Fakulta elektrotechniky a informatiky

Fakulta elektrotechniky a informatiky

Katedra kybernetiky a umelej inteligencie

Predmet : **Objavovanie znalostí**
kurz 2024 / 2025

Zadanie:

Semestrálne zadanie (skupina PO 15:10)

Spracovaly:
Yehor Nepokrytyi, Mykhailo Ruzmetov

KKUI
Katedra kybernetiky a umelej inteligencie

Obsah

1 Úvod

Táto dokumentácia opisuje našu aplikáciu metodiky CRISP-DM na dátový súbor „UK Road Safety: Traffic Accidents and Vehicles“.

2 Pochopenie problému a cieľa

2.1 Biznis cieľ

Zlepšiť pochopenie faktorov, ktoré vedú k rôznym úrovniam závažnosti dopravných nehôd, s cieľom vyvinúť modely, ktoré by dokázali predikovať takéto situácie a mohli by tiež pomôcť pri budúcich zlepšeniach bezpečnosti cestnej premávky.

2.2 Cieľ KDD (Knowledge Discovery in Databases)

Identifikovať dôležité atribúty, ktoré ovplyvňujú závažnosť dopravných nehôd, a vytvoriť prediktívne/klasifikačné a popisné modely na analýzu pomocou existujúcich metód.

2.3 Zdroje

Dataset *UK Road Safety: Accidents and Vehicles* stiahnutý z Kaggle [1].

- Súbor `Accident_Information.csv` obsahuje informácie o nehodách.
- Súbor `Vehicle_Information.csv` obsahuje detaily o vozidlách zapojených do nehody.

3 Pochopenie a príprava dát

3.1 Vytvorenie datasetu

- Dáta boli skombinované na základe kĺúča `Accident_Index`.
- Stĺpce s vysokým percentom chýbajúcich hodnôt ($> 40\%$) boli odstránené, aby sa v ďalšom kroku nevytvorilo príliš veľa syntetických údajov. Samozrejme, čím viac syntetických údajov je v datasete, tým je menej presný vzhľadom na skutočné príznaky.

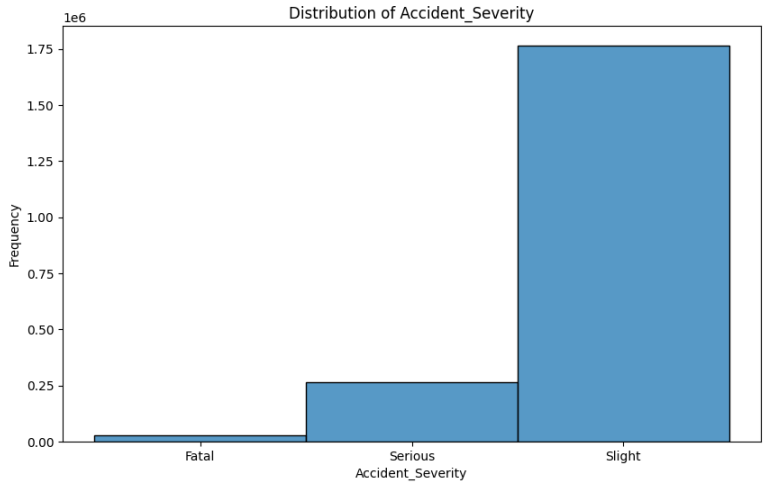
3.2 Spracovanie chýbajúcich hodnôt

Pokiaľ ide o spracovanie chýbajúcich hodnôt, rozhodli sme sa použiť niekoľko spôsobov doplnenia údajov:

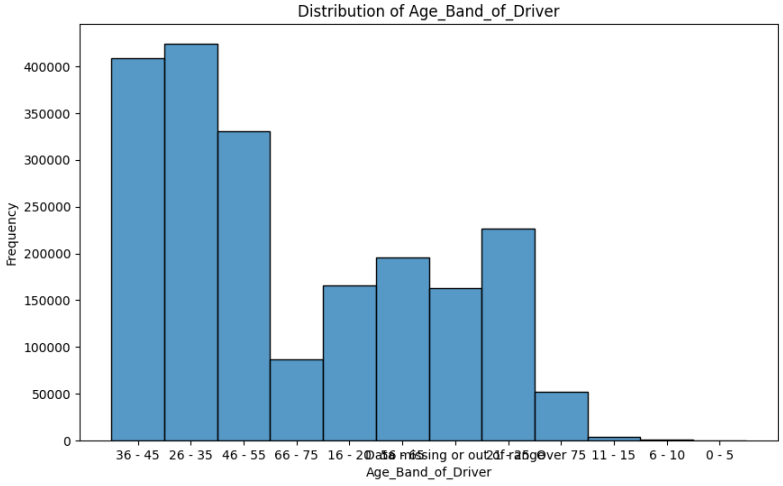
- SimpleImputer so stratégiou „most_frequent“ pre kategorické hodnoty a stratégiu „mean“ pre "Driver_IMD_Decile".
- KNNImputer pre pokročilejšie spracovanie numerických atribútov.

3.3 Oversampling dat pre klasifikačné modely

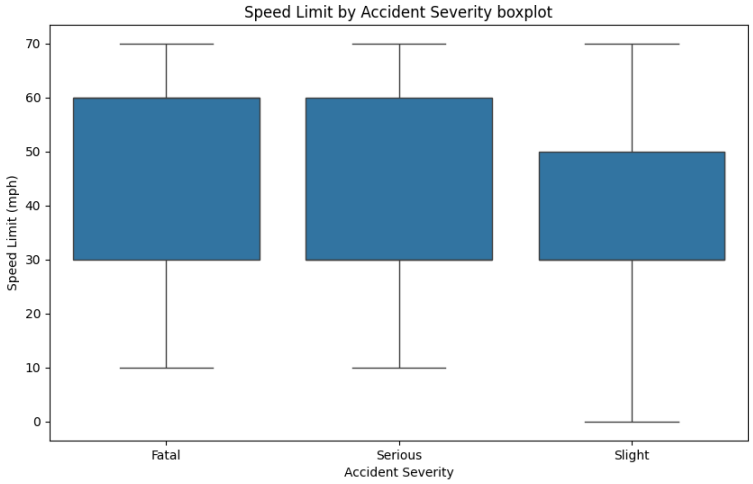
Ako vidíte na tomto obrázku, v našich údajoch veľmi chýbajú údaje o smrteľných a vážnych nehodách, ale vzhľadom na poznatky z hodín tohto predmetu sme sa rozhodli použiť vynikajúcu metódu SMOTE, ktorá nám poskytla chýbajúce údaje na lepšie predpovedanie nehodovej situácie.



Obr. 1: Distribúcia Accident_Severity: je to náš cieľový atribút a ako vidíte, je nevyvážený.



Obr. 2: Age_Band_of_Driver distribúcia: zaujímavá vizualizácia, ktorá zobrazuje počet nehôd s vodičmi rôzneho veku.



Obr. 3: Obmedzenie rýchlosti podľa závažnosti nehody - boxploty: grafy „serious“ a „fatal“ sú takmer rovnaké, ale „slight“ ukazuje, že limit rýchlosti bol v priemere nižší.

4 Modelovanie

4.1 Klasifikačné modelovanie

Po predspracovaní dátového súboru sme sa rozhodli skúsiť definovať triedy pomocou viacerých modelov:

- Random Forest
- SVM
- Gradient Boosting
- XGBoost
- MLPClassifier

Došli sme k rozhodnutiu, že na zvládnutie tejto úlohy sú najlepšie dva modely, a to Random Forest a XGBoost. Preto sa zameriame na výber najlepších hyperparametrov pre tieto modely pomocou GridSearchCV [2].

- Po algoritme GridSearchCV sú naše modely nasledovné:
  - Random Forest: Nastavené parametre zahŕňajú maximálnu hĺbku 20 a 300 stromov.
  - XGBoost: Optimalizované parametre ako learning rate (0.1), maximálna hĺbka (15) a počet stromov (300).
- Vyhodnotenie modelov:

  - Tabulka kde ukazane precision, recall, accuracy, macro avg, weighted avg, support a f1-score. Vytvorene pomocou sklearn.metrics.classification_report.

4.2 Popisné modelovanie

- Na túto úlohu sme sa na základe poznatkov získaných z prieskumu a spracovania dát rozhodli použiť:
  - Identifikáciu anomálií pomocou K-Means clusteringu a Mahalanobisovej vzdialenosti.
- Výsledky sú znázornené pomocou 3 grafov, ktoré predstavujú:
  - Detekciu anomálie pomocou Mahalanobisovej vzdialenosti
  - Hustotu anomálnych dátových bodov
  - Hustotu normálnych dátových bodov

5 Vyhodnotenie

Celý program spolu s predspracovaním údajov, trénovaním modelu a vyhľadávaním anomálií trvá 2 hodiny a 37 sekúnd, čo nie je málo času, ale treba uvážiť, že väčšinu z neho zaberie trénovanie modelu.

5.1 Prediktívne modely

- Random Forest a XGBoost dosiahli nasledujúce úspesnosti:
  - RF:

	precision	recall	f1-score	support
0	0.635341168740889	0.8172803917980489	0.714916938340549	530069.0
1	0.6419855615967888	0.2801804692956789	0.39010722560482786	529730.0
2	0.5856451006902019	0.7480190294094308	0.6569475326054491	529286.0
accuracy	0.6151659602853214	0.6151659602853214	0.6151659602853214	0.6151659602853214
macro avg	0.6209906103426266	0.6151599635010528	0.5873238988502753	1589085.0
weighted avg	0.6210035475017984	0.6151659602853214	0.5873316650471748	1589085.0

Obr. 4: Random Forest report

– XGBoost:

	precision	recall	f1-score	support
0	0.8278558630584842	0.875157762479979	0.85084989797556	530069.0
1	0.7389238169581637	0.5289373831952127	0.6165413533834586	529730.0
2	0.6662237870975788	0.8175863332867297	0.7341848606618636	529286.0
accuracy	0.740567685177319	0.740567685177319	0.740567685177319	0.740567685177319
macro avg	0.7443344890380755	0.7405604929873072	0.7338587040069607	1589085.0
weighted avg	0.7443741312834271	0.740567685177319	0.7338835706448062	1589085.0

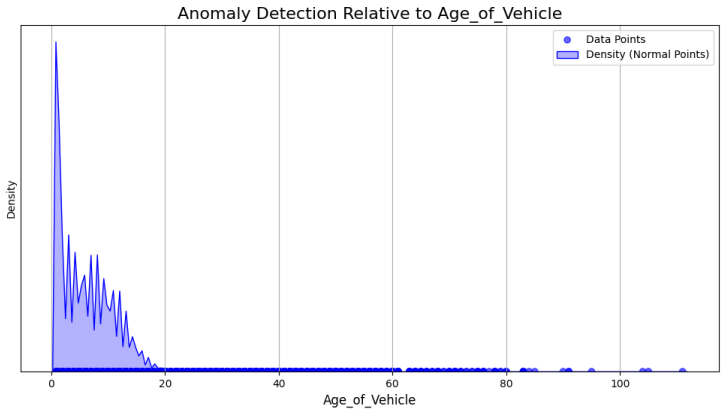
Obr. 5: XGBoost report

Ako ste mohli všimnúť, algoritmus XGBoost sa v našich údajoch celkom dobre orientuje a vykazuje pomerne dobré výsledky v klasifikácii na rozdiel od algoritmu Random Forest, ktorý síce svoju úlohu zrejme celkom nezvládol, ale stále je lepší ako ostatné modely, ktoré boli použité vo fáze fitovania modelu. Preto by sme pre tento súbor údajov odporúčali použiť XGBoost ako hlavný model na klasifikáciu zložitosti prípadov dopravných nehôd.

5.2 Popisné modely

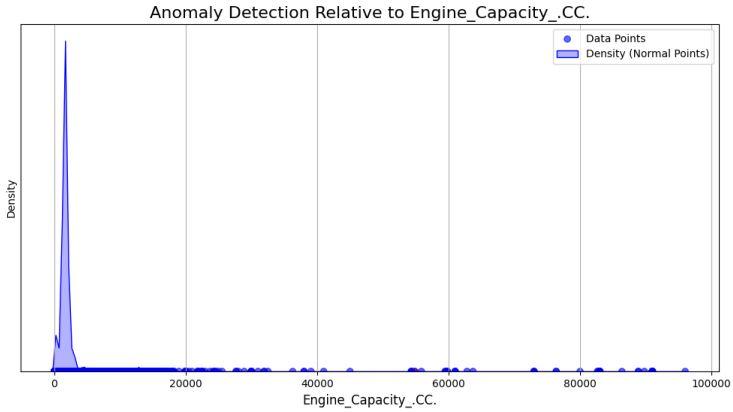
Po detekcii anomálie pre všetky atribúty sme si uvedomili, že naše výsledky najlepšie reprezentujú:

- Hustota (normalne body):
  - Age_of_Vehicle :
    - Graf znázorňuje rozloženie hustoty normálnych bodov (modrou farbou). Väčšina údajov je sústredená medzi 0 a 20 rokmi, pričom na začiatku je vrchol, ktorý naznačuje dominantný počet automobilov mladších ako 5 rokov.
    - Hustota po 20 rokoch prudko klesá, čo sa očakáva, keďže staršie autá sú menej rozšírené.



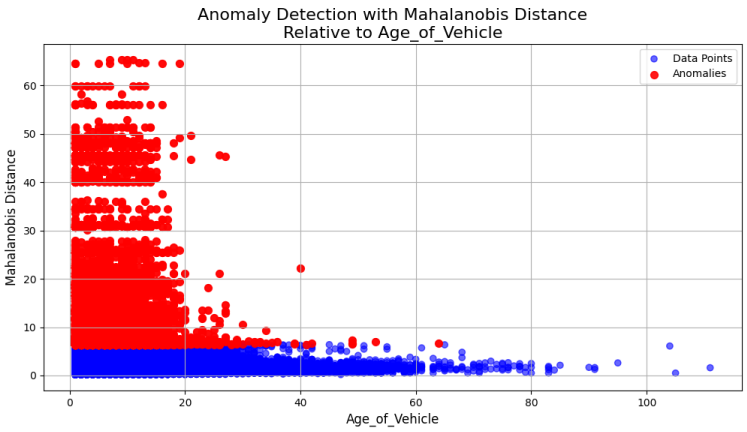
Obr. 6: Hustota (normalne body) pre Age_of_Vehicle

- Engine_Capacity_.CC. :
  - Normálne rozdelenie vykazuje vrchol v oblasti pod 3000 cm3. To odráža skutočnosť, že väčšina automobilov má motor tejto veľkosti.
  - Hodnoty nad 10 000 cm3 sa v normálnom rozdelení takmer nevyskytujú, čo poukazuje na ich zriedkavosť a možnosť anomálií.



Obr. 7: Hustota (normalne body) pre Engine_Capacity_.CC.

- Mahalanobisovej vzdialenosti:
  - Age_of_Vehicle :
    - Väčšina normálnych bodov (modré) sa nachádza v oblasti s nízkou vzdialenosťou (0-10), čo znamená, že patria do hlavnej skupiny.
    - Anomálie sú badateľné v oblasti automobilov starších ako 40 rokov a v oblastiach nových automobilov s neobvyklými charakteristikami.

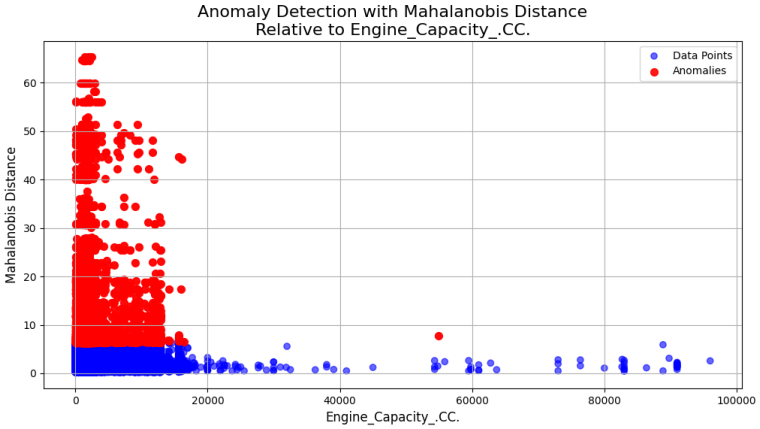


Obr. 8: Mahalanobisovej vzdialenosti pre Age_of_Vehicle

– Engine_Capacity_.CC. :

∗ Pozorovali sa mnohé anomálie s vysokými Mahalanobisovými vzdialenosťami (červené body), najmä pri hodnotách zdvihového objemu motora > 10 000 cm3.

∗ Normálne údaje sú sústredené v rámci nízkych vzdialeností (0-10), ktoré zodpovedajú motorom s menším objemom.

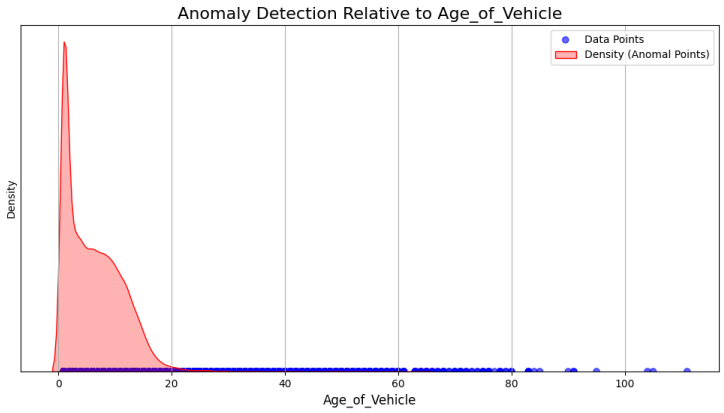


Obr. 9: Mahalanobisovej vzdialenosti pre Engine_Capacity_.CC.

- Hustota (anomalne body)

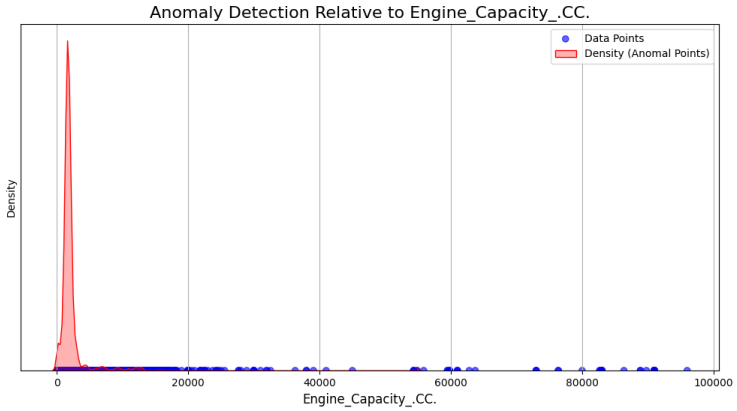
– Age_of_Vehicle :

∗ Anomálie (znázornené červenou hustotou) sa nachádzajú v oblasti nových automobilov, ako aj medzi extrémne starými automobilmi (staršími ako 20 rokov).



Obr. 10: Hustota (anomalne body) pre Age_of_Vehicle

– Engine_Capacity_.CC. :
  * Body anomálie zahŕňajú hodnoty pod typickými normálnymi hodnotami, ako aj extrémne vysoké hodnoty (napr. $>$ 10 000 cm3).
  * Rozloženie anomálií zdôrazňuje zriedkavosť automobilov s veľmi malými alebo obrovskými motormi.



Obr. 11: Hustota (anomalne body) pre Engine_Capacity_.CC.

6 Záver

Úloha bola úspešne vykonaná s ohľadom na všetky kroky metodiky CRISP-DM. Výsledky získané v priebehu tejto práce ukazujú všetky aspekty možnej práce s týmto datasetom a môžu byť použité na ďalšiu implementáciu v projektoch súvisiacich s bezpečnosťou cestnej premávky.

Referencie

[1] Andreas Tsiaras. *UK Road Safety - Accidents and Vehicles*. https://www.kaggle.com/datasets/tsiaras/uk-road-safety-accidents-and-vehicles. Accessed: 2024-12-11. 2024.

[2] GeeksforGeeks. *SVM Hyperparameter Tuning using GridSearchCV*. https://www.geeksforgeeks.org/svm-hyperparameter-tuning-using-gridsearchcv-ml/. Accessed: 2024-12-11. 2024.