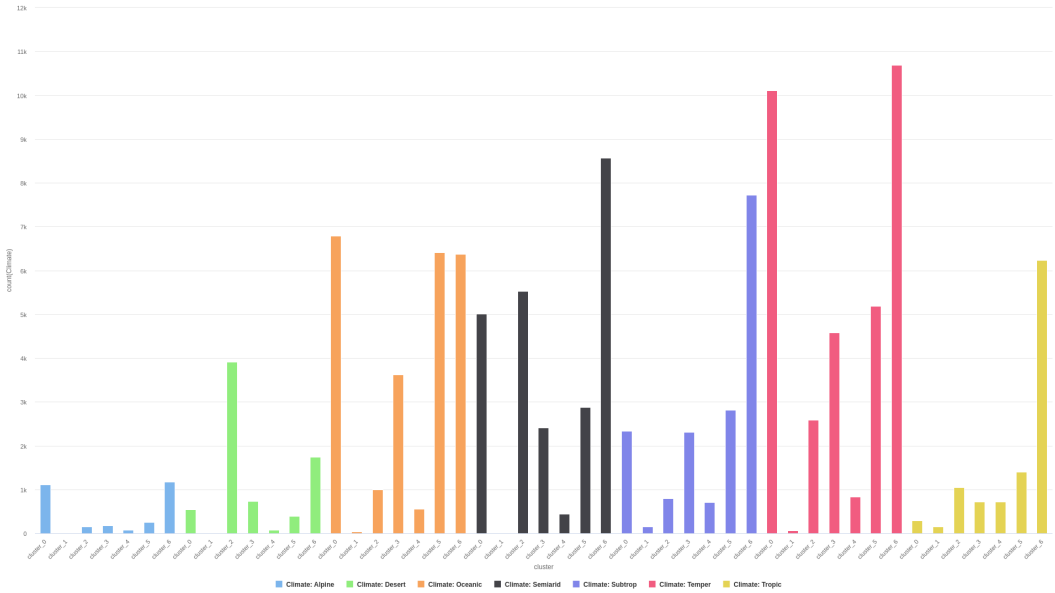
V hornej vetve z Obr. 1 boli pomocou operátorov vybrané rovnaké riadky, ktoré boli použité pri Clusteringu, no ostal z nich ponechaný iba atribút *Climate*. Tomu bolo pridelené ID a následne sa dve vetvy pomocou operátora **Join** spojili do jednej práve pomocou ID – znamená to, že do tabuľky obsahujúcej údaje z clusteringu bol pridaný atribút *Climate*.

V ideálnom prípade by nastala situácia, kde by sa podľa údajov o počasí každej klimatickej oblasti priradil práve jeden zo siedmich clusterov. Len z pár riadkovej ukážky je však vidieť, že sa tak nestalo (Obr. 2).

id	Climate	cluster	MinTemp	MaxTemp	Rainfall	WindGustS...	WindSpeed...	Humidity9am	Pressure9a...
1	Temper	cluster_5	13.400	22.900	0.600	44	20	71	1007.700
2	Temper	cluster_2	7.400	25.100	0	44	4	44	1010.600
3	Temper	cluster_2	12.900	25.700	0	46	19	38	1007.600
4	Temper	cluster_6	9.200	28	0	24	11	45	1017.600
5	Temper	cluster_5	17.500	32.300	1	41	7	82	1010.800

Obr. 2: Ukážka priradených clusterov ku klimatickému podnebiu.

Po odflitrovaní už nepotrebných atribútov bol použitý operátor **Aggregate**, kde bol spočítaný výskyt každého zo sedem clusterov pre každé zo siedmich podnebí. Ako už bolo spomenuté, v najlepšej situácii by jednému podnebí patril práve jeden cluster. Diagram zobrazujúci výskyt clusterov pre každé podnebie je na Obr. 3.



Obr. 3: Graf znázorňuje rozdelenie jednotlivých klastrov v rámci siedmich klimatických oblastí. Na základe grafu je možné pozorovať dominantné klastre pre niektoré klimatické oblasti, avšak medzi určitými oblasťami nie sú dostatočne výrazné rozdiely.

Dôvody, prečo clustering nevyšiel podľa predstáv (konkrétne percentuálne zobrazenie je ukázané v Tabuľke 1), sú napríklad nasledujúce:

- medzi klimatickými oblasťami nie sú dostatočne výrazné rozdiely,
- mohli byť pri čistení vymazané dôležité dáta, ktoré by umožnili lepšie rozlíšiť tieto oblasti,
- vybrané atribúty nemusia dostatočne reprezentovať rozdiely medzi klimatickými oblasťami alebo nezohľadňujú všetky relevantné faktory, ktoré by mohli ovplyvniť clustering.

Climate	**cluster_0**	**cluster_1**	**cluster_2**	**cluster_3**	**cluster_4**	**cluster_5**	**cluster_6**
Alpine	37.58	0.30	4.98	6.07	2.68	8.61	39.78
Desert	7.32	0.04	52.92	9.91	0.97	5.33	23.51
Oceanic	27.41	0.16	4.02	14.60	2.24	25.89	25.70
Semiarid	20.16	0.05	22.26	9.70	1.76	11.58	34.49
Subtrop	13.89	0.87	4.71	13.70	4.18	16.72	45.93
Temper	29.69	0.17	7.60	13.46	2.43	15.24	31.40
Tropic	2.72	1.40	9.89	6.81	6.80	13.26	59.13

Tabuľka 1: Percentuálne zastúpenie clusterov v jednotlivých klimatických oblastiach

Výsledky naznačujú, že pre presnejšie klastrovanie klimatických oblastí by bolo potrebné zvážiť použitie širšej škály atribútov alebo aplikáciu iných metód strojového učenia, ktoré by mohli lepšie zachytiť komplexné vzťahy medzi údajmi.

4 Dolovacia úloha č. 2 – Klasifikácia

Druhou úlohou je klasifikácia do 2 tried, *Yes* alebo *No* odpovedajúvca na otázku *Prší alebo neprší?*. Dataset obsahuje atribút *RainToday*, ktorý na túto otázku priamo odpovedá, preto bol tento atribút použití ako label. Binárny atribút *RainToday* je však vytvorený na základe spojitého atribútu *Rainfall* obsahujúceho hodnoty v rozmedzí 0 až 371. Ak bola hodnota spojitého atribútu *Rainfall* nižšia ako 1, tak atribút *RainToday* nadobúdal hodnotu *No*, v ostatných prípadoch nadobúdal hodnotu *Yes*.

Atribút *RainToday*	**Počet hodnôt**
No	110 319
Yes	31 880

Tabuľka 2: Pomer hodnôt pre atribút *RainToday*

Touto diskretizáciou bol dosiahnutý pomer, ktorý naznačuje, že dataset je silne nevyvážený. Avšak vzhľadom na vysoký počet atribútov a veľké množstvo hodnôt, ktoré môžu nadobúdať, neboli generované ďalšie vzorky, aby nedošlo k vygenerovaniu skreslených a nepravdivých vzoriek. Atribút *Rainfall* bol následne z datasetu v kroku prípravy dát odstránený.

Prvú iteráciu hľadania vhodného modelu pre klasifikáciu prší alebo neprší sme uskutočnili s pôvodným neupraveným datasetom. Využili sme funkciu *Auto Model*, ktorá pre zvolený dataset našla viacero modelov a porovanala ich (Tab. 3).

Model	**Accuracy**
Naive Bayes	80.3%
Generalized Linear model	79.1%
Deep Learning	83.9%
Decision Tree	80.1%
Random Forest	80.4%
Gradient Boosted Tree	83.5%
Support Vector Machine	79.8%

Tabuľka 3: Accuracy Auto Modelov pre neupravený dataset

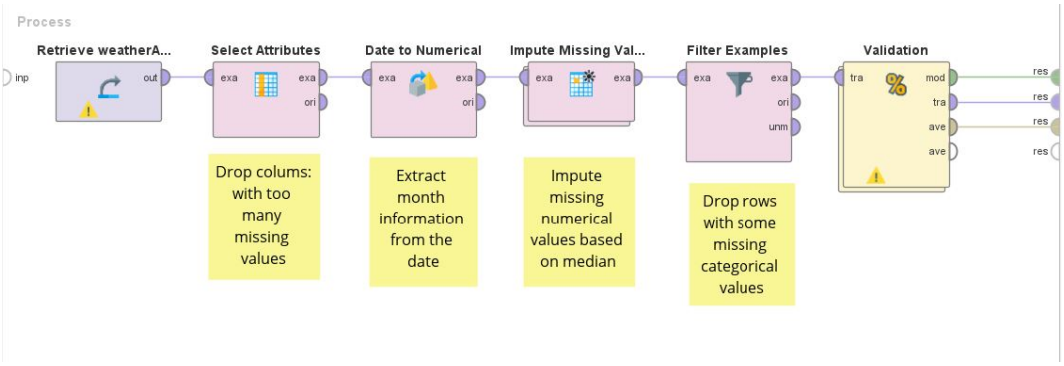
Na základe výsledkov z *Auto Model* sme následne skúšali nájsť tieto modely aj pre upravené varianty pôvodného datasetu.

4.1 Príprava dát

Aj po odstránení niektorých atribútov s veľkým množstvom chýbajúcich hodnôt sa v datasete stále nachádzali vzorky, kde hodnota pre niektorý z atribútov chýbala. Chýbajúce hodnoty boli ošetrené 2 spôsobmi, z čoho vznikli 2 rôzne datasety, s ktorými boli neskôr prevádzané experimenty.

4.2 Doplnenie chýbajúcich hodnôt

Prvý spôsob ošetrenia chýbajúcich hodnôt bolo ich doplnenie vhodne zvolenou hodnotou. V prípade numerických atribútov šlo o nahradenie mediánom daného atribútu. Medián bol zvolený z dôvodu väčšej odolnosti voči odľahlým hodnotám v datasete, ktoré sa tam nachádzajú. Nahradenie chýbajúcich hodnôt v kategorických atribútoch sme chceli ošetriť metódou KNN, avšak z dôvodu obrovského množsta záznamov, sa nám toto nahradenie nepodarilo previesť, keďže výpočet by vo vývojovom prostredí trval príliš dlho. Vzorky s chýbajúcimi hodnotami v kategorických atribútoch boli preto vymazané z datasetu (Obr. 4).

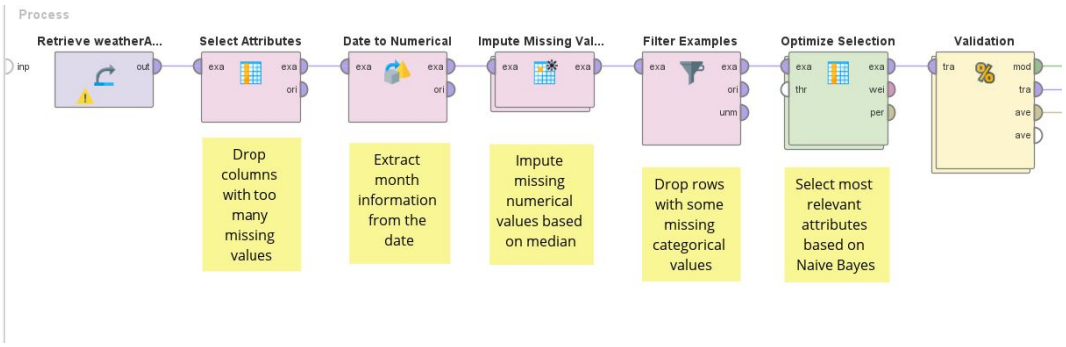


Obr. 4: Proces pre dataset s nahradenými chýbajúcimi hodnotami

Pre vytvorený dataset sme vytvorili modely zvolené podľa výsledkov zo záložky *Auto Model*. Model SVM sme však už použiť nemohli, pretože pri úprave dát vznikol polynomiálny atribút, ktorý tento model nepodporuje. Presnosti sa však výrazne nezlepšili, preto sme pristúpili ku kroku výberu len niektorých atribútov, ktoré sú najvhodnejšie na nájdenie modelu. Výber atribútov sme previedli pomocou operátora *Optimize selection* na základe metódy *Naive Bayes* (Obr. 5).

Pre vzniknuté datasety sme vytvorili zvolené modely (Tab. 5).

Na základe porovananí môžme súdiť, že nahradenie chýbajúcich hodnôt hodnotou mediánu daného atribútu nepomohlo nájsť model s lepšou presnosťou. Takmer všetky modely sa aspoň o trochu zhoršili. Po aplikácia operátora *Optimize selection* sa poväčšine presnosť pre modely zlepšila, avšak ani táto úprava nepomohla dosiahnúť znateľne lepšie výsledky oproti neupravenému datasetu. Na pôvodnom datasete najlepšie fungovala metóda Deep Learning s presnosťou 83.9%, na neoptimizovanom opäť metóda Deep Learning s presnosťou 83.4% a na optimizovanom sa ukázala byť najlepšia metóda Naive Bayes, čo však môže



Obr. 5: Proces pre dataset s nahradenými chýbajúcimi hodnotami a výberom atribútov

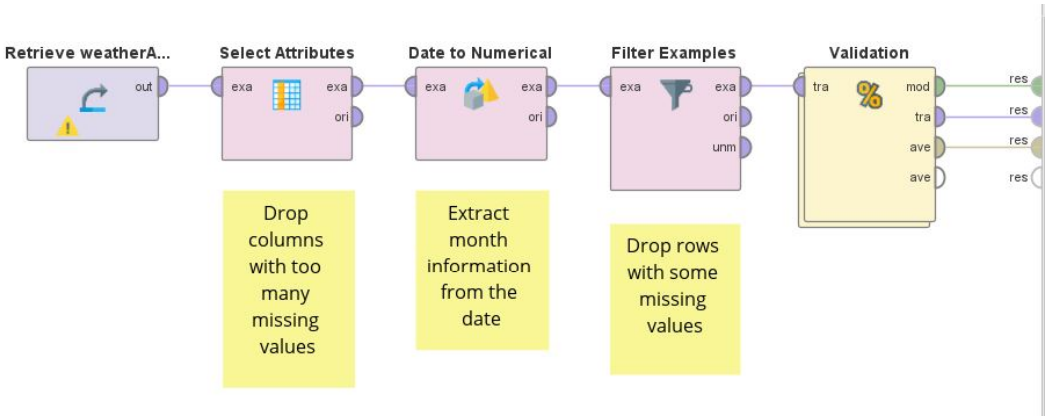
Model	Pôvodný	Neoptimizovaný	Optimizovaný
Naive Bayes	80.3%	77.57%	81.18%
Generalized Linear model	79.1%	80.36%	79.02%
Deep Learning	83.9%	83.40%	81.12%
Decision Tree	80.1%	79.99%	79.62%
Random Forest	80.4%	79.79%	80.63%
Gradient Boosted Tree	83.5%	80.71%	79.16%

Tabuľka 4: Presnosti pre neupravený dataset a datasety s nahradenými chýbajúcimi hodnotami

byť tiež ovplyvnené tým, že metóda Naive Bayes bola tiež použitá pre výber podmnožiny najvhodnejších atribútov.

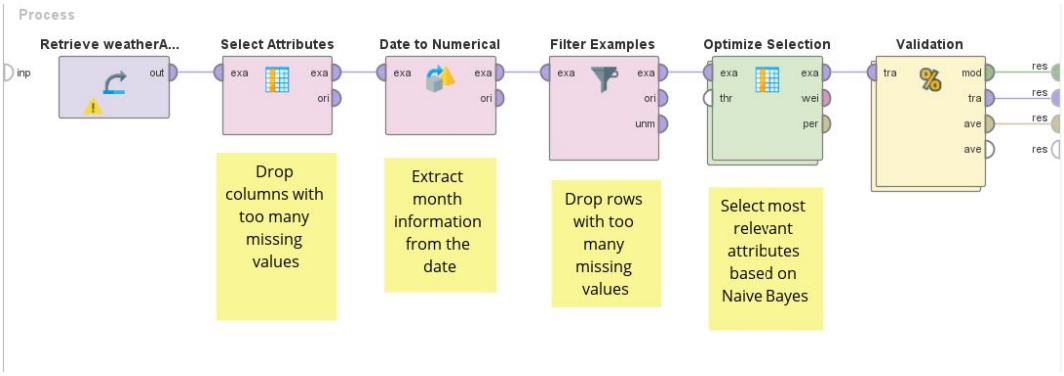
4.3 Vymazanie záznamov s chýbajúcimi hodnotami

Druhým spôsobom ošetrenia chýbajúcich hodnôt bolo odstránenie všetkých záznamov, v ktorých nejaké hodnoty chýbali. Ostatné kroky v postupe sa nezmenili a replikovali sme ich rovnako ako pri prvom postupe ošetrenia chýbajúcich hodnôt (Obr. 6).



Obr. 6: Proces pre dataset s vymazanými záznamami s chýbajúcimi hodnotami

Ďalej sme aplikovali operátor Optimize selection, ktorý mal za úlohu pomocou metódy Naive Bayes nájsť najvhodnejšiu podmnožinu atribútov pre zvolené modely (Obr. 7).



Obr. 7: Proces pre dataset s vymazanými záznamami a výberom atribútov

Pre tieto 2 vzniknuté datasety sme opäť vytvorili zvolené modely a porovanali ich presnosti.

Model	Pôvodný	Neoptimizovaný	Optimizovaný
Naive Bayes	80.3%	77.58%	81.47%
Generalized Linear model	79.1%	81.93%	80.27%
Deep Learning	83.9%	84.16%	81.79%
Decision Tree	80.1%	79.76%	79.85%
Random Forest	80.4%	79.65%	81.79%
Gradient Boosted Tree	83.5%	80.82%	80.70%

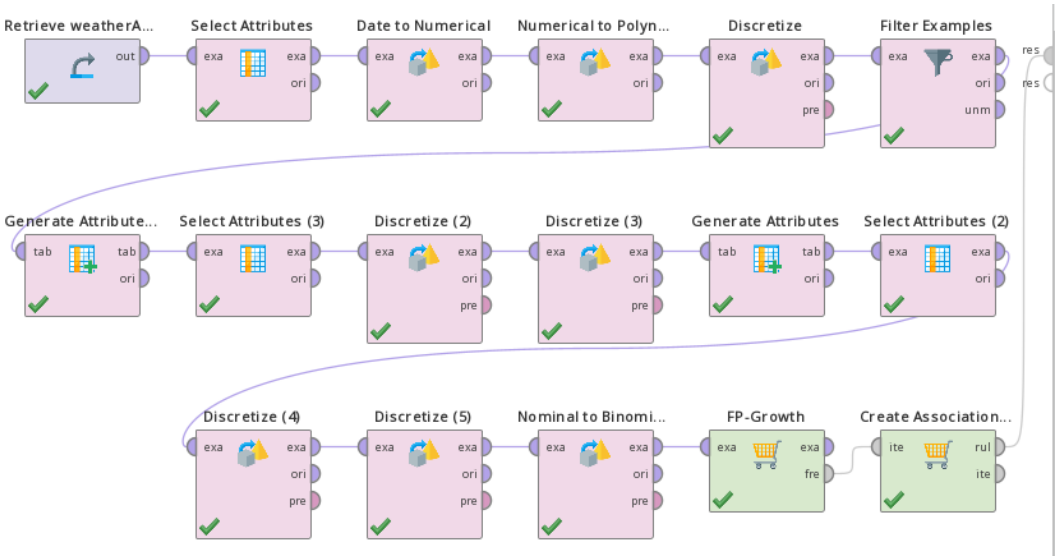
Tabuľka 5: Presnosti pre neupravený dataset a datasety s vymazanými chýbajúcimi hodnotami

Z porovnania môžme vidieť, že pre neoptimizovaný dataset s odstránený záznamami obsahujúcimi chýbajúce hodnoty sa opäť zdá byť najlepšia metóda Deep Learning s presnosťou 84.16% a po aplikácia operátora *Optimize selection* performovali rovanko dobre metódy Deep Learning a Random Forest s presnosťou 81.79%. Celkovo sa presnosť u väčšiny modelov po odstránení záznamov s chýbajúcimi hodnotami a po výbere vhodnej podmnožiny atribútpv zlepšila, avšak toto zlepšenie je len veľmi mierne. Podobnú situáciu sme mohli pozorovať tiež pri postupe s nahradením chýbajúcich hodnôt mediánom, z čoho súdime, že ani jeden z týchto postupov nemal výrazne lepši vplyv oproti druhému postupu na zlepšenie presnosti zvolených modelov.

5 Dolovacia úloha č. 3 – Asociačné pravidlá

Asociačné pravidlá sú jednou z metód dolovania dát , ktoré sa používajú na identifikáciu vzťahov alebo závislostí medzi jednotlivými položkami v dátových súboroch, či na nájdenie skrytých vzorov v údajoch. Pre naše data chceme nájsť rôzne závislosti, ktoré by mohli

popísať počasie v Austrálii. Pre ich nájdenie využijeme viaceré operátory, ktoré sú ukázané na Obr. 8.



Obr. 8: Ukážka použitých operátorov pre vytvorenie asociačných pravidiel v nástroji *RapidMiner*.

Mnoho údajov muselo byť kvôli tvorbe asociačných pravidiel diskretizovaných na kategorické. Pomocou operátoru **Filter Example** boli odfiltrované riadky, ktoré obsahovali jeden a viac chýbajúcich záznamov. Ďalšie konkrétne operátory, ktoré boli použité nad datovou sadou sú nasledujúce:

- **Select Attributes (1)** – v tomto kroku boli vynechané atribúty s veľkým množstvom chýbajúcich dat – konkrétne údaje o oblačnosti, miere odparovania, slnečnosti a informáciu o tom, či dnes prší (tá sa dá odvodiť z atribúty zrážky).
- **Date to numerical** – všetky dátumy sme sa rozhodli kategorizovať na mesiace.
- **Discretize** – pre vlhkosť bolo zvolených podľa hodnoty 5 kategógií – *very_low, low, optimal, high, very_high*.
- **Generate Attributes (2)** – vytvorenie nového atribútu pre priemernú teplotu celého dňa.
- **Discretize (2)** – teplota bola rozdelená do šiestich kategórií – *freezing, low, cool, optimal, high, very_high*.
- **Discretize (3)** – pre silu vetra bolo tak isto zvolených 6 kategórií – *calm, light_air, light_breeze, gentle_breeze, moderate a strong*.
- **Generate Attributes** – vytvorenie nového atribútu pre priemerný denný tlak.

- **Discretize (4)** – Priemerný denný tlak bol kategorizovaný do katerógií *very_low, low, optimal, high, very_high*.
- **Discretize (5)** – Denné zrážky boli kategorizované ako *no_rain, light_rain, moderate_rain, heavy_rain, extreme*.
- Operátor **FP-Growth** potreboval na svoj vstup binomické data, ktoré boli zabezpečené operátorom **Nominal to Binominal**.
- Posledný operátor **Create Association Rules** je napojený na výstup. Ako kritérium bola nastanvená *condifence* na hodnotu *0.7*.

Ukážka niektorých výsledkov z *RapidMiner* je ukázana na Obr. 9.

No.	Premises	Conclusion	Support	Confidence	Lift	Gain
3	WindGustSpeed = calm	NewPressure = optimal	0.208	0.659	1.183	-0.422
4	WindGustSpeed = calm	Rainfall = no_rain	0.217	0.688	0.888	-0.413
5	NewPressure = high	TemperatureAllDay = cool	0.241	0.693	1.352	-0.455
6	Humidity3pm = optimal	Rainfall = no_rain	0.249	0.725	0.936	-0.437

Obr. 9: Ukážka pár výsledkov asociačných pravidiel.

Očakávali sme, že asociačné pravidlá odhalia vzory týkajúce sa extrémneho počasia. Výsledné pravidlá však ukázali predovšetkým známe závislosti, ktoré odrážajú bežné klimatické javy. Tento výsledok mohol byť ovplyvnený rôznorodosťou dát naprieč klimatickými podnebiami a potenciálne aj obmedzeniami diskretizácie atribútov, ktorá mohla viesť k strate jemnejších vzorcov. Pre podrobnejšie analýzy by mohlo byť užitočné v budúcnosti zvážiť detailnejšie kategorizácie alebo pokročilejšie metódy spracovania dát.

6 Záver

Na datasete obsahujúcom dáta o počasí v Austrálii sme vyskúšali 3 dolovacie úlohy, konkrétne clustering, klasifikáciu do 2 tried a asociačné pravidlá. Každá z metód vyžadovala osobitnú úpravu atribútov a použitie rôznych operátorov v prostredí *RapidMiner (Altair AI Studio)*. Dataset síce obsahuje veľké množstvo záznamov, avšak mnoho z nich bolo kvôli chýbajúcim hodnotám nepoužitých. Výsledky dolovacích úloh boli dostatočné, avšak ani úprava datasetu nepomohla k dosiahnutiu ešte lepších výsledkov. Náročnosť práce s dátami prisudzujeme tiež silnej nevyváženosti datasetu, kde prevažuje trieda *RainToday : No*. Výpočty modelov boli tiež zťažené veľkým počtom spojitých a tiež kategorických atribútov, ktoré nadobúdali širokú škálu hodnôt. Po analýze dát a zostavení modelov sa viaceré atribúty ukázali ako nie nevyhnutné pre nájdene modelov a boli z datasetu vypustené. Pripravené dolovacie úlohy splnili svoj účel, dosiahli sme s nimi výsledky, ktoré by už boli akceptovateľné. Na základe veľkosti datasetu sme však očakávali lepšie výsledky alebo zaujímavejšie zistenia. Myslíme si, že okrem nevyváženosti tried zohrala rolu v nájdení

nie až tak prekvapujúcich výsledov aj veľkosť datasetu, konkrétne množstvo atribútov a rôznorodosť hodnôt jednotlivých atribútov pre záznamy, ktoré patria do rovnakej triedy, čo znižovalo množstvo informácií získaných z vytvorených modelov.