

Izveštaj o Analizi Podataka i Modeliranju

1. Pregled Skupa Podataka

Skup podataka sadrži 400 redova sa 9 karakteristika vezanih za predikciju prijema na postdiplomske studije:

- **Karakteristike:** GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research, Chance of Admit.
- Svi podaci su potpuni (nema nedostajućih vrednosti).

Ključna statistika:

- **GRE Score** se kreće između 290 i 340, sa prosekom od ~316,81.
- **TOEFL Score** se kreće od 92 do 120, sa prosekom od ~107,41.
- **Chance of Admit** varira od 0,34 do 0,97, sa prosekom od ~0,724.

2. Istraživačka Analiza Podataka (EDA)

- **Histogrami:** Vizualizacija distribucija svih numeričkih kolona radi identifikacije obrazaca i outliera.
 - Primer: GRE Scores pokazuje približno normalnu distribuciju sa blagim pomakom ka višim vrednostima.
- **Scatter Plots:** Grafički prikaz svake karakteristike u odnosu na indekse radi otkrivanja trendova.
- **Box Plotovi:** Identifikacija outliera u svakoj karakteristici.
 - Primer: CGPA i LOR imaju minimalne outliere, što ukazuje na kvalitet podataka.

3. Obrada Podataka

- Uklonjena kolona "Serial No." kao irelevantna za analizu.
- Sve numeričke kolone su normalizovane korišćenjem **StandardScaler** radi boljih performansi modela.

4. Linearni Regrsioni Model

- Skup podataka podeljen je na trening i test skupove (80%/20%).
- Trenirali smo **Linear Regression** model za predviđanje "Chance of Admit" na osnovu ostalih karakteristika.

Rezultati:

- **Srednja apsolutna greška (MAE):** ~0,048
- **Srednja kvadratna greška (MSE):** ~0,0046
- **R-kvadrat:** ~0,821, što pokazuje da model objašnjava 82,1% varijanse u šansama za prijem.

5. K-Fold Kros Validacija

- Primijenjeno je **K-Fold Cross-Validation** sa 5 podela za procenu generalizacije modela.
- Prosečne metrike:
 - MAE: ~0,046
 - MSE: ~0,0042
 - R-kvadrat: ~0,789

6. Vizualizacija Rezultata

- Upoređene su stvarne i predviđene vrednosti pomoću scatter plot-a. Tačke su blizu idealne dijagonalne linije, što ukazuje na dobre predikcije.

7. Ključni Uvidi i Zapažanja

- **GRE Score i CGPA** imaju najveći uticaj na šanse za prijem prema analizama.
- Model je stabilan kroz sve iteracije K-Fold validacije, što potvrđuje njegovu pouzdanost.

8. Zaključak

Analiza i modeliranje uspešno pružaju okvir za razumevanje i predikciju šansi za prijem na studije. Sa R-kvadratom od ~82%, model pokazuje solidne prediktivne sposobnosti.