**SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU**

**FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA**

**Procjena cijene rabljenog automobila**

**Seminarski rad**

**Raspoznavanje uzoraka i strojno učenje**

**Igor Delić**

**Osijek, 2023.**

Sadržaj

[1. UVOD 2](#_Toc127725952)

[2. NADZIRANO UČENJE 3](#_Toc127725953)

[2.1. Regresija 4](#_Toc127725954)

[3. ANALIZA PODATAKA 5](#_Toc127725955)

[4. REZULTATI 14](#_Toc127725956)

[4.1. Linearnu regresija 14](#_Toc127725957)

[4.2. Random Forest Regressor 15](#_Toc127725958)

[4.3. Stablo odlučivanja 16](#_Toc127725959)

[4.4. K-najbližih susjeda 17](#_Toc127725960)

[5. STREAMLIT APLIKACIJA 18](#_Toc127725961)

[6. ZAKLJUČAK 22](#_Toc127725962)

# 1. UVOD

Procjena cijene automobila važan je zadatak u automobilskoj industriji i tržištu rabljenih vozila. Zbog raznolikosti karakteristika i stanja automobila, procjena cijene može biti složen proces koji zahtijeva visoku razinu stručnosti i iskustva. Srećom, danas postoji značajna količina podataka o prodaji rabljenih vozila, što olakšava izgradnju modela koji mogu procijeniti vrijednost automobila na temelju njihovih karakteristika. Stoga se strojno učenje i umjetna inteligencija sve više koriste u ovom području kako bi se poboljšale preciznost i učinkovitost u procjeni cijena automobila.

U ovom seminaru koristit ćemo strojno učenje za procjenu cijene automobila. Proučit ćemo ključne značajke koje utječu na cijenu automobila i kako se te značajke koriste za izgradnju modela. Također ćemo pregledati različite algoritme strojnog učenja koji se koriste za regresijsku analizu i kako se primjenjuju na skupu podataka automobila.

# 2. NADZIRANO UČENJE

Nadzirano učenje je vrsta strojnog učenja u kojoj se model uči na temelju ulaza i pripadnih izlaza ili oznaka. Cilj je naučiti model koji može predviđati izlaz za nove ulaze koji mu nisu bili prethodno poznati. U nadziranom učenju, skup podataka koji se koristi za treniranje modela obično se sastoji od ulaznih vrijednosti i odgovarajućih oznaka ili izlaza, a model koristi te podatke za učenje. Na primjer, u zadatku klasifikacije slika, model bi se trenirao na skupu podataka slika i pripadnih oznaka koje ih klasificiraju u određene klase. Ključna prednost nadziranog učenja je da modeli mogu biti vrlo precizni i da se mogu koristiti za predviđanje novih vrijednosti koje nisu bile prisutne u skupu podataka za treniranje. Međutim, to također znači da ovisimo o kvaliteti i raznolikosti podataka za treniranje, te se problemi mogu javiti ako skup podataka nije dovoljno velik ili reprezentativan za svrhu učenja. Nadzirano učenje koristi se u mnogim područjima, uključujući računalni vid, obradu prirodnog jezika, prepoznavanje govora, medicinsku dijagnostiku, financije i mnoge druge industrije i grane znanosti.

## 2.1. Regresija

Regresija je statistička metoda koja se koristi za predviđanje numeričkih vrijednosti (kontinuiranih varijabli) između dvije ili više varijabli. Cilj regresijske analize je pronalaženje matematičke funkcije koja najbolje odgovara odnosu između nezavisne varijable (ulaza) i zavisne varijable (izlaza). Ova funkcija se zatim može koristiti za predviđanje izlaza na temelju novih ulaznih vrijednosti. Regresija se koristi u mnogim područjima kao što su ekonomija, poslovna analiza, medicina, inženjerstvo, znanost o podacima i strojno učenje. Primjeri regresije uključuju predviđanje cijene nekretnina, prognoziranje prodaje proizvoda, predviđanje potrošnje energije, predviđanje broja prodanih proizvoda i tako dalje. Postoji mnogo algoritama koji se koriste u regresiji, a neki od najčešćih su linearna regresija, regresija drveća odluke, k-NN regresija, random forest regresija i gradient boosting regresija. Odabir algoritma ovisi o prirodi podataka i vrsti problema koji se rješava.

# 3. ANALIZA PODATAKA

Podaci za strojno učenje su skup podataka koji se koriste za treniranje, testiranje i evaluaciju modela strojnog učenja. Podaci mogu biti strukturirani ili nestrukturirani. Strukturirani podaci su podaci koji se nalaze u tabličnom formatu i lako se mogu obrađivati ​​pomoću matematičkih funkcija. Ovi podaci se najčešće koriste u strojnom učenju za klasifikaciju i regresiju. Nestrukturirani podaci su podaci koji nemaju jasnu strukturu i teško se obrađuju.

Za potrebu izrade projektnog zadatka korišten je Dataset “car\_price\_prediction” spremljen kao CSV datoteka s atributima:

ID, Price, Levy, Manufacturer, Model, Prod. Year, Category, Leather interior, Fuel type, Engine volume, Mileage, Cylinders, Gear box type, Drive wheels, Doors, Wheel, Color i Airbags

Graphical user interface, text, application

Description automatically generated

Slika 3.1. Kod za učitavanje podataka

A picture containing table

Description automatically generated

Slika 3.2. Prikaz prvih 5 učitanih podataka

Table

Description automatically generated

Slika 3.3. Prikaz općih informacija

Porebno je odraditi provjeri nalazi li se u našim podacima neka null vrijednost i ako postoji, probrojati oliko ih ima. (Slika 3.4.)

Table

Description automatically generated

Slika 3.4.

Također je potrebno provjeriti nalaze li u našim podacima duplikati te ih je potrebno ukloniti.

Graphical user interface, text, application

Description automatically generated

Slika 3.5.

Table

Description automatically generated

Slika 3.6. Proizvođači automobila

Table

Description automatically generated

Slika 3.7. Kategorije automobila

Table

Description automatically generated

Slika 3.8. Boje automobila

Graphical user interface, text

Description automatically generated

Slika 3.9. Vrsta mjenjača i pogona

Kako bi pripremili podatke za daljnju obradu bilo potrebno je napraviti neke izmjene. Uklonjen je stupac 'ID' jer nam je ID automobila nepotreban za analizu. Zamijenjeni su nedostajući podaci u stupcu 'Levy' s NaN i konvertiran je u tip float, tako da se njime može lakše raditi. U stupcu 'Engine volume' obrisani su svi stringovi 'Turbo' iz podataka te je konvertiran u tip float. U stupcu 'Mileage' obrisani su svi stringovi 'km' iz podataka te je konvertiran u tip int. U stupcu 'Doors', stringovi su zamijenjeni odgovarajućim brojevima, kako bi se omogućila daljnja obrada u modelu strojnog učenja.(Slika 3.10.)

Graphical user interface, text, application

Description automatically generated

Slika 3.10. Prikaz opisanog koda

Slike prikazuju dio koda koji se koristi za identificiranje stupaca koji najviše utječu na cijenu automobila na temelju apsolutne vrijednosti korelacijskih koeficijenata. (Slika 3.11.)

Table

Description automatically generated

Slika 3.11.

Table

Description automatically generated

Slika 3.12.

Nadalje podaci su podijeljeni u kategorije temeljem tipa podataka u svakom stupcu. Varijabla x sadrži sve podatke iz skupa podataka, osim prvog stupca koji je ciljna varijabla, dok je y ta ciljna varijabla. Stvaraju se dvije liste X\_cats i X\_nums koje se popunjavaju stvarnim vrijednostima podataka iz skupa podataka X koji odgovaraju imenima stupaca u odgovarajućim listama. X\_cats sadrži sve kategoričke varijable, dok X\_nums sadrži sve numeričke varijable.(Slika 3.13.)

Text

Description automatically generated

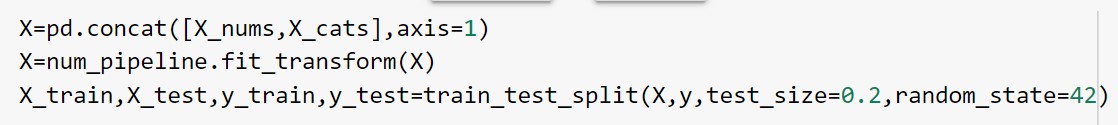
Slika 3.13.

Prethodno obrađeni numerički i kategorički podaci kombiniraju se kako bi se stvorila konačna matrica značajki X. Num\_pipeline je objekt Pipeline koji spaja dva transformatora: SimpleImputer i StandardScaler.(Slika 3.14.) Transformator SimpleImputer se koristi za zamjenu nedostajućih vrijednosti s medijanom respektivne značajke, a transformator StandardScaler koristi se za standardizaciju vrijednosti značajki oduzimanjem srednje vrijednosti i dijeljenjem s vrijednosti standardne devijacije. Nakon primjene num\_pipeline na numeričke značajke u X\_nums, rezultirajući niz se kombinira s one-hot enkodiranim kategoričkim značajkama u X\_cats pomoću funkcije concat. Zatim se cijela matrica značajki X transformira pomoću num\_pipeline. Konačno, funkcija train\_test\_split se koristi za podjelu podataka na skupove za učenje i testiranje, pri čemu se 20% podataka rezervira za testiranje. (Slika 3.15.)

Text

Description automatically generated

Slika 3.14.



Slika 3.15.

# 4. REZULTATI

U ovom poglavlju ćemo proći kroz rezultate učenja i vidjeti koji model je najuspješniji.

## 4.1. Linearnu regresija

Linearna regresija je jedna od najosnovnijih tehnika regresijskog strojnog učenja. Koristi se za modeliranje linearnih veza između ulaznih varijabli i izlazne varijable na temelju podataka za učenje. Cilj je pronaći linearnu funkciju koja najbolje opisuje odnos između ulaza i izlaza. Linearna regresija pokušava pronaći optimalne koeficijente koji minimiziraju kvadratnu razliku između predviđene vrijednosti i stvarne vrijednosti.

Text

Description automatically generated

Slika 4.1.

Prvo se model trenira na podacima za obuku pomoću LinearRegression() funkcije, a zatim se model primjenjuje na testne podatke i predviđanja se uspoređuju s stvarnim vrijednostima. Imamo tri različite metrike za evaluaciju performansi regresijskog modela: Mean Squared Error (MSE), Mean Absolute Error (MAE) i R-squared (R2) vrijednosti. MAE je prosječna apsolutna vrijednost razlike između stvarnih vrijednosti i predviđenih vrijednosti, a MSE je prosječna kvadratna razlika između stvarnih vrijednosti i predviđenih vrijednosti. R2 je vrijednost koja predstavlja omjer varijabilnosti u predviđenim vrijednostima i stvarnim vrijednostima. R2 vrijednost je u rasponu od 0 do 1, gdje 1 označava savršeno prilagođeni model. (Slika 4.1.)

## 4.2. Random Forest Regressor

Random Forest Regressor utemeljen je na ansamblu stabala odlučivanja (decision trees) za zadatke regresije. Stvara više stabala odlučivanja na temelju različitih podskupova podataka za treniranje. Svako stablo se trenira na podskupu podataka te svako stablo daje svoju predikciju, a krajnja predikcija je prosjek svih predikcija pojedinačnih stabala. Popularan je zbog svoje jednostavnosti i fleksibilnosti, te dobre sposobnosti generalizacije i izbjegavanja problema overfittinga. Također, može rukovati velikim brojem ulaznih značajki i podataka te se može koristiti za rješavanje različitih problema regresije.

Text

Description automatically generated

Slika 4.2.

Stvaramo instancu modela regresije slučajnih šuma, koji se zatim trenira na X\_train i y\_train podacima pomoću metode fit. Nakon treniranja, model se koristi za predviđanje vrijednosti na X\_test podacima pomoću metode predict, čije se vrijednosti uspoređuju s pravim vrijednostima y\_test pomoću nekoliko metrika. Mean Squared Error (MSE), Mean Absolute Error (MAE) i R-squared (R2) vrijednosti. (Slika 4.2.)

## 4.3. Stablo odlučivanja

Stabla odlučivanja se koriste za predviđanje kontinuirane vrijednosti. Temelji se na konstrukciji stabla odlučivanja gdje svaki unutarnji čvor predstavlja testiranje na određenoj značajci, dok su listovi predviđanja vrijednosti. Tijekom predviđanja, algoritam prolazi kroz stablo odlučivanja s pitanjima koja testiraju značajke i na kraju dolazi do jednog od listova koji predstavlja predviđanu vrijednost. Ovaj algoritam ima tendenciju overfittinga, što znači da se može prilagoditi podacima za učenje vrlo precizno, ali može imati lošiju točnost na novim podacima.

Text

Description automatically generated

Slika 4.3.

Na temelju zadanih podataka, gradi se stablo odlučivanja koje se sastoji od niza pitanja koja su postavljena kako bi se donijela konačna odluka. Stablo je izgrađeno na temelju atributa koji su najkorisniji u predviđanju tražene vrijednosti, a razgranavanje stabla ovisi o odgovorima na postavljena pitanja. Zatim, kada se stablo izgradi, podaci se mogu unijeti u stablo da bi se dobila predviđena vrijednost za traženu varijablu. Izvještaj prikazuje srednju kvadratnu pogrešku, apsolutnu pogrešku te koeficijent determinacije.(Slika 4.3.)

## 4.4. K-najbližih susjeda

KNeighborsRegressor je model regresije zasnovan na metodi k-najbližih susjeda (k-nearest neighbors). Ovaj model predviđa vrijednost za neku varijablu cilja (target variable) tako što traži k najbližih primjera (eng. instances) u skupu podataka i uzima prosječnu vrijednost ciljne varijable tih primjera kao predviđenu vrijednost za novi primjer. Metrika udaljenosti (eng. distance metric) se koristi za određivanje koje primjere uzeti u obzir kao k-najbliže primjere. U ovisnosti o odabranoj metrici i broju susjeda, model može biti osjetljiv na podatke i moguće imati tendenciju preprilagođavanja (eng. overfitting) na skup podataka za učenje. (Slika 4.4.)

Text

Description automatically generated

Slika 4.4.

# 5. STREAMLIT APLIKACIJA

Streamlit je Python biblioteka koja omogućuje jednostavno i brzo stvaranje web aplikacija za strojno učenje, vizualizaciju podataka i prototipiranje. Ova biblioteka nam omogućuje da u nekoliko linija koda izgradimo aplikaciju koja korisniku omogućuje interakciju s treniranim modelom. Kako bi smo koristili svoj trenirani model u streamlit aplikaciji, prvo moramo učitati model pomoću joblib biblioteke. Nakon toga, definiramo funkciju predict koja prima unos podataka i koristi prethodno učitani model za predviđanje izlaza za te podatke. Nakon što smo definirali funkciju, možemo koristiti streamlit komponente (npr. text\_input, button, selectbox itd.) kako bismo omogućili korisniku interakciju s aplikacijom. Napravljena je streamlit aplikacija koja predviđa cijenu automobila na temelju nekoliko značajki. Nakon učitavanja modela, korisniku prikazujemo ulazne okvire (input-e) u koje može upisati nekoliko značajki kao što su godina proizvodnje, prijeđeni kilometri, broj vrata i sl. Nakon što korisnik unese podatke, može kliknuti na gumb za predviđanje (button), a aplikacija će koristiti funkciju predict za predviđanje cijene automobila na temelju unešenih podataka.

Graphical user interface, application, email

Description automatically generated

Slika 5.1. Streamlit aplikacija

Table

Description automatically generated

Slika 5.2. Streamlit aplikacija

Graphical user interface, application

Description automatically generated

Slika 5.3. Streamlit aplikacija

Table

Description automatically generated with medium confidence

Slika 5.4. Kod streamlit aplikacije

A picture containing text

Description automatically generated

Slika 5.4. Kod streamlit aplikacije

# 6. ZAKLJUČAK

Na temelju navedenih mjera evaluacije, možemo zaključiti da model Random Forest Regressor ima najmanju vrijednost MSE i MAE, što znači da pokazuje najbolju točnost u predviđanju cijene automobila među ovim modelima. Također, R2 ocjena od 0,75 pokazuje da je model uspješan. Model linearne regresije ima najlošije rezultate u usporedbi s drugim modelima, s najvećim vrijednostima MSE i MAE te najnižom R2 ocjenom dok K-najbližih susjeda i stablo odlučivanja imaju slične vrijednosti.