# ANALIZA VELIKIH PODATAKA

## školska 2024/2025 godina

## Vežba 1: Uvod u predmet i instalacija neophodnog softvera

**Tema 1: Rad sa numeričkim podacima pomoću NumPy biblioteke**

Uvod u NumPy – kreiranje nizova (arrays), matematičke operacije, indeksiranje i reshaping. Prikaz razlike između običnih Python lista i NumPy nizova. Uvešćemo i rad sa manjim datasetovima kako bismo pripremili teren za rad sa realnim podacima sa Kaggle platforme u narednim vežbama.

**Tema 2: Rad sa tabelarnim podacima pomoću Pandas biblioteke**

Učitavanje i analiza podataka pomoću pandas – DataFrame i Series strukture. Uvoz CSV fajlova, prikaz prvih redova, filtriranje i sortiranje podataka. Studenti će početi da rade sa realnim podacima koji se preuzimaju sa Kaggle, uz praktične primere iz e-trgovine, recenzija, filmova itd.

**Tema 3: Čišćenje i priprema podataka za analizu**

Identifikovanje i uklanjanje nedostajućih vrednosti, duplikata, kao i konverzija tipova podataka. Rad sa vremenskim kolonama i transformacija tekstualnih podataka u brojeve. Objasnićemo važnost kvalitetne pripreme podataka za dalje analize i modeliranje.

**Tema 4: Vizualizacija podataka pomoću Matplotlib i Seaborn**

Pravljenje osnovnih grafika – linijski dijagrami, stubičasti grafikoni, histogrami i scatter plotovi. Kombinovanje Matplotlib i Seaborn biblioteka za efikasno prikazivanje veza između varijabli. Studenti će vizualno analizirati stvarne podatke i otkrivati obrasce ponašanja.

**Tema 5: Grupisanje i agregacija podataka u Pandas-u**

Rad sa groupby() metodom i pivot tabelama za kategorizaciju i agregaciju. Analiza vremenskih trendova i sezonskih obrazaca u podacima. Studentima će biti pokazano kako da efikasno analiziraju velike skupove podataka po grupama (po zemlji, kategoriji ili vremenu).

**Tema 6: Klaster analiza i uvod u ne-nadzirano učenje**

Uvod u ne-nadzirano učenje kroz algoritam K-Means. Normalizacija podataka i vizualizacija klastera pomoću scatter plot-ova. Objasnićemo kako algoritmi prepoznaju sličnosti između podataka bez prethodnog označavanja. Takođe, studenti će naučiti kako da interpretiraju rezultate klaster analize i kako da koriste ovu tehniku za otkrivanje skrivenih obrazaca u podacima.

**Tema 7: Linearna regresija i analiza zavisnosti između varijabli**

Objašnjenje koncepta linearne regresije i predikcija numeričkih vrednosti. Analiza metrika uspešnosti modela (MAE, MSE, $R^2$). Studenti će primeniti model na dataset iz stvarnog sveta i naučiti kako se gradi prediktivni model korak po korak. Takođe, biće prikazane tehnike za interpretaciju koeficijenata modela i kako da donesu zaključke na osnovu rezultata regresije.

**Tema 8: Logistička regresija i binarna klasifikacija**

Uvod u logističku regresiju kao osnovni klasifikacioni model. Studenti će raditi na binarnim klasifikacionim zadacima – npr. predikcija da li će korisnik otkazati uslugu, da li je komentar pozitivan/negativan i slično. Pokazujemo kako model funkcioniše i kako se ocenjuje (accuracy, confusion matrix, ROC AUC).

**Tema 9: Višestruka regresija i značaj osobina (feature importance)**

Primena više ulaznih karakteristika za poboljšanje preciznosti predikcije. Vizuelizacija uticaja pojedinačnih osobina na izlaznu promenljivu. Uvešćemo i koncept one-hot enkodiranja i skaliranja ulaznih podataka.

**Tema 10: Analiza teksta i osnove NLP**

Uvod u obradu teksta (NLP) – čišćenje, tokenizacija, uklanjanje stop-reči i osnovni rad sa tekstualnim kolonama. Primena VADER alata za analizu sentimenta recenzija. Rad sa tekstualnim datasetovima preuzetim sa Kaggle-a ili otvorenih izvora.

**Tema 11: Random Forest**

Upoznavanje sa konceptom Random Forest modela i njegovom primenom u klasifikaciji. Objašnjenje rada algoritma kroz drveće odlučivanja i slučajni izbor osobina. Studenti treniraju model na binarnim klasifikacionim problemima i analiziraju metrike uspešnosti (confusion matrix, precision, recall, F1-score). Poseban akcenat se stavlja na feature importance i tumačenje modela.

**Tema 12: XGBoost**

Uvod u XGBoost – efikasni gradijentni boosting model. Prikaz razlike u odnosu na Random Forest. Studenti će optimizovati hiperparametre i upoređivati performanse modela na istom datasetu. Posebno ćemo obraditi praktične izazove poput overfitting-a i vremena treniranja na velikim skupovima podataka.

**Tema 13: Uvod u neuronske mreže**

Objašnjenje osnovne arhitekture veštačkih neuronskih mreža (Multi-layer Perceptron). Uvod u Keras i TensorFlow, pravljenje jednostavnog modela za binarnu klasifikaciju. Prikaz kako mreže uče i kako se podešavaju parametri kroz epohe. Studenti će analizirati rezultate i naučiti osnovne pojmove kao što su aktivacione funkcije, epohe, batch size i loss funkcija.

Sajt za skupove podataka: https://www.kaggle.com/datasets

Okruženje gde ćemo raditi: https://colab.research.google.com/ Za o'line rad,

Visual Studio Code sa Python dodatkom ili PyCharm

https://www.python.org/downloads/ i https://www.jetbrains.com/pycharm/

Načini polaganja:

- 2 kolokvijuma po 40 poena + lab vežbe i prisustvo 20 poena
- Pismeni ispit 80 poena + lab vežbe i prisustvo 20 poena
- Projekat 80 poena + lab vežbe i prisustvo 20 poena

Korisni linkovi:

https://www.w3schools.com/datascience/ds_python.asp

https://www.youtube.com/watch?v=wUSDVGivd-8 https://www.coursera.org/learn/python-for-applied-data-science-ai