# Technická univerzita v Košiciach

# Katedra kybernetiky a umelej inteligencie

# Zadanie z predmetu Objavovanie znalostí

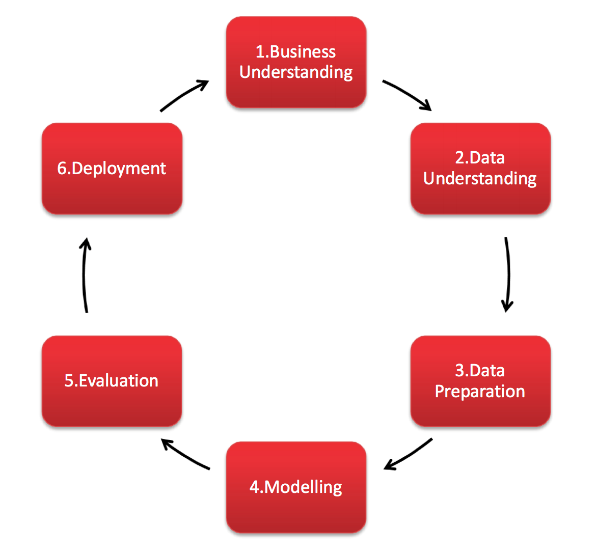
Jakub Geľo

Ľudovít Hiľovský

2019/2020

# Metodológia CRISP-DM

Nasledujúce zadanie budeme vypracovávať podľa rozšírenej metodológie CRISP-DM. CRISP-DM je skratka medzisektorového procesu získavania údajov. Táto Metodika poskytuje štruktúrovaný prístup k plánovaniu objavovania znalostí v dátach. Je to robustná a osvedčená metodika. Model CRISP-DM je zobrazený na Obr. 1 .



Obr. 1 CRISP-DM model

Tento model je idealizovanou sekvenciou udalostí. V praxi je možné mnohé úlohy vykonávať v inom poradí a často bude potrebné ustúpiť od predchádzajúcich úloh a opakovať určité akcie.

# Pochopenie cieľa

Táto časť CRISP-DM modelu sa zameriava na pochopenie obchodných alebo iných cieľov a požiadaviek z manažérskeho hľadiska a následne ich pretransformovať na definíciu úlohy pre získavanie znalostí z databáz. V tejto fáze sa vykonáva aj inventúra zdrojov, hodnotia sa možné rizika, náklady a prínos použitia metód KDD.

## KDD a Biznis cieľ

KDD cieľ nášho zadania je vekové rozdelenie zamestnancov do 2 kategórií. Znalosti, ktoré získame z modelov, ktoré vznikli použitím viacerých metód následne použijeme na splnenie nášho biznis cieľa.

Biznis cieľ je v našom prípade zamerať sa na konkrétnu vekovú kategóriu ľudí, ktorých budeme školiť a upozorňovať na možné nebezpečie na cestách. Hlavným cieľom je teda prevencia akejkoľvek nepríjemnosti či nehody, ktorá môže vzniknúť v každodennej cestnej premávke.

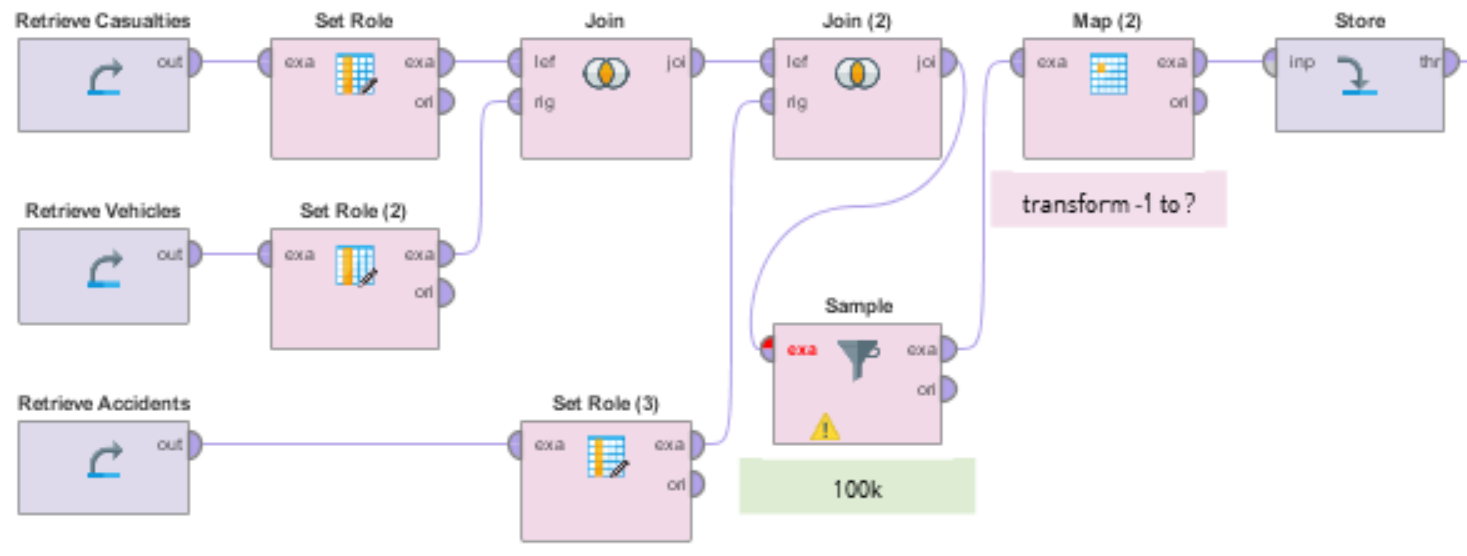
## Zdroj dát

Dáta, ktoré sme použili na získanie znalostí sú verejne dostupné a poskytli ich policajné sily Spojeného kráľovstva, ktoré zhromažďujú údaje o každej zrážke s vozidlom vo Veľkej Británii na formulári s názvom Stats19. Súbory na tomto formulári poskytujú podrobné údaje o bezpečnosti cestnej premávky o okolnostiach nehôd na cestách vo VB od roku 1979, o typoch zúčastnených vozidiel a o následných nehodách. Štatistika sa vzťahuje iba na nehody vzniknuté na verejných cestách. V tejto sade sú 3 .csv súbory obsahujúce údaje o

1. Obetiach
2. Zúčastnených vozidlách
3. Samotných nehodách

# Načítanie dát

Dáta sme načítavali z troch súborov .csv. Samotný proces načítavania dát je znázornený na Obr. 2.



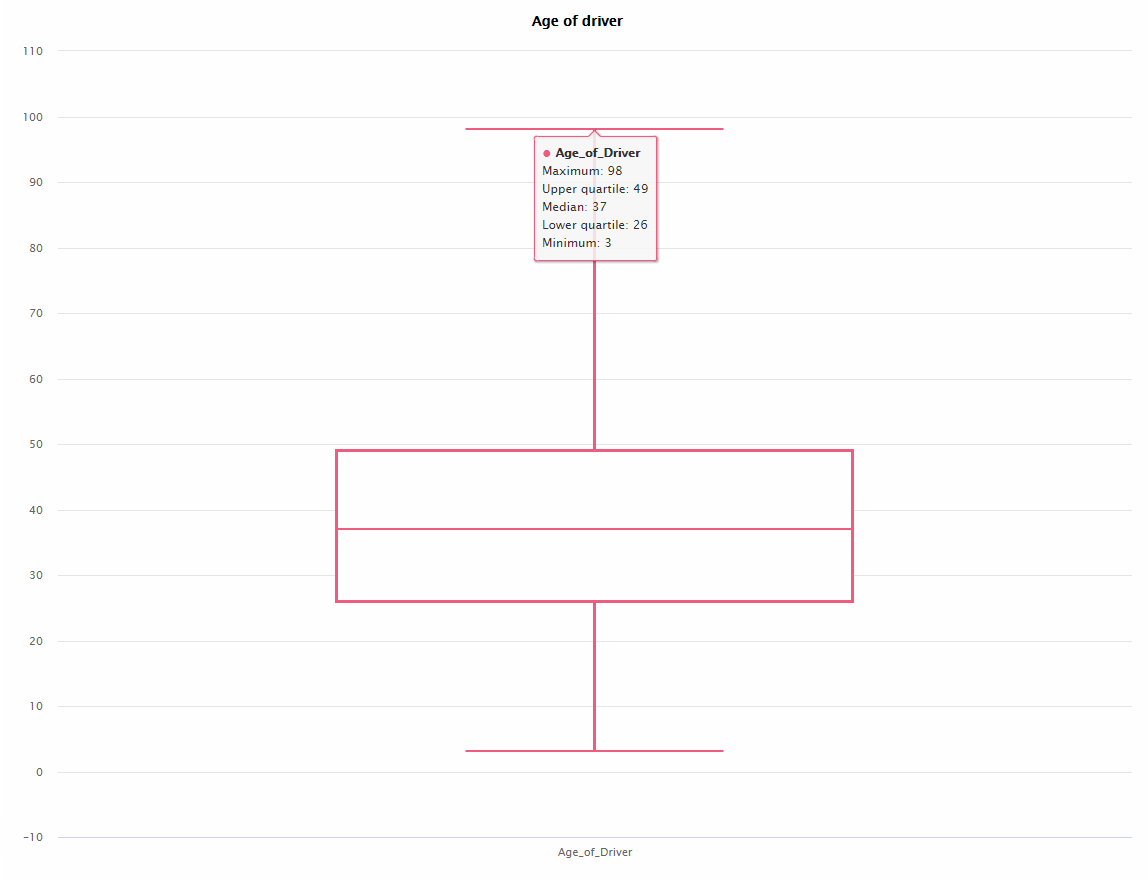
Obr. 2 Načítanie vzorky - 100 000 dát

Pri procese načítavania dát z týchto súborov sme z dát vybrali vzorku o veľkosti 100 000. Vstup dát sme realizovali pomocou funkčného bloku *Retrieve.* Následne sme tieto súbory spojili na základe jedinečného *id,* ktoré sme nastavili v bloku *Set Role*. Dáta sme spojili pomocou funkcie *Join* a vymenili hodnoty pri chýbajúcich atribútoch z *-1* na *?* pomocou funkčného bloku *Map.*

## Exploračná analýza a štatistické charakteristiky

Metódy exploračnej analýzy slúžia na objavenie štruktúr, vytvorenie hypotéz, rozoznanie osobitostí a znázornenie fenoménov. Východiskovým bodom každej analýzy dát sú samotné dáta. Dáta nemusia spĺňať určité podmienky, ako sa žiada v inferenčnej analýze (napr. že dáta museli byť získané náhodným výberom). Ide hlavne o to, rozličnými spôsobmi znázorniť tieto dáta, rozoznať pravidelnosti a nepravidelnosti, štruktúry, vzory a osobitosti. V exploračnom procese hľadáme v dátach zaujímavé konfigurácie a vzťahy.

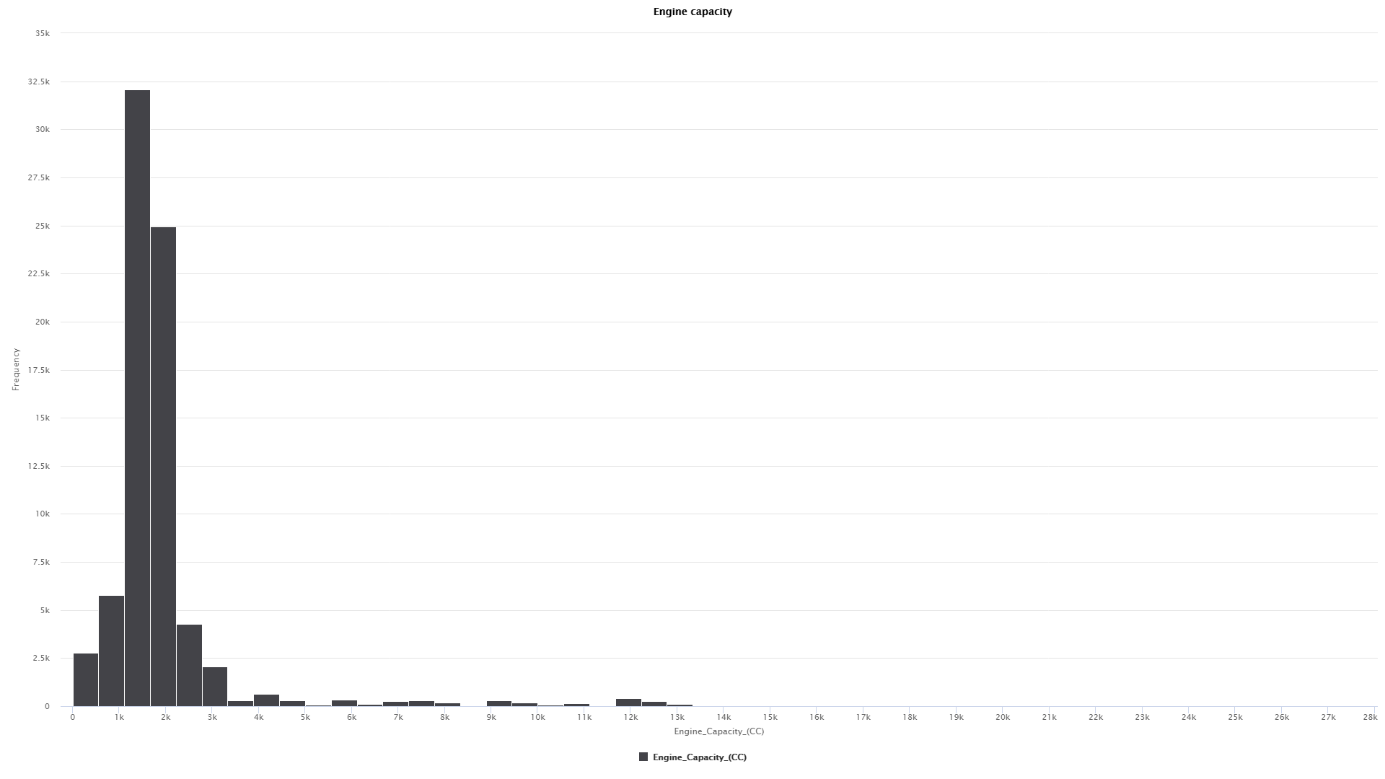
V tejto časti si načítané dáta vizualizujeme pomocou grafov, kde sa budeme snažiť pochopiť dané údaje. Keďže je našim cieľom klasifikovať vek vodiča, bolo by dobré zistiť pomocou grafickej analýzy extrémne hodnoty tohto údaju. Zisťovali sme to pomocou boxplotu, ktorý je znázornený na Obr. 4.



Obr. 3 Boxplot atribútu Age\_of\_Driver

Z grafu je zrejmé, že najstarší vodič má 98 rokov, najmladší 3 roky a najväčšia vzorka vodičov má vekové rozpätie približne 27-50 rokov. Je jasné, že takéto dáta budeme musieť upraviť, keďže extrémne hodnoty môžeme považovať za anomálie a sú v našom prípade zavádzajúce.

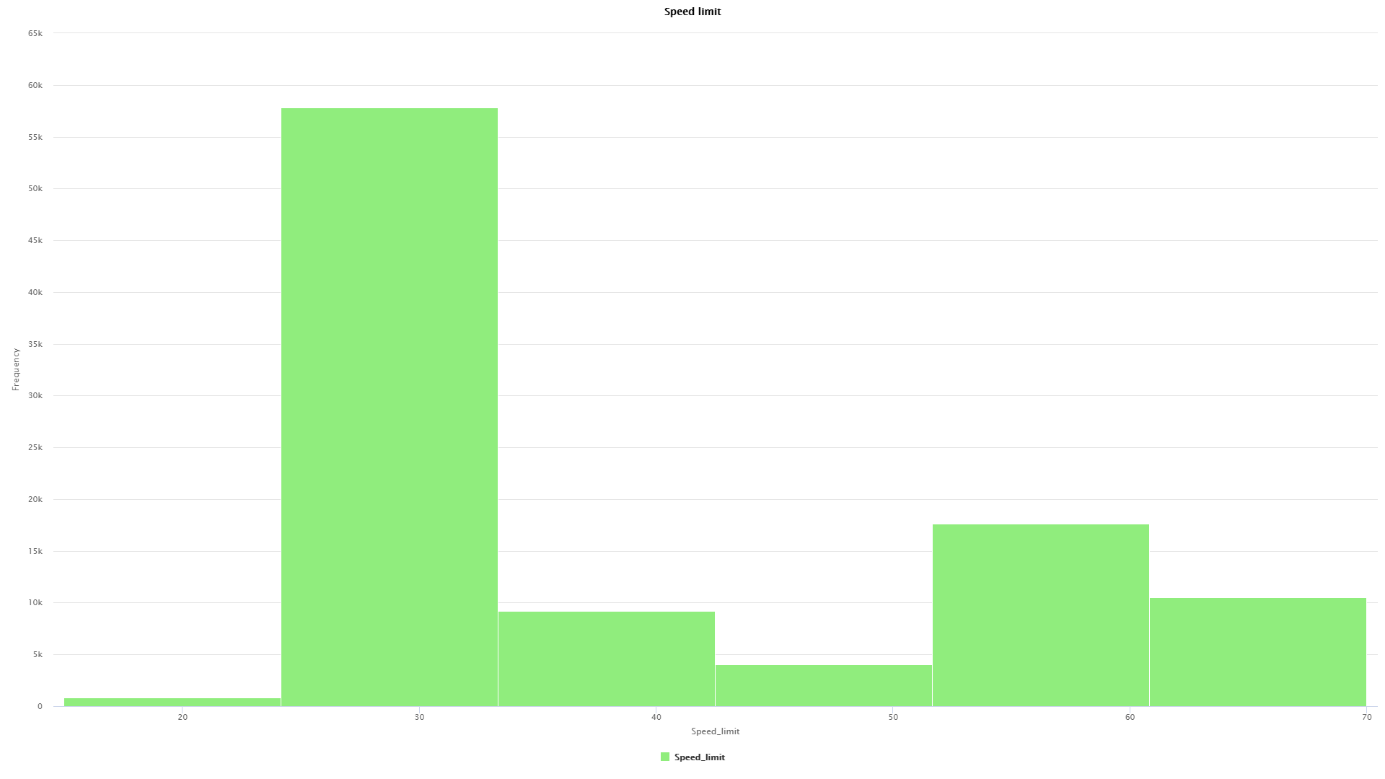
Nasledujúci histogram, ktorý je znázornený na Obr. 4 nám graficky znázorňuje objem valcov motora áut, ktoré spôsobili nehodu.



Obr. 4 Histogram atribútu Engine\_Capacity\_(CC)

Z tohto grafu je zrejmé, že najpočetnejšia vzorka áut mala objem na škále medzi 1.0l – 2l. Tento objem je u bežných, respektíve priemerných áut, z čoho vyplýva, že za nehodu nie je nutne zodpovedné auto s predimenzovaným výkonom motora.

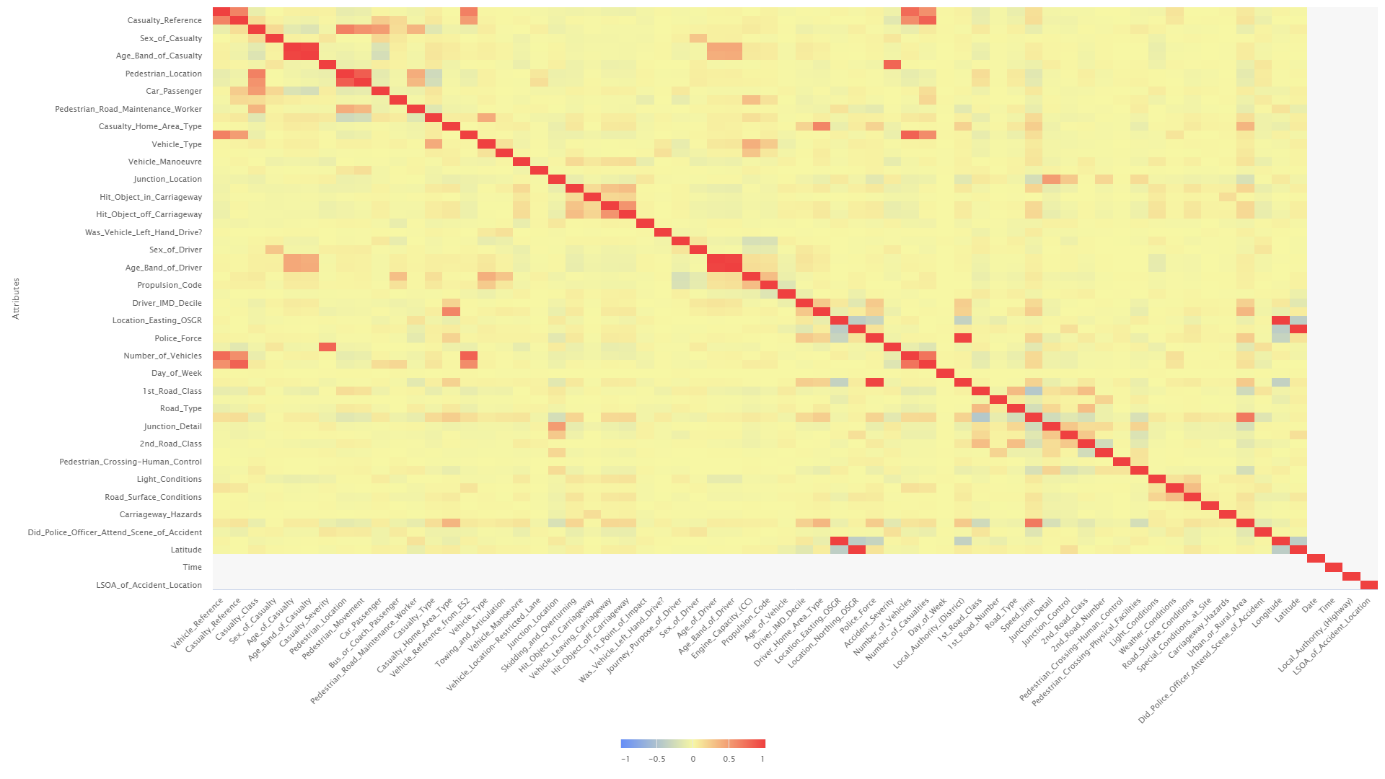
Ďalší graf bude taktiež histogram a bude nám znázorňovať grafickú analýzu atribútu *rýchlostné obmedzenie.* Graf bude znázornený na Obr. 5.



Obr. 5 Histogram atribútu Speed\_limit

Z daného grafu je zrejmé, že najčastejšie došlo k nehodám v oblastiach s rýchlostným obmedzením 30-40 MPH, čo zodpovedá mestským oblastiam.

Korelácia vzájomných atribútov je znázornená korelačnou maticou - Obr. 7.

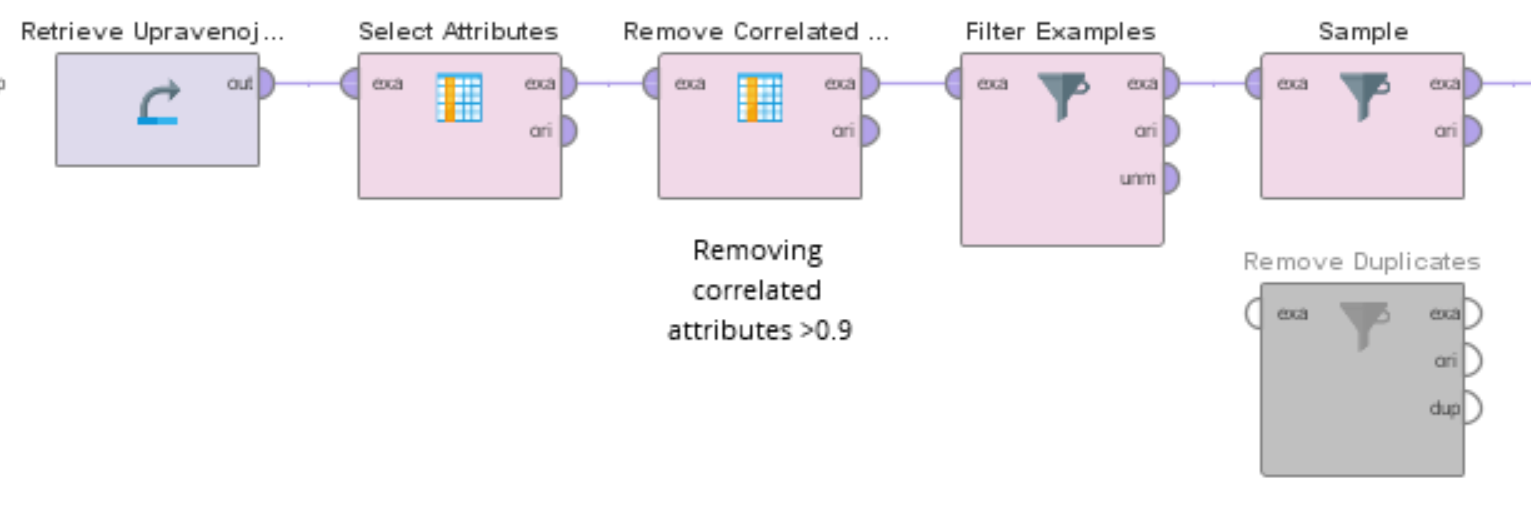


Obr. 6 Korelačná matica

Načítané dáta potrebujeme následne upraviť. Tento proces je veľmi dôležitý, pretože bez neho by sa naše klasifikačné a predikčné úlohy riešili veľmi ťažko, respektíve by naše výsledné modely boli nepresné.

# Príprava dát

Súčasťou prípravy dát bolo dstránenie údajov, ktoré medzi sebou korelovali na úrovni od 0.9. Proces je znázornený na Obr. 7.



Obr. 7 Odstránenie dát s vysokou koreláciou

Pri procese prípravy dát sme využili aj funkčný blok *Remove Duplicates,* ktorý však neovplyvnil výslednú kvalitu dát ani v malej miere. Pre menšiu časovú náročnosť spracovania projektu sme sa rozhodli funkčný blok nezakomponovať do konečného procesu.

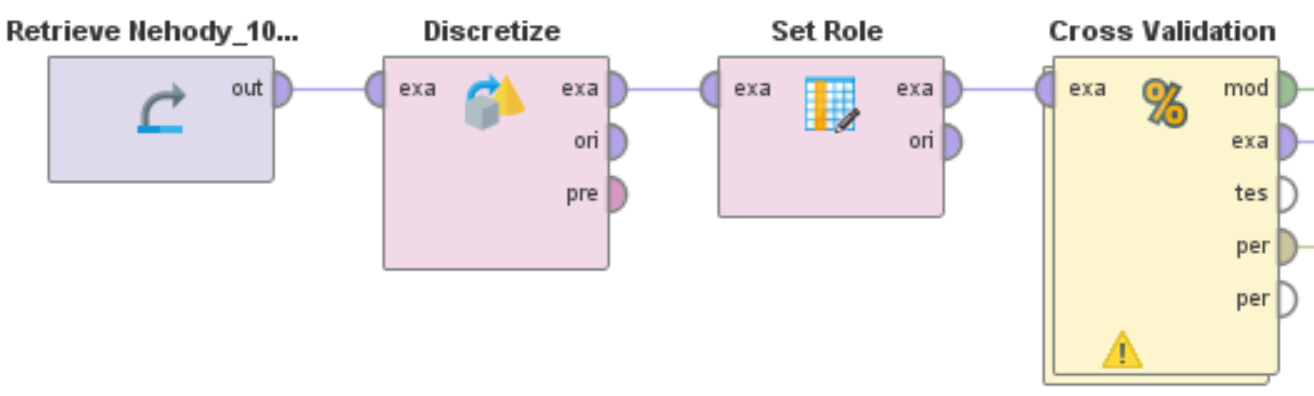
Takto pripravené dáta sme používali na nasledujúce úlohy.

# Klasifikačné úlohy

Cieľom klasifikačnej úlohy je predikovať kategorické označenia tried (predikovaný atribút je nominálny).

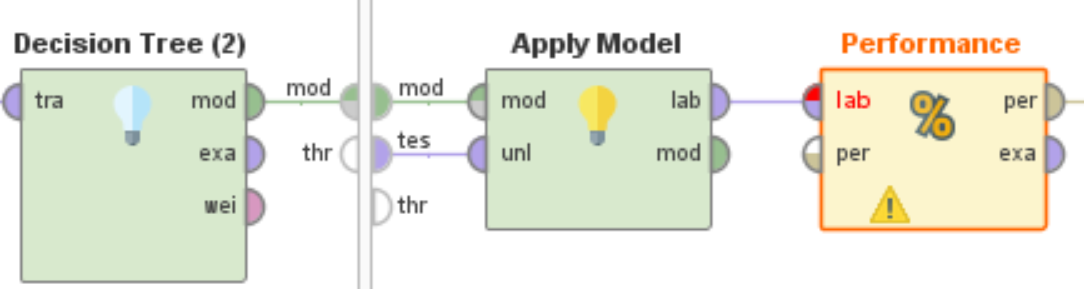
## Rozhodovací strom

Pre úlohu klasifikácie sme použili rozhodovací strom, kde sme klasifikovali vek vodiča podľa nami vybranej podmienky. Vo funkčnom bloku *Discretize* sme diskretizovali numerický atribút na binominálny – starý a mladý vodič. Hranica medzi týmito dvoma skupinami je 45 rokov. Pre spresnenie výsledného modelu sme použili viacnásobnú krížovú validáciu. Proces tvorby rozhodovacieho stromu je znázornený na Obr. 8.



Obr. 8 Rozhodovací strom

Rozdelenie dát na trénovaciu a testovaciu množinu sme pri tejto metóde realizovali pomocou už spomínanej krížovej validácie(Obr. 9).



Obr. 9 Krížová validácia - rozhodovací strom

Parametrizácia jednotlivých funkčných blokov je zhrnutá v Tab. 1.

Tab. 1 Parametrizácia funkčných blokov – krížová validácia a rozhodovací strom

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Pokus |  |  | | | |
|  | *Number of folds* | *Criterion* | Max. Depth | Pru. conf. | Pre-pru conf. |
| 1 | 10 | Inf. gain | 5 | 0.1 | 0.001 |
| 2 | 5 | Inf. gain | 5 | 0.1 | 0.001 |
| 3 | 3 | Inf. gain | 5 | 0.1 | 0.001 |
| 4 | 3 | Accur. | 5 | 0.1 | 0.001 |
| 5 | 3 | Gain ratio | 5 | 0.1 | 0.001 |
| 6 | 3 | Gini index | 5 | 0.1 | 0.001 |
| 7 | 3 | Gini index | 3 | 0.1 | 0.001 |
| 8 | 5 | Gini index | 3 | 0.1 | 0.001 |

Pri daných parametroch sme získali modely s nasledujúcimi vlastnosťami(Tab. 2)

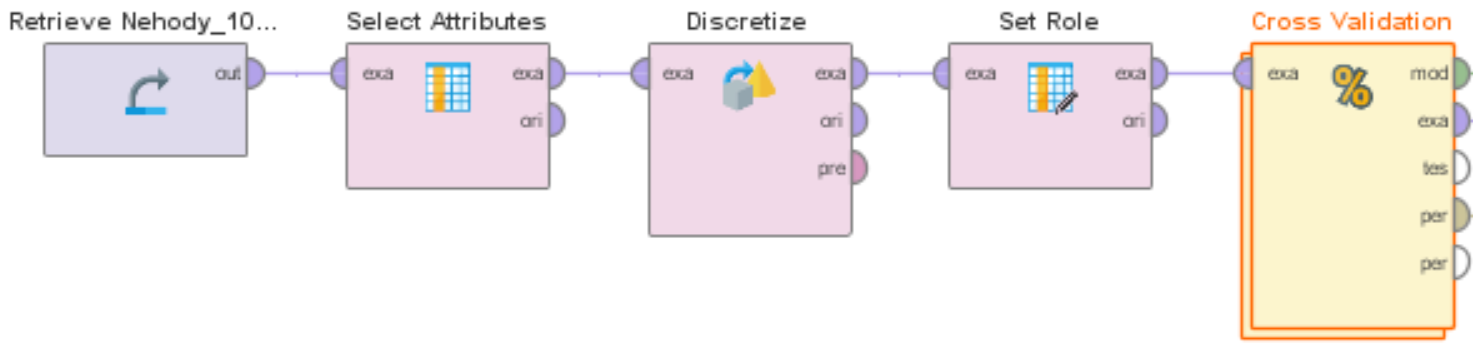
Tab. 2 Výsledné parametre modelov

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pokus | Accuracy | AUC | Sensitivity | Specificity |
| 1 | 74.21%±1.16% | 0.758±0.021 | 57.85%±3.48% | 83.44%±2.01% |
| 2 | 74.73%±1.03% | 0.764±0.015 | 57.10%±3.16% | 84.67%±1.77% |
| 3 | 74.37%±0.43% | 0.763±0.002 | 56.27%±1.54% | 84.58%±0.39% |
| 4 | 74.26%±0.33% | 0.702±0.005 | 54.49%±3.93% | 84.41%±1.62% |
| 5 | 74.47%±0.24% | 0.711±0.011 | 56.85%±0.62% | 84.41%±0.67% |
| 6 | 74.26%±0.33% | 0.749±0.017 | 57.54%±2.80% | 83.69%±1.07% |
| 7 | 74.56%±0.25% | 0.708±0.004 | 55.46%±1.65% | 85.33%±1.10% |
| 8 | 74.49%±0.87% | 0.708±0.013 | 54.88%±2.73% | 85.55%±0.82% |

V Tab. 1 a Tab. 2 sú farebne označené(červenou) 2 najlepšie parametrizácie procesu, pri ktorých sme dostali najkvalitnejšie výsledky.

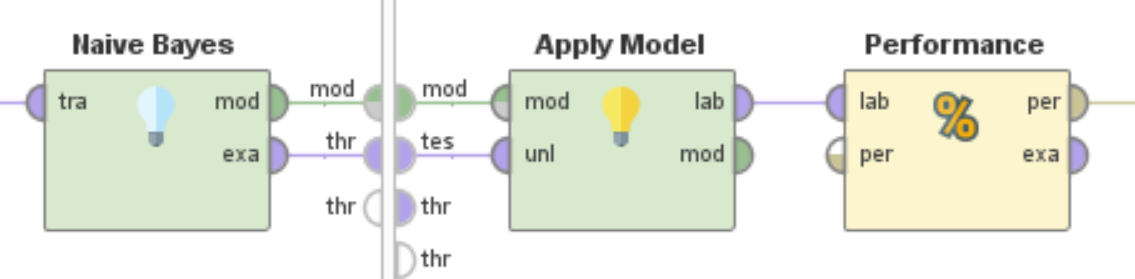
## Naivný bayes

Bayesovské klasifikátory sú štatistické klasifikátory, ktoré predikujú pravdepodobnosti, s ktorými daný príklad patrí do tej – ktorej triedy. Vychádzajú pritom z určenia podmienených pravdepodobností jednotlivých hodnôt atribútov pre rôzne triedy. Naivný Bayesovský klasifikátor vychádza z predpokladu nezávislosti atribútov medzi sebou. To znamená, že efekt, ktorý má hodnota každého atribútu na danú triedu, nie je ovplyvnený hodnotami ostatných atribútov. Kvôli tomuto zjednodušeniu je tento klasifikátor nazývaný ako „naivný“. Túto metódu sme vytvorili s cieľom porovnania klasifikačných metód(v našom prípade Rozhodovací strom). Tak ako pri predošlej metódy, aj pri tejto sme zmenili numerický atribút na binominálny. Proces je znázornený na Obr. 10.



Obr. 10 Naivný Bayes

Pri tejto metóde sme použili viacnásobnú krížovú validáciu. Proces s použitím tejto metódy je znázornený na Obr. 11.



Obr. 11 Naivný Bayes - krížová validácia

Parametrizácia funkčných blokov a výsledné hodnoty modelu sú zhrnutá v Tab. 3 Parametrizácia a model - Naivný Bayes.

Tab. 3 Parametrizácia a model - Naivný Bayes

|  |  |  |  |
| --- | --- | --- | --- |
| Pokus |  |  | |
|  | *Number of folds* | *Accuracy* | *AUC* |
| 1 | 10 | 63.25%±0.77% | 0.726±0.016 |
| 2 | 5 | 63.15%±0.49% | 0.726±0.013 |
| 3 | 3 | 63.18%±0.78% | 0.726±0.005 |
| 4 | 15 | 63.28%±1.32% | 0.726±0.018 |

Výsledný model s najlepšími parametrami je vyznačený červenou farbou.

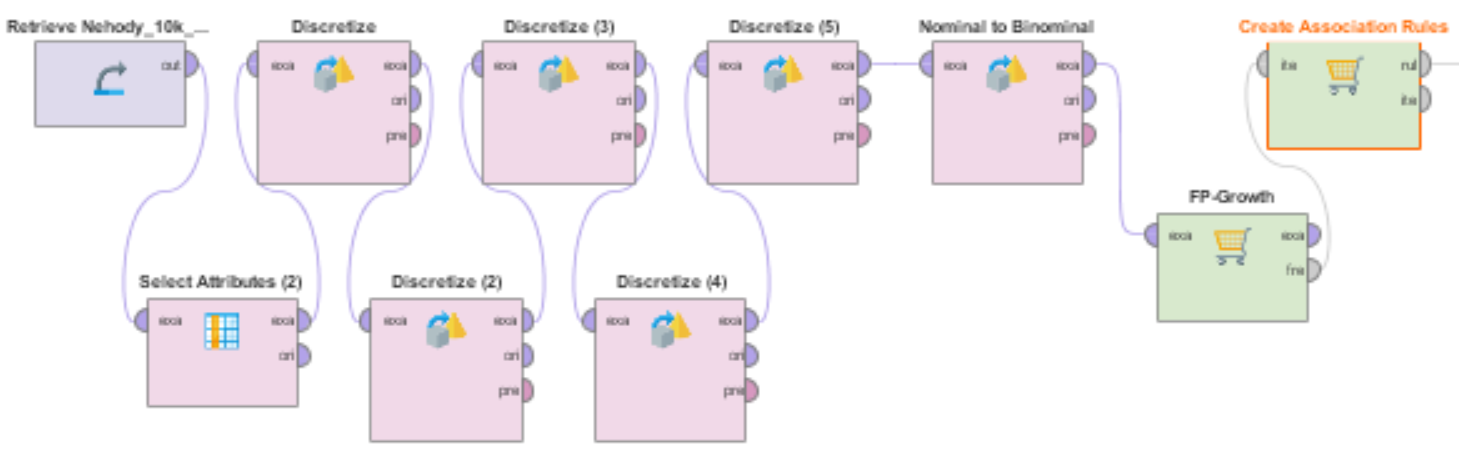
# Popisné úlohy

Popisné dolovanie v dátach sa snaží popísať určitú dátovú množinu stručným a výstižným spôsobom. Ide teda vlastne o generalizáciu (zovšeobecňovanie).

## Asociačné pravidlá

Dolovanie asociačných vzorov, resp. asociačných pravidiel znamená hľadanie frekventovaných a zaujímavých asociácií (korelácií alebo kauzálnych vzťahov) medzi skupinou atribútov v dátach. Pravidlá sú vo forme “Telo ⇒ Hlava [podpora, spoľahlivosť]”. Tento problém je klasicky definovaný v kontexte transakčných dát zo supermarketu, v ktorých sa hľadajú často spolu nakupované skupiny tovarov (frequent itemsets, frequent patterns), avšak bol zovšeobecnený na mnohé iné typy dát, dokonca aj na dáta so závislosťami

Proces tvorby asociačných pravidiel je znázornený na Obr. 12.



Obr. 12 Asociačné pravidlá

Asociačné pravidlá sme tvorili s nasledujúcimi atribútmi:

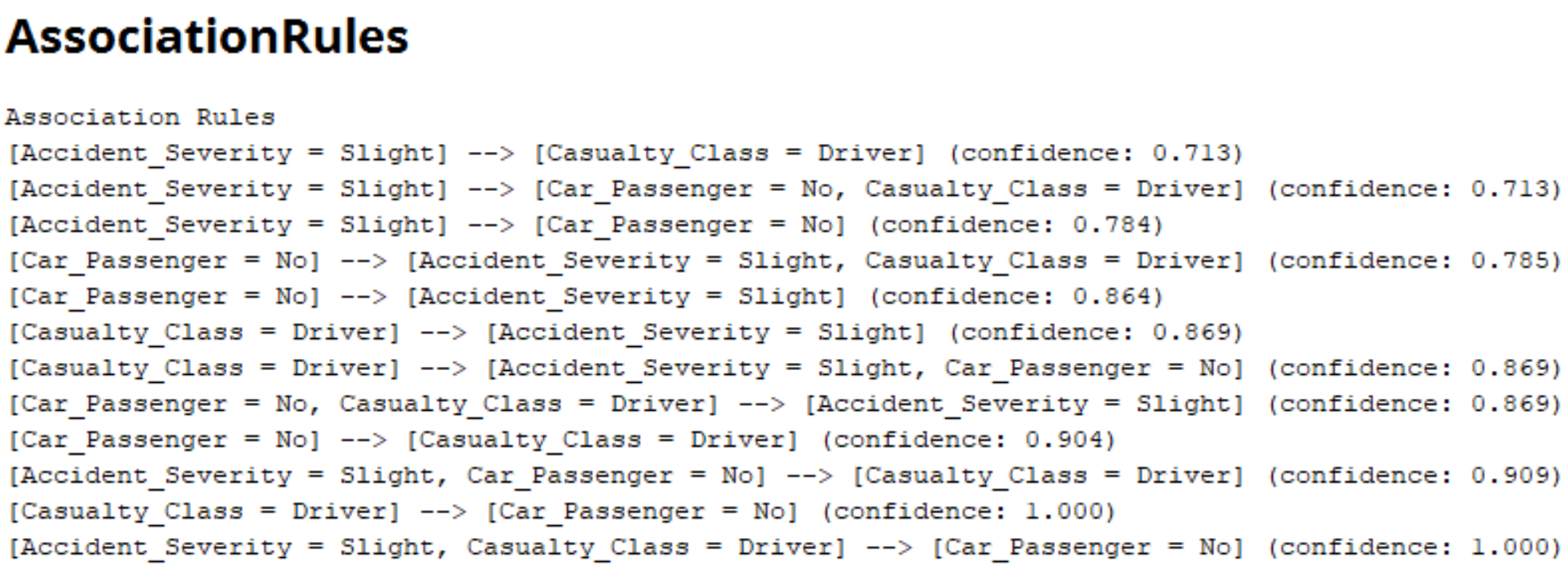
* Accident\_everity
* Age\_of\_Casualty
* Car\_Passenger
* Casualty\_Class
* Sex\_of\_Casualty
* Accident\_Index

Tieto atribúty sme vybrali pomocou funkčného bloku *Select Atributes.* Následne sme každý atribút diskretizovali osobitne pomocou bloku *Discretize.* Ako ďalší krok sme zvolili transformáciu nominálnych atribútov na binominálne. Parametre zvyšných funkčných blokov zhrnieme v Tab. 4.

Tab. 4 Parametrizácia funkčných blokov –časté vzory a asociačné pravidlá

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Pokus |  | | | | | | |  | |
|  | *Input format* | *Min. req.* | *Min. support* | | *Min/Max per itemset* | *Max num. Of itemsets* | | *Criterion* | *Min Confidence* |
| 1 | default | support | 0.5 | 1/0 | | | 1x106 | *Confidence* | 0.7 |

Výsledné asociačné pravidlá sú zobrazené na Obr. 13.



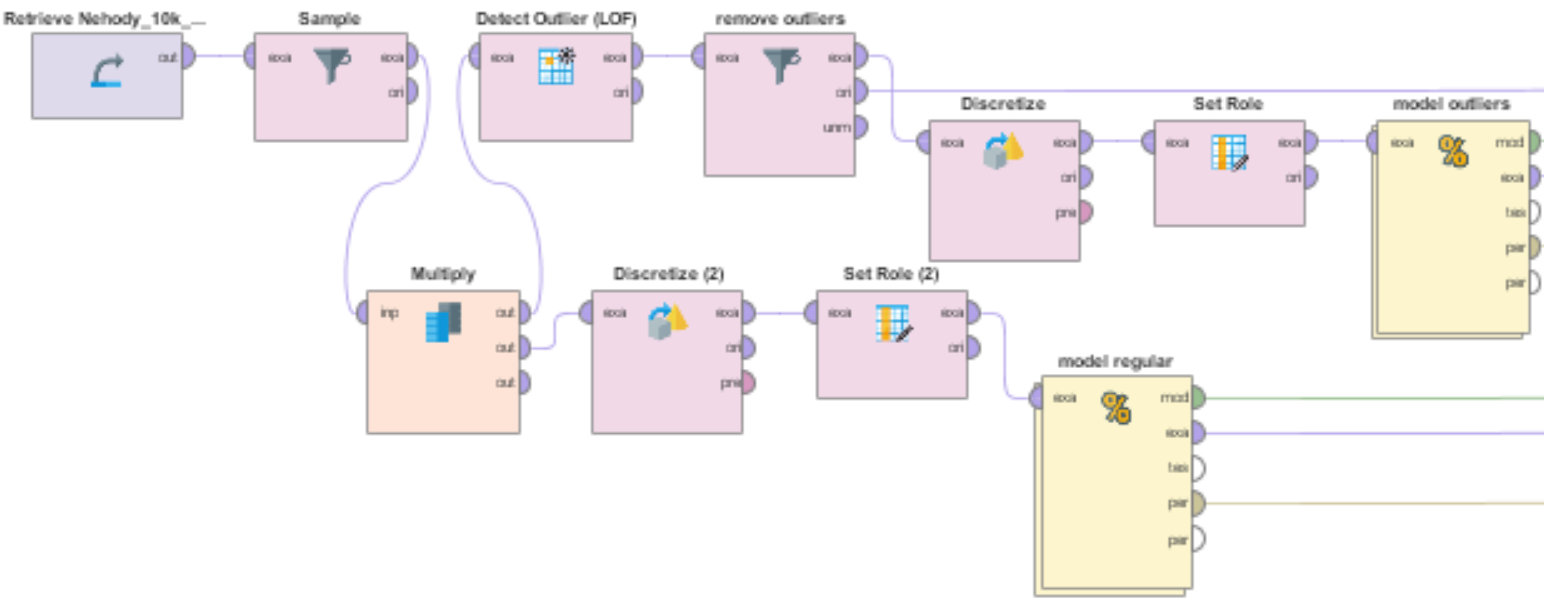
Obr. 13 Výsledný model - asociačné pravidlá

Z asociačných vzťahov nám vyplývajú určité pravidlá, ktoré môžeme využiť v našom biznis cieli. Najdôležitejšia informácia je neprítomnosť spolujazdca v nehodách s malými celkovými škodami na zdraví. To znamená, že treba obzvlášť dbať na zdravie spolujazdcov.

# Detekcia anomálií

Anomália (outlier) je dátový bod ktorý je veľmi odlišný od ostatných bodov. Aplikácie anomálií v dolovaní v dátach sa využívajú pri čistení dát – anomálie predstavujú šum v dátach, preto sa zvyknú v procese predspracovania odstrániť. Podvody s kreditnými kartami sú takisto typickým príkladom ich využitia – netypické vzory v aktivite kreditnej karty môžu poukazovať na jej zneužitie. Anomálie môžeme detegovať ako prieniky do sietí – prevádzka rôznych typov sietí môže byť chápaná ako prúd multidimenzionálnych záznamov. Anomálie sú často definované ako netypické záznamy v tomto prúde, alebo netypické zmeny v základných trendoch.

My sme využili detekciu anomálii pre porovnanie klasifikačnej metódy rozhodovacieho stromu. Proces si znázorníme na Obr. 14.



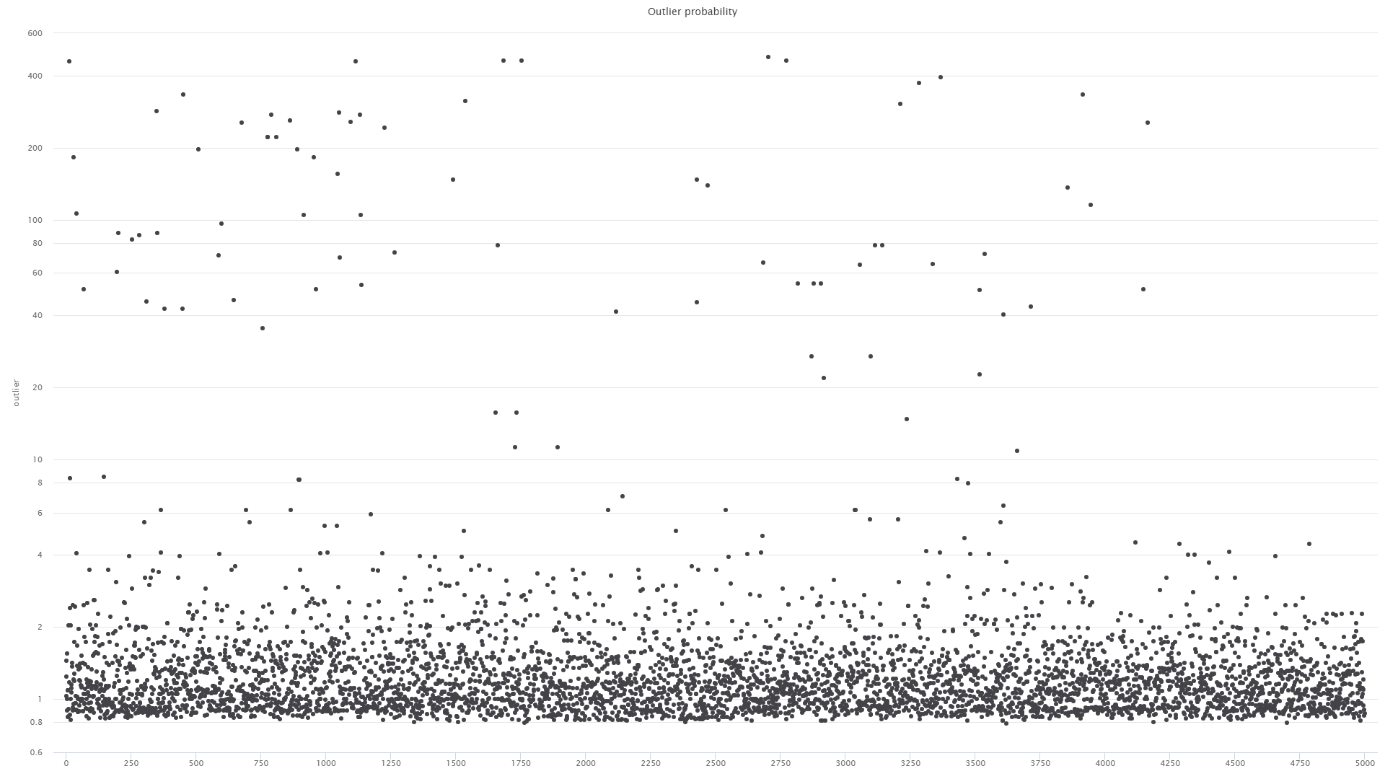
Obr. 14 Detekcia anomálii a ich následne odstránenie

Pre časovú náročnosť sme zúžili rozsah dát na 5000. Dáta sme rozdelili do dvoch smerov pomocou bloku *Multiply.* Smer dátového toku viedol k detekcii a odstráneniu anomálií pomocou funkčných blokov *Detect Outliers(LOF) a remove outliers.* Diskretizovali sme rovnako ako pri predchádzajúcom modely(Rozhodovací strom) *.* Následnú parametrizáciu funkčných blokov si zhrnieme v Tab. 5.

Tab. 5 Parametrizácia funkčných blokov - detekcia anomálií

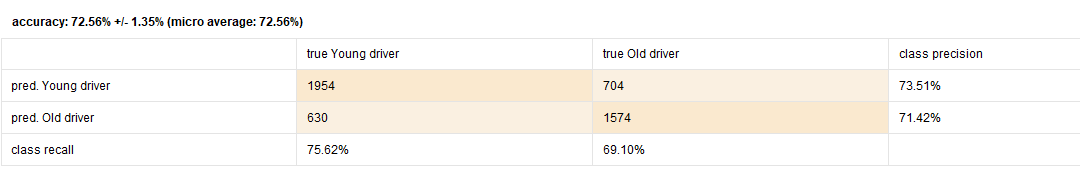
|  |  |  |
| --- | --- | --- |
| Funkčný blok | Parameter | Hodnota |
|  | *Minimal points lower bounds* | 8 |
| *Minimal poionts upper bound* | 15 |
| *Distance function* | *Euclidian distance* |
|  | *Number of folds* | 5 |
| *Sampling type* | *Automatic* |
|  | *Criterion* | *Information\_gain* |
| *Maximal depth* | 5 |
| *Confidence(apply pruning)* | 0.1 |
| Minimal gain(apply prepruning) | 0.001 |
| Minimal leaf size(applyprepruning) | 2 |

Anomálie si znázorníme na rozptylovom grafe – tento typ grafu je najvhdonejší pre ich vizualizáciu. Graf je znázornený na Obr. 15.



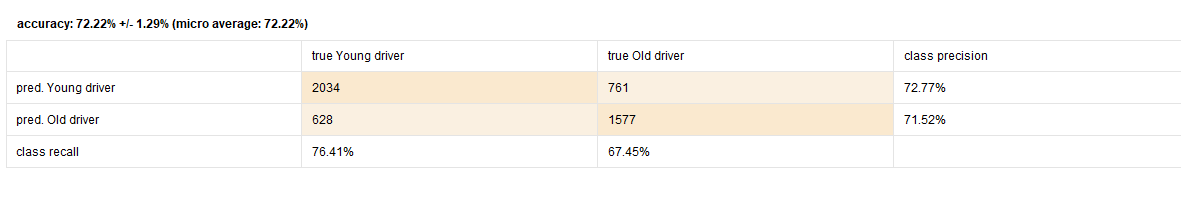
Obr. 15 Pravdepodobnosť anomálií

Porovnanie presností výsledných modelov si znázorníme výslednými kontigenčnými tabuľkami modelov. Začneme znázornením presnosti modelu bez anomálií - Obr. 16.



Obr. 16 Presnosť rozhodovacieho stromu po odstránení anomálií

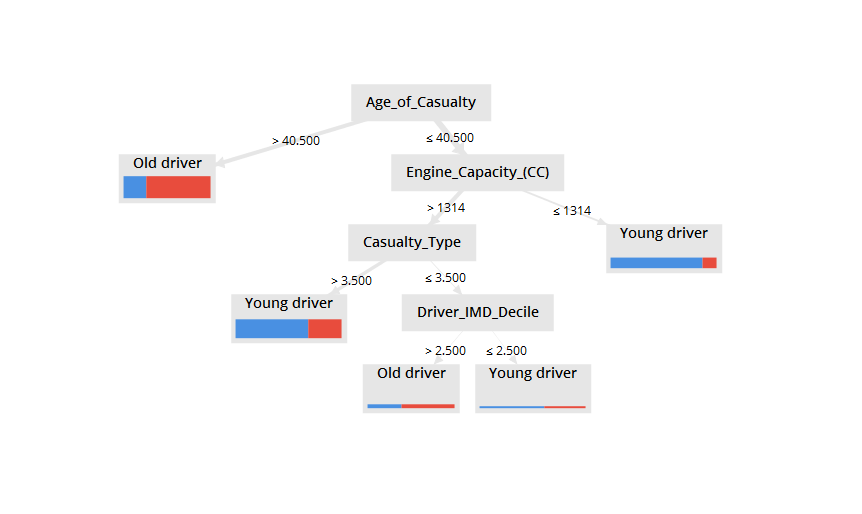
Bez odstránenia anomálií mal tento model nasledujúce výsledky - Obr. 17.



Obr. 17 Presnosť rozhodovacieho stromu po odstránení anomálií

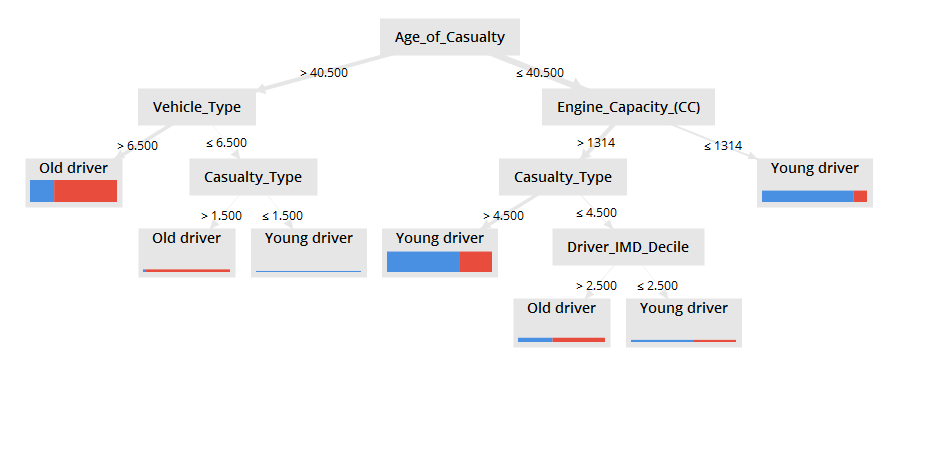
Z výsledkov môžeme usúdiť, že – paradoxne - model s anomálií klasifikoval náš atribút s väčšou presnosťou ako bez nich.

Rozhodovací strom modelu bez anomálií si znázorníme na Obr. 18.



Obr. 18 Rozhodovací strom - bez anomálií

Rozhodovací strom modelu s anomáliami si znázorníme na Obr. 19.



Obr. 19 Rozhodovací strom - s anomáliami

Na modeloch rozhodovacích stromov môžeme zreteľne vidieť, ktoré atribúty dopomohli ku úspešnej klasifikácií modelu najviac.

# Záver

Úspešnosť našich modelov sa pohybovala na úrovni okolo 73%, čo hodnotíme ako pozitívny výsledok, keďže sme vylúčili duplicitný atribút *Age\_of\_Band,* ktorý by skresľoval výsledky našich pokusov. Najťažšie sa nám pracovalo s popisnými metódami, keďže zhlukovanie nám nevykazovalo potrebnú presnosť (aspoň 60%) ani po niekoľkých pokusoch a využití viacerých metód. Klasifikačné metódy pokladáme za úspešné a pracovalo sa nám s nimi nie veľmi obtiažne. Výsledné modely teda môžeme využiť pre náš biznis model.