**Univerzitet u Kragujevcu**

A logo of a school

Description automatically generated

**UVOD U NAUKU O PODACIMA**

**Analiza seta podataka:** polovni automobili

**STUDENT:** **MENTOR:**

Aleksić Filip prof Nenad Filipovic

**KRAGUJEVAC 2023**

Sadržaj:

**Uvod:**

U ovom radu će se govoriti o analizi seta podataka, u oblasti polovnih automobile. Poenta rada jeste da se ceo set podataka prođe i propusti kroz programski jezik Python i da se dođe do određenih zaključaka i da se ti zaključci posle kroz ovaj dokument predstave.

Sam rad će u svom sklopu sadržati određene isečke iz programa Python kojima će se prikazivati ceo set podataka koji se obrađuje. Ovim radom treba da damo određen osvrt na naš skup podataka, u sledećem smislu:

* Pre svega da odredimo koje biblioteke ćemo koristiti,
* Da uvezemo skup podataka,
* Od čega je sačinjen,
* Koji je tip podataka koji obrađujemo,
* Koji sve tipovi podataka postoje u posmatranom setu ,
* Koliko je praznih ćelija, redova, kolona,
* Da identifikujemo tačno koje su to prazne ćelije ili kolone,
* Da odredimo broj redova i kolona,
* Prikažemo podatke kroz grafike,
* Prikažemo podatke pre i posle obrade seta podataka,
* Da obrišemo prazna polja ili popunimo odgovarajućim vrednostima
* Da definišemo i promenimo tip podatka/kolone,

Sve gore što je navedeno treba da se odradi po segmentima i uz praćenje adekvatne hronologije. Takođe jako je bitno ukrštati podatke kako bismo uspeli da ispratimo analogiju samog seta podataka. Ukrštanjem ćemo dobiti kretanja različitih podataka uz praćenje zahteva koji mi postavimo kroz program.

* **Opis problema**

U ovom konkretnom slučaju kada govorimo o skupu podataka vezano za polovne automobile, imamo podatke koje kada ukrstimo možemo dobiti brojne informacije, kako o kolima, tako i o prodaju. Naš zadatak je da upravo kroz prikazivanje i ukrštanje podataka i rezultata iz ovog skupa podataka dođemo do nekih zaključaka o kretanju tržišta automobila i podacima koje su vezane i za samu marku automobile. Detaljnije o ovome u sledećem poglavlju.

* **Skup podataka i analiza**

Dobili smo skup podataka koji je sačinjen od 8128 redova i 13 kolona:

A blue rectangle with white lines

Description automatically generated  
slika1

Kolone od kojih je sačinjen ovaj set podataka su:

* Ime
* Godina proizvodnje
* Prodajna cena
* Pređeni kilometri
* Gorivo koje koristi
* Tip prenosa/menjač
* Broj vlasnika
* Potrošnja
* Kubikaža
* Snaga
* Obrtni moment
* Broj sedišta

Set podataka će se analizirati primenom nekoliko biblioteka kao što su:

* Pandas
* Matplotlib
* Missingno

Podaci koji su upisani u data setu, su različitog tipa što se vidi iz sledeće slike:

A screenshot of a computer program

Description automatically generated  
slika2

Ovde vidimo da je set podataka sačinjen od 3 int (cela broja), 1 float (decimalni broj), i 9 tekstualnih kolona. Ono što treba napomenuti jeste da je to sastav ovoga seta podataka na početku analize što ne znači da će na dalje ostati isti.

Postoje kolone koje su ispunjene brojevima ali u nastavku imaju dodatak slovni pa onda se karatkerisu kao object, a u stvari treba da se posmatraju kao brojevi. Zbog toga moramo da vršimo transformaciju kolona u brojevne vrednosti.

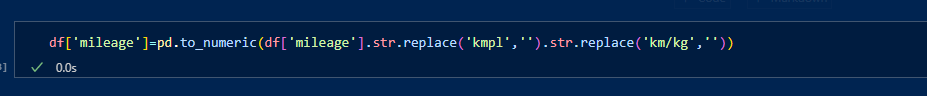
**Pre svega ćemo to pokazati na sledećoj slici:**

A screenshot of a computer

Description automatically generated  
slika3

Iz gornje slike vidimo da kolona mileage ili potrošnja nosi u nastavku kmpl, zbog toga ova kolona mileage se prepoznaje kao object što se vidi na slici 2. Pa smo dužni da u nastavku analize otklonimo sve što je tekst i da pretvorimo ovu kolonu u numeričku.

To ćemo odraditi na sledeći način:

  
slika4

Ovde vidimo da smo kolonu mileage putem funkcije pd.to\_numeric preveli iz stringa u intiger u ovom slučaju float 64 što se vidi iz sledeće slike:

A screenshot of a computer screen

Description automatically generated  
slika5

Kad uporedimo sliku 5 i sliku 2 videćemo da je na slici 5 osma stavka mileage u koloni dtype float 64 oznaka, što znači da smo je preveli iz object tip podatka u float 64 tačnije u decimlani zapis. I sada može da se putem pandas funkcija koristi kao brojevna vrednost.

Transformaciju podataka iz kolone mileage smo uradili na sledeći način. Prvo smo izvršili brisanje svih nedostajućih vrednosti tj null vrednosti. Null vrednosti su se nalazile u poslednjih 5 kolona. Kada smo odstranili nedostajuće vrednosti onda smo vršili promenu vrednosti kolona iz object u float.

Bitno je ispoštovati redosled kako ne bismo imali situaciju u kojoj ne možemo da prevedemo iz object u float zbog null vrednosti. Poslednjih 5 kolona su kolone koje moramo da dovedemo u numeričku vrednost kako bismo mogli da ih koristimo u analaizi a trenutno svi podaci u tih pet kolona se vode kao object, izuzev poslednje, broj sedišta, ona je float. To je grafički prikazano na slici 2 pa na slici 5 nakon brisanja nedostajućih vrednosti.

Ove nedostajuće vredosti se pre brisanja mogu prikazati grafički korišćenjem missingno biblioteke. Takođe grafik će biti prikazan i nakon brisanja nedostajućih vrednosti radi poređenja:

A screenshot of a computer

Description automatically generated  
slika 6: grafički prikaz null vrednosti u datom skupu podataka

A screenshot of a computer

Description automatically generated  
slika 7: grafički prikaz nakon sređivanja skupa podataka (data cleaning)

Za opciju dropna sam se opredelio iz razloga što nedostajući podaci ne prelaze 10% vrednosti skupa podataka. Jer se broj null vrednosti kretao u poresku oko 230 po koloni. Što smatram da nije od presudnog značaja i da će negativno uticati na rezultat i analizi datog seta podataka.

Nakon rešavanja kolone milage i prebacivanja u numeričke vrednosti, nastavio sam i za preostale kolone kao što su engine, max power, torque.

Kolona name je kolona koju treba razdvojiti na car i model, time ćemo dobiti dve nove kolone, ali takođe ukloniti postojeću kolonu iz data seta. To ćemo uraditi na sledeći način:

A screenshot of a computer

Description automatically generated

Pošto smo razdvojili postojeće kolone, sada je red da vidimo koji auto je najprodavaniji i to je Suzuki Maruti. Pored toga vidimo koeficijent korelacije između pređenih kilometara i prodajne cene da je negativan što dalje implicira da je između te dve varijable vlada obrnuta proporcija tačnije, što auto više prelazi kilometara to je jeftinije. Što dokazuje i grafik ispod. Da bi bio lepši pregled dinare sam pretvorio u evre kako bi grafik bio pregledniji.

Što se tiče deskriptivne statistike:

A screenshot of a computer screen

Description automatically generated

Vidimo da većina automobile 75% je mlađe od 2017 godišta, a cena isto manja od 5882 eura. Potrošnja je 75% automobile je manja od 22,3 milja po kilometru. Vidmo takođe da se starost automobile kreće u rasponu od 94 do 2020 godišta.

Automobil sa najviše konja je volvo xc90t8 sa 400hp, benzinski motor, dok je Maruti omni auto sa najmanje hp-32.8

Nakon razdvajanja kolone torque dobili smo 2 nove kolone torque value i torque rpm i pojavilo se novih praznih ćelija. Pa sam morao opet da očistim data set od istih.