Analiza podataka, njihovih veza i mogućnost primene algoritma mašinskog učenja nad skupom podataka radi određivanja realne cene

Opis problema

Problem koji je izabran da se rešava je analiza podataka sa nemačkog sajta *AutoScout24*, jednog od najvećih u Evropi za nove i polovne automobile. Sakupljeni su podaci o automobilima koji su oglašavani, a koji imaju godinu proizvodnje od 2011. do 2021. godine. Razlog izabranog problema je taj što se svakodnevno oglasi ogroman broj automobila, i bilo bi veoma korisno ako bismo mogli da prikažemo sve te podatke na grafički način i da pronađemo vezu između unetih podataka kako bismo predvideli cenu. To bi olakšalo kupovinu automobila kupcima, omogućavajući im realan prikaz vrednosti određenog automobila, čime bi se sprečilo da prodavci prodaju vozila po mnogo većoj ceni.

Skup podataka

Koristiće se osnovni podaci o automobilima koji su oglašavani, kao što su model, kilometraža i slično. Podaci su sledeći :

- Mileage - kilometraža
- Make – proizvođač automobila
- Model – model automobila
- Fuel – gorivo koje koristi
- Gear – vrsta menjača
- OfferType – stanje automobila (nov, korišćen, demo vozilo, itd.)
- Price – cena
- Hp – snaga motora
- Year – godina proizvodnje

Podaci su preuzeti sa sledećeg linka.

Tehnologije

Tehnologija koja će se koristiti prilikom realizacije problema je Python. Razlog odabira ovog programskog jezika je korišćenje mnogobrojnih biblioteka koje olakšavaju rad sa podacima.
Biblioteke koje će se koristiti su:

- Pandas – efikasne strukture podataka za rad sa podacima
- Matplotlib – koristi se za vizualizaciju podataka
- Sklearn – koristi se za mašinsko učenje

Okruženje u kojem će se pisati kod je Visual Studio Code.

Algoritmi

Neki od algoritama koji su korišćeni :

- RandomForestRegressor - koristi se za predviđanje kontinuiranih vrednosti
- LinearRegression - algoritam za regresiju koji se koristi za modeliranje linearnih odnosa između nezavisnih i zavisnih varijabli.

Cilj

Cilj projekta je analiza podataka i njihovo prikazivanje. Nakon analize podataka i identifikacije veza koje se pojavljuju, potrebno je kreirati grafički prikaz dobijenih rezultata. Aplikacija će posedovati korisnički interfejs (UI) koji će omogućiti korisniku da pronađe željene analize podataka i prikaže ih grafički. Potrebno je implementirati mogućnost da se na osnovu podataka o vozilu predvidi cena korišćenjem algoritma linearne regresije. Nakon identifikacije najefikasnijeg algoritma za predviđanje cene, potrebno je kreirati model i odrediti njegovu tačnost predikcije. Dodatno, potrebno je prikazati koji automobil pređe najviše kilometara svake godine pre nego što se proda.

Nikola Miljković E5-3/2023