Zadanie Objavovanie znalostí

Naše dáta zaznamenávajú údaje o nehodách v rokoch 2005 až 2015. Pochádzajú z Ministerstva dopravy Spojeného kráľovstva. Dáta zozbierali Policajné sily Spojeného kráľovstva, lebo oni zhromažďujú údaje o každej zrážke s vozidlom v Spojenom kráľovstve vo formulári s názvom Stats19.

Dataset obsahuje 10 000 záznamov s jedným špeciálnym atribútom a 65 atribútmi, ktoré sme získali spojením troch tabuliek Accidents, Casualties a Vehicles na základe špeciálneho atribútu Accident\_Index. Špeciálny atribút obsahuje 12 znakov (prvých 5 sú čísla, potom dve písmená a posledné sú čísla, napr. 200501BS00001). Atribúty jednotlivých tabuliek sú vypísané v nasledujúcich tabuľkách:

**Tabuľka Accidents**

|  |  |
| --- | --- |
| Police Force | Speed Limit |
| Accident Severity | Junction Detail |
| Number of Vehicles | Junction Control |
| Number of Casualties | 2nd Road Number |
| Date | 2nd Road Class |
| Day of Week | Pedestrian Crossing-Human Control |
| Time | Pedestrian Crossing-Physical Facilities |
| Location Easting OSGR | Light Conditions |
| Location Northing OSGR | Weather Condition |
| Longitude | Road Surface Condition |
| Latitude | Special Condition at Site |
| Local Authority (District) | Carriageway Hazards |
| Local Authority | Urban or Rural Area |
| 1st Road Class | Did Police Officer Attend Scene of Accident |
| 1st Road Number | Lower Super Ouput Area of Accident\_Location |
| Road Type |  |

**Tabuľka Vehicles**

|  |  |
| --- | --- |
| Vehicle Reference | 1st Point of Impact |
| Vehicle Type | Was Vehicle Left Hand Drive |
| Towing and Articulation | Journey Purpose of Driver |
| Vehicle Manoeuvre | Sex of Driver |
| Vehicle Location-Restricted Lane | Age of Driver |
| Junction Location | Age Band of Driver |
| Skidding and Overturning | Engine Capacity |
| Hit Object in Carriageway | Vehicle Propulsion Code |
| Vehicle Leaving Carriageway | Driver IMD Decile |
| Hit Object off Carriageway | Driver Home Area Type |
| Age of Vehicle (manufacture) |  |

**Tabuľka Casualties**

|  |  |
| --- | --- |
| Vehicle Reference | Pedestrian Movement |
| Casualty Reference | Car Passenger |
| Casualty Class | Bus or Coach Passenger |
| Sex of Casualty | Pedestrian Road Maintenance Worker |
| Age of Casualty | Casualty Type |
| Age Band of Casualty | Casualty IMD Decile |
| Casualty Severity | Casualty Home Area Type |
| Pedestrian Location |  |

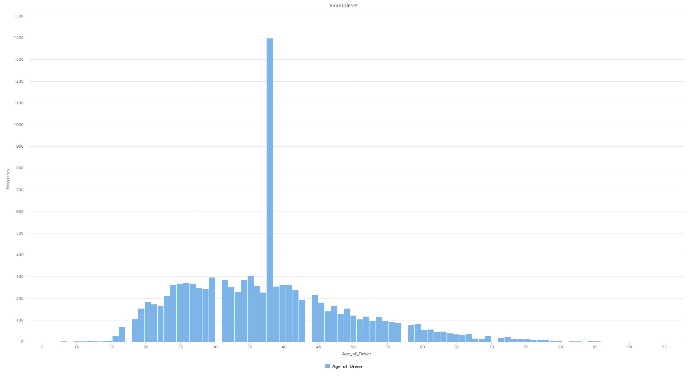
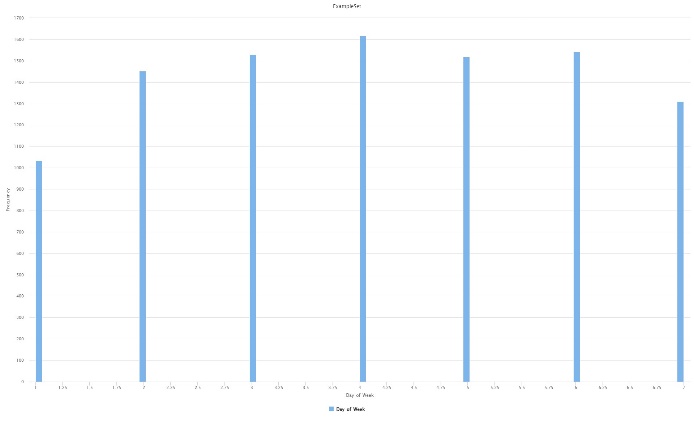
Po spojení tabuliek sme začali skúmať, aké dáta máme k dispozícií, aké hodnoty obsahujú jednotlivé atribúty, t.j. pozreli sme sa na štastické údaje.

**Štatistické údaje**

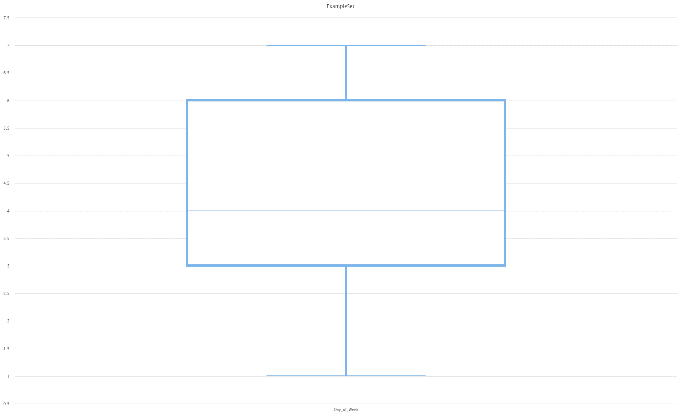
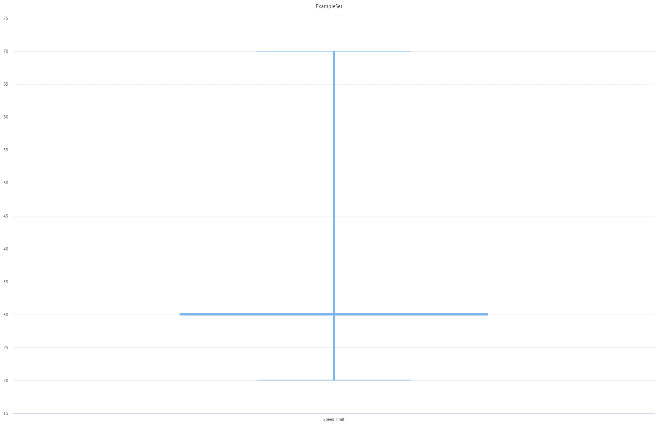
|  |  |  |  |
| --- | --- | --- | --- |
| **Atribút** | **Min** | **Max** | **Priemer** |
| Location Easting | 503 870 | 559 570 | 534 749,289 |
| Location Northing | 170 370 | 195 950 | 183 636,049 |
| Longitude | -0,506 | 0.301 | -0,059 |
| Latitude | 51,417 | 51,645 | 51,535 |
| Police Force | 1 | 48 | 2,457 |
| Accident Severity | 1 | 3 | 2,897 |
| Number of Vehicles | 1 | 8 | 1,768 |
| Number of Casualties | 1 | 10 | 1,162 |
| Day of Week | 1 | 7 | 4,1 |
| Local Authority (District) | 1 | 570 | 26,425 |
| 1st Road Class | 1 | 6 | 3,859 |
| 1st Road Number | 0 | 5 210 | 493,266 |
| Road Type | 1 | 9 | 5,194 |
| Speed Limit | 20 | 70 | 31,119 |
| Junction Detail | 0 | 9 | 2,859 |
| Junction Control | -1 | 4 | 2,130 |
| 2nd Road Class | -1 | 6 | 3,126 |
| 2nd Road Number | -1 | 5 210 | 208,529 |
| Pedestrian Crossing-Human Control | 0 | 2 | 0,003 |
| Pedestrian Crossing-Physical Facilities | 0 | 8 | 1,193 |
| Light Conditions | 1 | 7 | 1,904 |
| Weather Conditions | 1 | 9 | 1,188 |
| Road Surface Conditions | 1 | 5 | 1,169 |
| Special Conditions at Site | 0 | 7 | 0,043 |
| Carriageway Hazards | 0 | 7 | 0,014 |
| Urban or Rural Area | 1 | 2 | 1,015 |
| Did Police Officer Attend Scene of Accident | 1 | 3 | 1,199 |
| Vehicle Reference | 1 | 5 | 1,223 |
| Casualty Reference | 1 | 6 | 1,007 |
| Casualty Class | 1 | 3 | 1,607 |
| Sex of Casualty | 1 | 2 | 1,345 |
| Age of Casualty | -1 | 97 | 32,920 |
| Age band of Casualty | -1 | 11 | 5,825 |
| Casualty Severity | 2 | 3 | 2,903 |
| Pedestrian Location | 0 | 10 | 1,160 |
| Pedestrian Movement | 0 | 9 | 0,810 |
| Car Passenger | 0 | 2 | 0,093 |
| Bus or Coach Passenger | 0 | 4 | 0,164 |
| Pedestrian Road Maintenance Worker | -1 | -1 | -1 |
| Casualty Type | 0 | 90 | 5.294 |
| Casualty Home Area Type | -1 | 3 | 0.534 |
| Vehicle Type | 1 | 90 | 8.839 |
| Towing and Articulation | 0 | 5 | 0.007 |
| Vehicle Manoeuvre | 1 | 18 | 12.758 |
| Vehicle Location – Restricted Lane | 0 | 9 | 0.032 |
| Junction Location | 0 | 8 | 3.082 |
| Skidding and Overturning | 0 | 5 | 0.048 |
| Hit Object in Carriageway | 0 | 11 | 0.283 |
| Vehicle Leaving Carriageway | 0 | 8 | 0.109 |
| Hit Object off Carriageway | 0 | 10 | 0.191 |
| 1st point of Impact | -1 | 4 | 1.957 |
| Was Vehicle Left Hand Drive | 1 | 2 | 1.001 |
| Journey Purpose of Driver | 1 | 15 | 13.25 |
| Sex of Driver | 1 | 3 | 1.269 |
| Age of Driver | -1 | 97 | 33.228 |
| Age Band of Driver | -1 | 11 | 5.774 |
| Engine Capacity(CC) | -1 | 15 000 | 1 396.915 |
| Propulsion Code | -1 | 8 | 0.577 |
| Age of Vehicle | -1 | 42 | 3.579 |
| Driver IMD Decile | -1 | 10 | 2.893 |
| Driver Home Area Type | -1 | 3 | 0.585 |

|  |  |  |
| --- | --- | --- |
| **Atribút** | **Najmenej** | **Najviac** |
| Time | 23:56 | 15:30 |
| Local Authority | E09000027 | E09000033 |
| LSOA | E01004754 | E01000004 |

Atribút Date má najstarší dátum 1.1.2005 a najskorší 31.12.2005. Jednotlivé atribúty sa dajú analyzovať aj pomocou grafov. My sme si znázornili viacero grafov. Ukážky sú na nasledujúcich obrázkoch



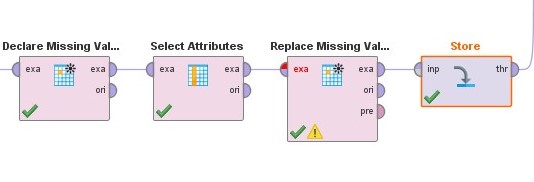
Histogram pre atribúty Day of Week a Age of Driver



Boxplot pre atribúty Speed limit a Day of Week

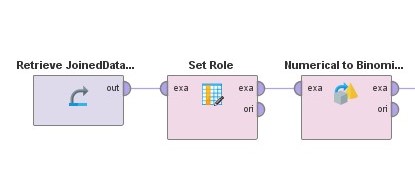
Príprava dát

Ako prvé sme odstránili atribút Pedestrian Road Maintenance Worker, lebo neobsahoval žiadne dáta. Za ďalšie sme chýbajúce hodnoty, ktoré boli označené ako -1, nahradili priemernou hodnotou daného atribútu.



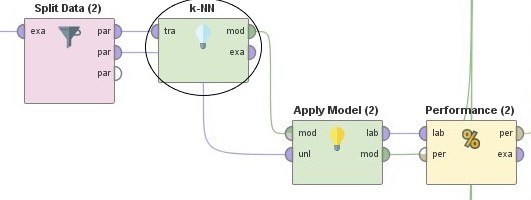
Proces nahradenia chýbajúcich dát

Následne sme mohli naše dáta klasifikovať. Za náš cieľový atribút sme si vybrali Sex of Casualty, ktorý obsahoval čísla 1 (Male) a 2 (Female). Tento atribút vlastne hovorí o tom, že či obeť nehody bola muž alebo žena. To sme sa pokúsili klasifikovať. Na začiatok sme previedli tento atribút na binominálny, aby sme ho mohli použiť na klasifikáciu.



Proces binomilizácie

Po tomto sme mohli použiť rôzne klasifikačné metódy .Na nasledujúcom obrázku je vidieť metódu k-NN. Namiesto toho krúžku sme následne dosadili Naivný Bayes, rozhodovací strom alebo Random Forest. Dostupnú množinu dát je potrebné rozdeliť na trénovaciu a testovaciu množinu. Použili sme algoritmus rozdelenia podľa pomeru 70% trénovacia množina, 30% testovacia množina a pre porovnanie potom aj pomer 80% trénovacia a 20% testovacia množina.



Proces klasifikácie

Presnosti spomenutých metód vypíšeme v tabuľke.

|  |  |  |
| --- | --- | --- |
| **Metóda** | **Presnosť (pre 70/30)** | **Presnosť (pre 80/20)** |
| Rozhodovací strom | 80,23% | 80.50% |
| k-NN | 59,97% | 57.20% |
| Naivný Bayes | 65,47% | 66.40% |
| Random Forest | 81,30% | 81.15% |

Ako môžeme vidieť v tabuľke tak najlepšiu presnosť mal Random Forest a najhoršiu metóda k-NN. Z časového hľadiska najrýchlejšie zbehol rozhodovací strom, ale najdlhšie Random Forest. Pre rozhodovací strom a Random Forest sme použili ako kritérium gini index a pri k-NN sme nastavili hodnotu k na 5.

Preskúmali sme aj možnosť použitia krížovej validácie na spomenutých algoritmoch ako rozhodovací strom alebo Random Forest.

|  |  |
| --- | --- |
| **Krížová validácia** | **Presnosť** |
| Rozhodovací strom | 77,40% |
| k-NN | 58,84% |
| Naivný Bayes | 66,72% |
| Random Forest | 78,52% |

Z tabuľky je vidieť, že najlepšiu presnosť mal Random Forest a najhoršiu k-NN.

Rôzne metódy sa vyhodnocujú aj pomocou kontigenčných tabuliek. My sme vypísali pre tie metódy, kde sme množinu rozdelili v pomere 70/30.

**Kontigenčné tabuľky**

**Rozhodovací strom**

|  |  |  |  |
| --- | --- | --- | --- |
|  | True false | True true | Class precison |
| Pred. false | 1795 | 422 | 80.97% |
| Pred. true | 171 | 612 | 36.89% |
| Class recall | 91.30% | 59.19% |  |

Pri klasifikácii rozhodovacím stromom bolo správne klasifikovaných 1795 negatívnych a 612 pozitívnych prípadov. Chyba prvého druhu nastala u 171 prípadov. Chyba druhého druhu nastala u 422 prípadov.

**k-NN**

|  |  |  |  |
| --- | --- | --- | --- |
|  | True false | True true | Class precison |
| Pred. false | 1564 | 799 | 66.19% |
| Pred. true | 402 | 235 | 36.89% |
| Class recall | 79.55% | 22.73% |  |

Pri klasifikácii k-NN bolo správne klasifikovaných 1564 negatívnych a 235 pozitívnych prípadov. Chyba prvého druhu nastala u 402 prípadov. Chyba druhého druhu nastala u 799 prípadov.

**Naivný Bayes**

|  |  |  |  |
| --- | --- | --- | --- |
|  | True false | True true | Class precison |
| Pred. false | 1515 | 585 | 72.14% |
| Pred. true | 451 | 449 | 49.89% |
| Class recall | 77.06% | 43.42% |  |

Pri klasifikácii Naivným Bayesom bolo správne klasifikovaných 1515 negatívnych a 449 pozitívnych prípadov. Chyba prvého druhu nastala u 451 prípadov. Chyba druhého druhu nastala u 585 prípadov.

**Random Forest**

|  |  |  |  |
| --- | --- | --- | --- |
|  | True false | True true | Class precison |
| Pred. false | 1849 | 444 | 80.64% |
| Pred. true | 117 | 590 | 83.45% |
| Class recall | 94.05% | 57.06% |  |

Pri klasifikácii Random Forest bolo správne klasifikovaných 1849 negatívnych a 590 pozitívnych prípadov. Chyba prvého druhu nastala u 117 prípadov. Chyba druhého druhu nastala u 444 prípadov.

**Rozhodovací strom (krížová validácia)**

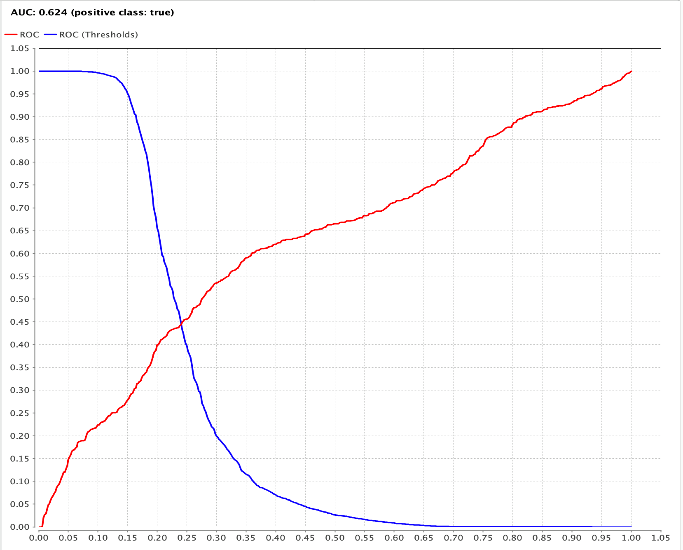
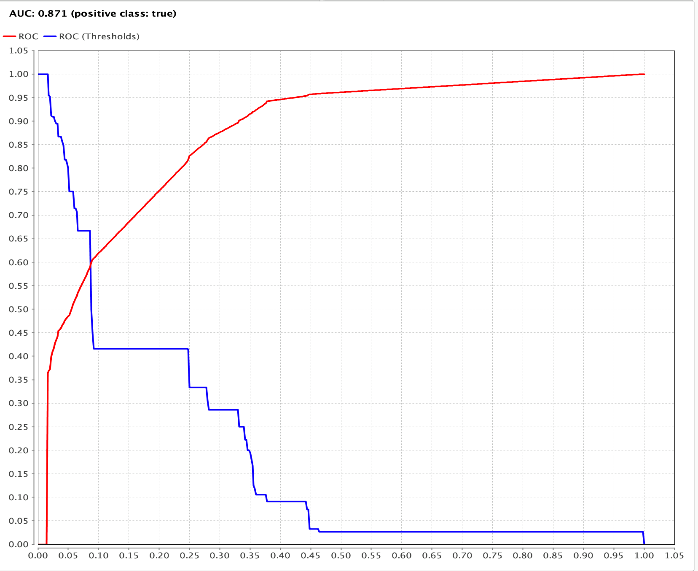
|  |  |  |  |
| --- | --- | --- | --- |
|  | True false | True true | Class precison |
| Pred. false | 5958 | 1458 | 80.34% |
| Pred. true | 117 | 590 | 76.93% |
| Class recall | 90.91% | 57.69% |  |

Pri klasifikácii Random Forest bolo správne klasifikovaných 5958 negatívnych a 590 pozitívnych prípadov. Chyba prvého druhu nastala u 117 prípadov. Chyba druhého druhu nastala u 1458 prípadov.

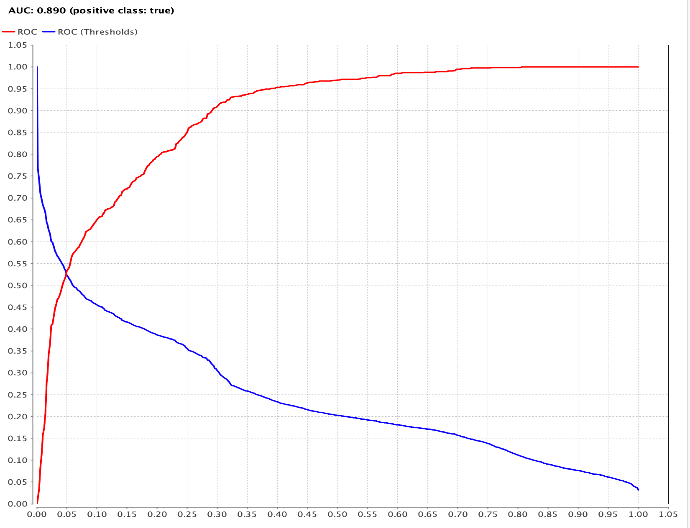
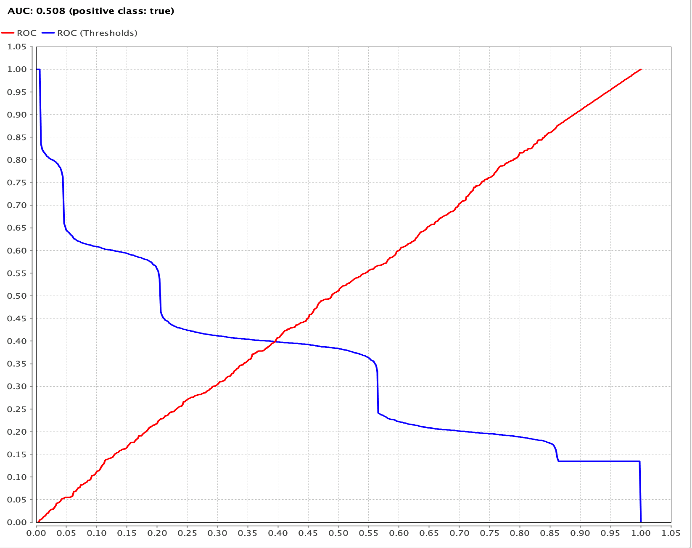
Modely sa môžu vyhodnocovať aj pomocou ďalších parametrov ako senzitivita alebo špecifickosť si ukážeme v nasledujúcej tabuľke.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model klasifikácie** | **Senzitivita** | **Špecifickosť** | **AUC** |
| Rozhodovací strom | 59.19% | 91.30% | 0.871 |
| k-NN | 22.73% | 79.55% | 0.508 |
| Naivný Bayes | 43.42% | 77.06% | 0.624 |
| Random Forest | 57.06% | 94.05% | 0.890 |

Model sa tiež dá vyhodnotiť pomocou ROC krivky, ktoré si znázorníme na nasledujúcich obrázkoch.

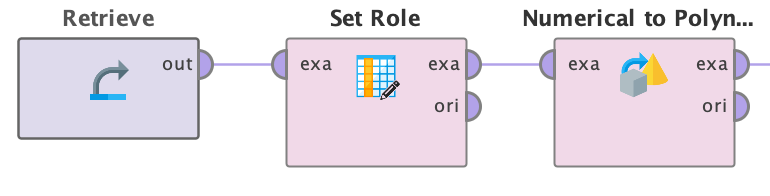


ROC krivka pre rozhodovací strom ROC krivka pre Naivný Bayes



ROC krivka pre k-NN ROC krivka pre Random Forest

Preskúmali sme aj ďalšie možnosti klasifikácie, konkrétne pre cieľový atribút Accident Severity. Atribút hovorí o závažnosti nehody, t.j. 1 = smrteľná, 2 = vážna, 3 = nezávažná, To sme sa pokúsili klasifikovať. Tento atribút sme previedli na polynominálny atribút.



Proces polynomizácie

Použili sme rovnakú sadu klasifikačných algoritmov, dosiahnuté presnosti sú v tabuľke.

|  |  |
| --- | --- |
| **Metóda** | **Presnosť** |
| Rozhodovací strom | 99,40% |
| k-NN | 88,90% |
| Naivný Bayes | 96,87% |
| Random Forest | 99,57% |
| k-nn (krížová validácia) | 88,64% |

Pre tento atribút mal najlepšiu presnosť algoritmus Random Forest. Krížová validácia dáva opäť výsledky porovnateľné s rozdelením podľa pomeru 70%/30%. Ako ďalšie si znázorníme kontigenčné tabuľky.

**Kontigenčné tabuľky**

**Rozhodovací strom**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | True 2 | True 3 | True1 | Class precision |
| Pred. 2 | 292 | 3 | 0 | 80.97% |
| Pred. 3 | 15 | 2689 | 0 | 36.89% |
| Pred. 1 | 0 | 0 | 0 | 0.00% |
| Class recall | 95.11% | 99.89% | 0.00% |  |

**k-NN**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | True 2 | True 3 | True1 | Class precision |
| Pred. 2 | 5 | 31 | 0 | 13.89% |
| Pred. 3 | 302 | 2661 | 0 | 89.81% |
| Pred. 1 | 0 | 0 | 0 | 0.00% |
| Class recall | 1.63% | 98.85% | 0.00% |  |

**Naivný Bayes**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | True 2 | True 3 | True1 | Class precision |
| Pred. 2 | 288 | 75 | 0 | 79.34% |
| Pred. 3 | 19 | 2617 | 0 | 99.28% |
| Pred. 1 | 0 | 0 | 0 | 0.00% |
| Class recall | 93.81% | 97.21% | 0.00% |  |

**Random Forest**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | True 2 | True 3 | True1 | Class precision |
| Pred. 2 | 294 | 0 | 0 | 100% |
| Pred. 3 | 13 | 2692 | 0 | 99.52% |
| Pred. 1 | 0 | 0 | 0 | 0.00% |
| Class recall | 95.77% | 100% | 0.00% |  |

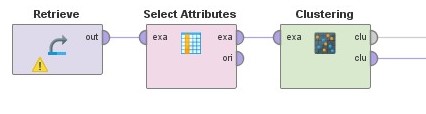
**k-NN (krížová validácia)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | True 2 | True 3 | True1 | Class precision |
| Pred. 2 | 8 | 118 | 0 | 6.35% |
| Pred. 3 | 1017 | 8856 | 1 | 89.69% |
| Pred. 1 | 0 | 0 | 0 | 0.00% |
| Class recall | 0.78% | 98.69% | 0.00% |  |

Z kontigenčných tabuliek je aj vidieť, že našťastie veľa smrteľných nehôd sa v našom datasete ani nenachádzalo.

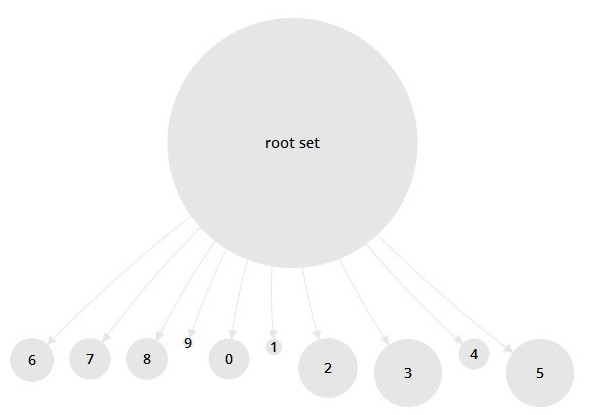
**Popisné modely**

Na popisnú úlohu sme si vybrali zhlukovanie, ktorého proces vyzerá:



Proces zhlukovanie

Následne sme nastavili počet zhlukov najprv na 10. Výsledkom bude rozdelenie do 10 zhlukov (čísluje od 0 po 9) ako je na nasledujúcom obrázku.



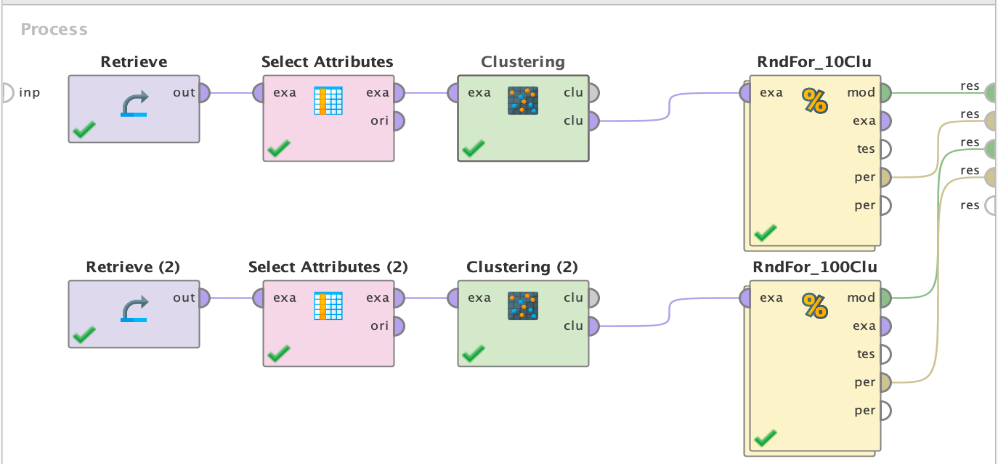
Vytvorené zhluky

|  |  |  |  |
| --- | --- | --- | --- |
| **Zhluk** | **Počet záznamov** | **Zhluk** | **Počet záznamov** |
| Zhluk 0 | 972 | Zhluk 5 | 1673 |
| Zhluk 1 | 381 | Zhluk 6 | 1047 |
| Zhluk 2 | 1449 | Zhluk 7 | 991 |
| Zhluk 3 | 1668 | Zhluk8 | 1000 |
| Zhluk 4 | 733 | Zhluk 9 | 86 |

Za ďalšie sme skúsili dať až 100 zhlukov. Časovo oveľa dlhšie zhlukovalo ako pri 10 zhlukoch a taktiež centroid so zhlukmi nie je taký prehľadný. Vypísať všetky zhluky s počtom záznamou by nemalo žiaden zmysel, preto len na znázornenie vypíšeme aspoň pár zhlukov.

|  |  |  |  |
| --- | --- | --- | --- |
| **Zhluk** | **Počet záznamov** | **Zhluk** | **Počet záznamov** |
| Zhluk 0 | 104 | Zhluk 5 | 139 |
| Zhluk 1 | 65 | Zhluk 6 | 150 |
| Zhluk 2 | 111 | Zhluk 7 | 164 |
| Zhluk 3 | 123 | Zhluk8 | 37 |
| Zhluk 4 | 82 | Zhluk 9 | 112 |

Na proces zhlukovania sme napojili krížovú validáciu, ktorá obsahovala Random Forest.



Proces zhlukovanie a Random Forest v krížovej validácií

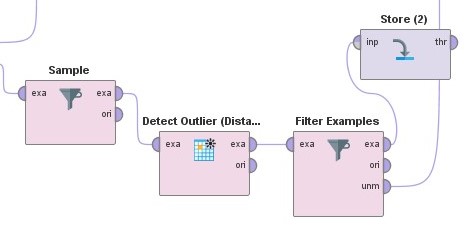
Taktiež sme vyskúšali namiesto krížovej validácie samostatne Random Forest a rozhodovací strom. Výsledky sa nachádzajú v tabuľke.

|  |  |  |
| --- | --- | --- |
| **Metóda** | **Presnosť pre 10 zhlukov** | **Presnosť pre 100 zhlukov** |
| Random Forest (krížová validácia) | 92.77% | 50.65% |
| Random Forest | 94.27% | 43.34% |
| Rozhodovací strom | 95.37% | 33.53% |

Z tabuľky je vidieť, že pri všetkých metódach je lepšia presnosť pre menší počet zhlukov. Najlepšiu presnosť mal rozhodovací strom pre 10 zhlukoch. Na druhej strane pri 100 zhlukoch mal najhoršiu presnosť.

**Detekcia anomálií**

Na detekciu anomálii sme použili operátor Detect Outlier (Distances), celý proces je znázornený na obrázku.



Proces detekcie anomálií

Pri tomto procese sme však museli zmenšiť náš dataset, lebo pre 10 000 záznamov to vôbec nezbehlo. Zmenšili sme na 7500 záznamov a potom proces zbehol.

TU ESTE DACO K ANOMALIAM

!

!

!

Po detekcii anomálií sme získané dáta opäť využili pre klasifikáciu, kde ako cieľový atribút sme použili Sex of Casualty. Výsledky procesov zobrazíme v tabuľke.

|  |  |
| --- | --- |
| **Metóda** | **Presnosť** |
| Rozhodovací strom | 77,70% |
| k-NN | 57,27% |
| Naivný Bayes | 65,97% |
| Random Forest | 80,04% |
| Random Forest (krížová validácia) | 76,73% |

Po detekcií anomálií najlepšiu presnosť mal Random Forest a najhoršiu k-NN.