Specifikacija projekta iz predmeta Računarska inteligencija

Student: Marija Živanović, SV19/2021

Naziv teme: Predikcija cena laptopa na osnovu karakteristika putem mašinskog učenja

Definicija problema: Ovaj projekat ima za cilj predviđanje cena laptopa putem mašinskog učenja. Problem se može formulisati kao regresioni problem, gde cilj predikcije jeste cena laptopa na osnovu različitih atributa odnosno karakteristika samog laptopa.

Motivacija: Predikcija cena laptop računara može imati široku praktičnu primenu u e-trgovini, prodajnim lancima, kao i u procenjivanju vrednosti laptopa u različitim uslovima tržišta. Ovo rešenje omogućava kupcima da dobiju realističnu predstavu o ceni proizvoda pre nego što ga kupe, a prodavcima pomaže u postavljanju konkurentnih cena.

Skup podataka: Skup podataka će biti izvorno preuzet sa jedne od najrelevantijih platformi za skupove podataka koji se mogu koristiti za izradu AI modela pod nazivom Kaggle. Link do dataseta: https://www.kaggle.com/datasets/muhammetvarl/laptop-price?resource=download. Broj instanci 1303.

Atributi skupa podataka: **laptop_ID:** jedinstveni identifikator laptopa; **Company**: proizvođač laptopa (npr. Dell, HP, Lenovo); **Product**: model laptopa; **TypeName**: tip laptopa (npr. Gaming, Ultrabook, Notebook); **Inches**: veličina ekrana laptopa u inčima; **ScreenResolution**: rezolucija ekrana; **Cpu**: specifikacija procesora; **Ram**: količina RAM memorije u laptopu; **Memory**: tip i kapacitet skladišta (npr. HDD, SSD, 256GB, 1TB); **Gpu**: grafička kartica; **OpSys**: operativni sistem koji se isporučuje sa laptopom; **Weight**: težina laptopa; **Price_euros**: cena laptopa u evrima (ciljno obeležje).

Najznačajniji atributi: Company, Product, TypeName, Inches, Cpu, Ram, Memory, Gpu, OpSys, Weight. Ovi atributi su od suštinskog značaja za predikciju cene laptopa. Informacije o proizvođaču, modelu, specifikacijama hardvera i softvera, kao i težini laptopa mogu značajno uticati na njegovu cenu.

Price_euros: Ovaj atribut predstavlja cenu laptopa izraženu u evrima i predstavlja **ciljno obeležje**, tj. vrednost koja se predviđa.

Pošto je cena laptopa kontinualna numerička vrednost, problem predstavlja regresioni problem. Opseg vrednosti za cene laptopa (“Price_euros”) će zavisiti od raspona cena koje su zastupljene u skupu podataka. Očekuje se da će opseg obuhvatiti širok spektar vrednosti kako bi model mogao efikasno da se generalizuje na različite cene.

Način pretprocesiranja podataka: Podaci će biti podvrgnuti standardnim postupcima pretprocesiranja kao što su uklanjanje duplikata, tretiranje nedostajućih vrednosti, normalizacija ili standardizacija numeričkih atributa, enkodiranje kategoričkih atributa, kao i podela skupa podataka na trening, validacioni i testni skup.

Metodologija:

Ulaz u model predstavljaju podaci o karakteristikama laptopa, dok je izlaz predviđena cena tog laptopa.

Koraci za rešavanje ovog problema će uključivati:

- Učitavanje podataka
- Pretprocesiranje podataka
- Podela podataka na trening, validacioni i testni skup
- Odabir modela

Različiti modeli se razmatraju i testiraju kako bi se odabrao najbolji za ovaj problem. Ja ću razmatrati modele linearne regresije, Decision Tree, Random Forest, Extra Trees.

U svakom modelu će se analizirati relevantnost svakog atributa na ciljno obeležje (putem kolrelacija, feature_importances kod decision tree modela, Recursive Feature Elimination- kod linearnih modela koje budem koristila kako bi se iterativno uklanjali najmanje bitni atributi i evaluirala promena performansi modela i drugo).

- Treniranje modela na trening skupu

Treniranje modela obuhvata podešavanje parametara i prilagođavanje modela trening podacima kako bi se minimizovala greška predikcije.

- Kod modela linearne regresije će se to postizati minimizacijom funkcije greške poput srednje kvadratne greške (Mean Squared Error) ili srednje apsolutne greške (Mean Absolute Error).
- Kod Decision i Extra Trees treniranje uključuje pronalaženje optimalnih granica za svaki čvor stabla kako bi se maksimizovala informacija ili smanjila entropija u svakom čvoru.
- Treniranje Random Forest modela uključuje kreiranje više stabala odlučivanja na slučajnim podskupovima trening podataka. Svako stablo se trenira nezavisno, a zatim se njihove predikcije kombinuju putem glasanja ili prosečnog vrednovanja kako bi se dobila konačna predikcija modela.
    - Evaluacija modela na validacionom skupu

Modeli se evaluiraju na validacionom skupu kako bi se uporedila njihova prediktivna moć i izabrali najbolji kandidati za dalje fine-tuning.

- Optimizacija (tjuning) hiperparametara

Parametri modela se podešavaju kako bi se postigla bolja predikcija (u mom projektu korišćenjem metoda poput GridSearchCV ili RandomizedSearchCV).

- Evaluacija konačnog modela na testnom skupu

Konačni model se evaluira na testnom skupu kako bi se procenila njegova sposobnost generalizacije na novim, neviđenim podacima.

Evaluacija rezultata: Rezultati će biti evaluirani na osnovu metrika kao što su srednja apsolutna greška (MAE), i koeficijent determinacije (R-squared).

Tehnologije: Za rešavanje ovog problema koristiću programski jezik Python i biblioteke kao što su scikit-learn, pandas, matplotlib, seaborn.

Relevantna literatura: Postoji nekoliko radova i resursa koji se bave sličnim temama kao što je predikcija cena proizvoda koristeći neuronske mreže. Prvo ću navesti neke od knjiga koje obuhvataju analiziranje tehnologija i modela koje ću koristiti:

- "Machine Learning Yearning" - Andrew Ng
- "Introduction to Machine Learning with Python" - Andreas C. Müller & Sarah Guido
- "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" - Aurélien Géron

Github i youtube linkovi na temu “Laptop Price Prediction With Python”:

- https://github.com/rajatrawal/laptop-price-predictor
- https://github.com/LuluW8071/Laptop-Price-Prediction
- https://www.youtube.com/watch?v=A1eU51jPpXQ&t=2141s
- https://www.youtube.com/watch?v=m1rY2J8ZlsY