

Završni projekt – Python kurs / Data & ML projekat

Opis zadatka

Zadatak svakog polaznika je da odabere dataset (poželjno sa Kaggle.com) i implementira cjelokupan proces obrade i analize podataka koristeći ETL pristup, SQL bazu, Python (Pandas) i osnovne algoritme mašinskog učenja. Projekat treba da bude autentičan, tehnički smislen i tematski koherentan.

Faze rada

1. Odabir i razumijevanje dataseta

- Dataset mora imati najmanje 500 redova i 5 kolona.
- Poželjno: stvarni podaci (npr. prodaja, zdravstvo, obrazovanje, sport...)
- Analizirati domenu, objasniti osnovni kontekst podataka (u README).

2. ETL proces

- E: Preuzeti i učitati podatke (CSV/Excel/API).
- T: Očistiti i transformisati podatke – ukloniti duplikate, raditi encoding, handle missing values, formatirati datume itd.
- L: Ubaciti transformisane podatke u SQL bazu

3. Rad sa SQL bazom

- Odabrani dataset ubaciti u SQL bazu
- Napraviti najmanje 5 SQL upita koji imaju poslovni smisao (npr. prosjek, suma, filtriranje, join).
- Upite integrisati u Python skriptu i rezultat prikazati preko Pandas dataframe-a.

4. Exploratory Data Analysis (EDA)

- Koristiti Pandas, Matplotlib, Seaborn (ili Plotly).
- Prikazati:
 - Opisne statistike (describe, info, value_counts, itd)
 - Korelacije (corr() + heatmap)
 - Histogrami, scatter plot, bar chart, boxplot (min. 6 grafa)
- Napraviti kratak komentar u vezi uočenih obrazaca i potencijalnih outliera.

5. Primjena algoritma mašinskog učenja

- Na osnovu tipa problema (npr. predikcija cijena, klasifikacija bolesti, grupisanje korisnika...), odabrati jedan od sljedećih algoritama:
 - Regresija: LinearRegression, Ridge, Lasso...
 - Klasifikacija: LogisticRegression, DecisionTree, RandomForest...
 - Clustering: KMeans, DBSCAN...
- Podijeliti podatke na trening i test set.
- Prikazati metrike performansi (npr. RMSE, R^2 , accuracy, precision, silhouette score...).
- Vizualizirati rezultate gdje je moguće.

Obavezni elementi

Element	Obavezno
Autentičnost rada	✓
Jupyter Notebook ili .py skripta sa logičnim komentarima	✓
SQL baza	✓
Pandas analiza + minimalno 6 grafičkih prikaza	✓
EDA i poslovni uvidi u tekstualnom obliku	✓
Primjena jednog ML algoritma i evaluacija modela	✓
README fajl sa uputstvom za pokretanje projekta	✓
Kod mora biti objektno orijentisan gdje je primjenjivo	✓

Preporučena dinamika

- 1. sedmica: Odabir dataseta + analiza
- 2. sedmica: ETL + SQL baza
- 3. sedmica: EDA + grafovi
- 4. sedmica: ML model + testiranje

Kriteriji za ocenjivanje

Stavka	Bodovi
Ispravno urađen ETL + SQL integracija	20
Kvalitetna EDA i vizualizacije	20
Ispravno primijenjen ML model + metrika	25
Struktura koda, čitljivost, komentari	10
Dokumentacija projekta (README, objašnjenja)	10
Prezentacija i odgovaranje na pitanja	15