report

May 23, 2020

0.1 Analiza prometnih nesre leta 2019 (v Sloveniji)

1 Projektna naloga pri predmetu Podatkovno rudarjenje

Mentor: doc. dr. Toma Curk

1.1 lani skupine

Ime in priimek	Vpisna tevilka
Bian Klannik	63180147
Jan Weissenbach	63180322
Gal Lindi	63180182
Tadej Lonikar	63180184
Obej Golob	63180105

1.2 Uvod

Pri projektni nalogi smo si izbrali podatkovno zbirko, ki belei razline podatke o prometnih nesreah v letu 2019. Nad podatki smo izvedli analizo, pri tem smo analizirali podatke, ki so se nam zdeli zanimivi. Rezultate smo predstavili tudi s pomojo razlinih vizualizacij.

1.3 Opis podatkovne zbirke

Za temo nae projektne naloge smo si izbrali podatkovno zbirko Prometne nesree v letu 2019.

Atributi se nahajajo v `podatki/pn2019.csv`.

Pomembni atributi, ki smo jih analizirali: UraPN, VzrokNesrece, TipNesrece, VremenskeOkoliscine, StanjeVozisca, Povzrocitelj, Starost, Spol, Drzavljanstvo, UporabaVarnostnegaPasu, VozniskiStazVLetih, VozniskiStazVMesecih, VrednostAlkotesta.

1.4 Cilji

- Vpliv alkohola, ure, spola, voznikega staa na tevilo povzroenih nesre
- Vpliv tipa nesree in uporabe varnostnega pasu na tip pokodbe
- Odkriti regije v Sloveniji, kjer je gostota nesre najveja
- Zgraditi uspeen klasifikacijski model, ki napove izzid nesree (tip pokodbe/materialna koda)

1.5 Predstavitev rezultatov

1.5.1 Vpliv alkohola na tevilo povzroenih nesre

Pri obdelavi podatkov smo ugotovili, da so v letu 2019 pijani vozniki povzroili **6.3%** vseh nesre. Od vseh nesre je imelo smrtno rtev **0.5%** nesre. Proti priakovanjem pa se je od vseh nesre, ki jih je povzroil pijan voznik, s smrtno rtvijo konalo samo **0.2%**.

Pri pijanih voznikih smo priakovali, da se bo tevilo nesre proti veeru vealo. Najveje tevilo nesre smo priakovali nekje med 1. in 3. uro zjutraj.

Po vizualizaciji smo ugotovili, da je najve nesre zaradi pijanih voznikov **ob 19. uri**. Presenetljivo po tem zane tevilo nesre padati.

1.5.2 Vpliv ure na tevilo povzroenih nesre

1.5.3 Vpliv spola na tevilo povzroenih nesre

Splono mnenje (mokih) je, da so enske slabe voznice, zato smo se odloili preveriti tevilo povzroenih nesre glede na spol..

Po vizualizaciji opazimo, da so moki povzroili priblino **dvakrat ve** nesre kot enske. Da bi lahko tono doloili, kdo procentualno povzroi ve nesre, bi rabili dostop do podatkov, ki povejo, koliko voznikov z izpitom je mokih in koliko ensk. Vendar pa takne odprte podatkovne zbirke nismo nali, zato smo pogledali tevilo mokih in ensk v Sloveniji in predpostavili, da je dele voznikov glede na spol podoben kot dele vsega prebivalstva glede na spol.

Po podatkih iz SiStat je leta 2019 v Sloveniji prebivalo 1,043,933.25 mokih in 1,043,211.5 ensk (povpreno), tako da je dele mokih in ensk razporejen enakomerno 50-50. e predpostavimo, da je tudi dele voznikov glede na spol priblino 50-50, ugotovimo, da moki procentualno povzroijo veliko ve nesre kot enske.

1.5.4 Vpliv doline voznikega staa na tevilo povzroenih nesre

Splono mnenje ljudi je, da mladi vozniki povzroijo ve nesre kot izkueni. Zato smo se odloili, da bomo to hipotezo preverili.

Kot vidimo na grafu, je tevilo ljudi z dolino voznikega staa 0 let sumljivo veliko. Zaradi tega je zgrajen regresijski model neuporaben (MSE je enak **87851.76**).

Predpostavili smo, da so v podatkovni zbirki vnosi z dolino staa 0 let in 0 mesecev neznane vrednosti, tako da smo jih ignorirali. Takih vnosov je 8542.

Nato smo nad novimi podatki zgradili linearni regresijski model in izraunali Pearsonov koeficient, ki je enak **-0.95**. To pomeni, da sta si dolina voznikega staa in tevilo povzroenih nesre **obratno sorazmerna**. MSE na tem regresijskem modelu je enak **1832.32**.

Izraunali smo tudi, da ljudje, mlaji od 21 let, povzroijo **8%** vseh prometnih nesre. Od vseh nesre s smrtnim izzidom, so jih **6.45%** povzroili ljudje, mlaji od 21 let.

1.5.5 Vpliv tipa nesree na pokodbe

Iz vizualizacije smo izpustili nesree, kjer so udeleenci brez pokodb oziroma z lajimi pokodbami. Kot vidimo, je vrsta pokodbe odvisna od tipa nesree. Iz grafa vidimo, da je pri elnem trenju najve smrtnih primerov, medtem ko pri povoenju ivali ni bilo nobene smrtne rtve (ivali not included).

1.5.6 Vpliv uporabe varnostnega pasu na tip pokodbe

Ko smo obdelali podatke, smo ugotovili naslednje:

- tevilo udeleencev prometnih nesre, ki so bili pripeti z varnostnim pasom: **26897**
- tevilo udeleencev prometnih nesre, ki so niso bili pripeti z varnostnim pasom: **3006**
- Odstotek ljudi, ki so bili pripeti: **89.9**

Vidimo, da je bilo **90%** ljudi, udeleenih v prometnih nesreah, pripetih z varnostnim pasom. Po vizualizaciji ugotovimo naslednje:

- Odstotek nesre s smrtnim izidom, kjer je bil udeleenec pripet z varnostnim pasom: **0.21%**
- Odstotek nesre s smrtnim izidom, kjer udeleenec ni bil pripet z varnostnim pasom: **1.23%**
- Odstotek nesre s hudimi telesnimi pokodbami, kjer udeleenec ni bil pripet z varnostnim pasom: **8.51%**
- tevilo mrtvih ljudi, ki niso bili pripeti z varnostnim pasom: **37**

Prej smo ugotovili, da je bilo **90%** udeleencev prometnih nesre pripetih z varnostnim pasom. Od ostalih **10%** ljudi, ki niso bili pripeti z varnostnim pasom jih je **1.23%** umrlo v prometni nesrei, kar je **1%** ve, kot pa pri udeleencih, kateri so bili pripeti. Hude telesne pokodbe jih je imelo okrog **9%**.

1.5.7 Gruenje nesre po koordinatah

K-means clustering Za zaetek smo eleli gruiti nesree po lokaciji tako, da bi dobili nekakne "psevdo" slovenske pokrajine, ki so definirane po lokaciji prometnih nesre. Za dosego tega cilja smo uporabili "k-means clustering". tevilo cluster-jev (gru) smo izbrali 12, saj je toliko slovenskih pokrajin. K-means smo lahko uporabili, saj smo poznali strukturo naih podatkov in tono vedeli, koliko cluster-jev potrebujemo.

Algoritem je gruil nesree v 12 pokrajin. Te pokrajine se sicer od dejanskih geografskih pokrajin rahlo razlikujejo, vendar pa so v veini podobne.

HDBSCAN clustering Za cilj smo si zadali odkriti regije v Sloveniji, kjer je gostota nesre najveja. Za dosego tega cilja smo se odloili za uporabo clustering algoritma HDBSCAN. HDBSCAN je "density based" algoritem. To pomeni, da tvori grue glede na gostoto regije. eleli smo si najti regije, kjer se je leta 2019 zgodilo vsaj 500 nesre.

Algoritem je tvoril le 5 gru, kjer je bila gostota dovolj velika in kjer je bilo v grui ve kot 500 nesre. Iz grafa lahko vidimo, da je najveja gostota nesre v letu 2019 bila v okolici Ljubljane, Celja, Maribora, Primorske in Grosuplja. Iz grafa lahko tudi razberemo, kje potekajo avtoceste in ostale veje ceste, saj so bolj intenzivne rne barve. Ker pa se na njih ni zgodilo ve kot 500 nesre in gostota nesre ni dovolj velika, jih je algoritem obravnaval kot nepomembne in jih zato ni gruil.

1.5.8 Gradnja klasifikacijskega modela

Za cilj smo si zadali, da zgradimo uspeen klasifikacijski model, ki napove izzid nesree.

Moni izzidi: z materialno kodo, z lajo telesno pokodbo, s hudo telesno pokodbo, s smrtnim izzidom.

Na vseh atributih smo zgradili 3 klasifikacijske modele: **Logistic Regression**, **Random Forest** in **MLP**.

Klasifikacijska tonost: * Logistic Regression: 0.6791 * Random Forest: 0.7293 * MLP: 0.6368 Najbolj uspeen je bil model Random Forest, zato smo uenje nadaljevali na tem modelu.

Nato smo model zgradili e na najboljih 6 atributih. Te atribute smo izbrali na 2 naina: **Chi-Square Test** in **Mutual Information**.

Klasifikacijska tonost: * Chi-Square Test: 73.54 * Mutual Information: 73.20

Najbolji model je bil zgrajen z atributi, ki jih je izbral Chi-Square Test. Zato smo za gradnjo konnega modela uporabili te atribute. Na konni model je napovedal izzid nesree s 73.54% tonostjo.

Vrednost veinskega klasifikatorja je