**PREDLOG PROJEKTA**

*Predikcija konačne ocene i verovatnoće da će student upisati visokoškolsko obrazovanje*

**OPIS PROBLEMA:**

Cilj projekta je predviđanje konačne ocene studenata na osnovu različitih parametara, kao i odluke da li će student upisati visokoškolsko obrazovanje, omogućiće razumevanje faktora koji utiču na akademski uspeh studenata.

**SKUP PODATAKA:**

Koristićemo .csv datoteku sa Kaggle platforme (link: [High School Student Performance and Demographics](https://www.kaggle.com/datasets/dillonmyrick/high-school-student-performance-and-demographics/data)). Skup podataka sadrži informacije o studentima, uključujući demografske podatke, rad porodice, podršku škole, vreme posvećeno učenju, zdravstveno stanje, konačne ocene itd… Ovaj skup podataka baziran je na dve srednje škole iz Portugala i specifično predstavlja učenike koji pohađaju časove portugalskog jezika.

**SPECIFIKACIJA ALGORITMA:**

Koristićemo multivarijabilnu regresiju s različitim regularizacionim normama za predviđanje konačne ocene studenata i verovatnoće upisa na visokoškolsko obrazovanje.

Razmatraćemo:

* L1 normu (lasso),
* L2 normu (ridge)
* Elastic net (l1 + l2)

kako bismo sprečili preprilagođavanje (overfitting) modela.

Za smanjenje broja karakteristika, primenićemo analizu korelacija između atributa, posebno uzimajući u obzir bitnu napomenu da ciljni atribut final\_grade ima snažnu korelaciju sa atributima grade\_2 i grade\_1. Ova korelacija proizlazi iz toga što je final\_grade ocena za završnu godinu (izdata u trećem periodu), dok grade\_1 i grade\_2 odgovaraju ocenama u prva dva perioda. Iako je izazovnije predviđati final\_grade bez grade\_2 i grade\_1, ove predikcije će biti od suštinskog značaja.

Dodatno, kao probna metoda, koristićemo Principal Component Analysis (PCA) kako bismo stvorili nove karakteristike na osnovu karakterističnih vrednosti korelacione matrice za dalju regresiju.

Takođe, razmotrićemo upotrebu naprednijih metoda iz biblioteke Sklearn (npr. Support Vector Regression i Gradient Boosting).

**FAZE PROJEKTA**

1. Prikupljanje i Analiza Podataka:

* Učitavanje i istraživanje skupa podataka o studentima.
* Vizualizacija raspodele ocena, demografskih podataka i drugih relevantnih informacija.

1. Priprema Podataka:

* Obrada nedostajućih vrednosti.
* Kodiranje kategoričkih podataka.
* Normalizacija numeričkih podataka.

1. Modeliranje:

* Podela skupa podataka na trening, validacioni i test set.(70% za trening, 15% za validaciju i 15% za testiranje.)
* Implementacija multivarijabilne regresije sa L1 normom (lasso), L2 normom (ridge) i Elastic net (l1 + l2).
* Korelaciona analiza i primena Principal Component Analysis za smanjenje dimenzionalnosti.

1. Treniranje i Evaluacija Modela:

* Treniranje modela na trening setu.
* Optimizacija parametara korišćenjem validacionog seta.
* Evaluacija performansi modela na test setu.

1. Analiza i Interpretacija Rezultata:

* Analiza važnosti obeležja dobijenih iz modela.
* Interpretacija uticaja različitih faktora na konačne ocene i odluku o upisu na visokoškolsko obrazovanje.

**SOFTVER I PROGRAMSKI JEZIK:**

Jupyter Notebook/IPython

Python sa bibliotekama NumPy, Matplotlib, Pandas i Scikit-Learn.

**LITERATURA I REFERENCE:**

<https://www.kaggle.com/datasets/dillonmyrick/high-school-student-performance-and-demographics>

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance.