Izveštaj o Analizi Podataka i Modeliranju

1. Pregled Skupa Podataka

Skup podataka koji je korišćen za izgradnju i evaluaciju modela sastoji se od dva glavna dela: trening skupa i test skupa. Skup podataka sadrži informacije vezane za predikciju ciljne promenljive (y), koja ima dve klase: "yes" i "no". Skup podataka za trening sadrži sledeće karakteristike:

- **Karakteristike**: Numeričke i kategorijske vrednosti, kao što su GPA, testni rezultati, i druge relevantne varijable koje mogu uticati na predikciju.
- **Ciljna promenljiva (y)**: Dve klase, gde je "yes" pozitivna klasa, a "no" negativna klasa.
- Skup podataka je u potpunosti očišćen i pripremljen za modeliranje, sa svim nedostajućim vrednostima zamenjenim odgovarajućim vrednostima.

2. Istraživačka Analiza Podataka (EDA)

- **Distribucija podataka**: Korišćeni su histogrami za analizu raspodele numeričkih karakteristika, što je omogućilo identifikaciju obrazaca i potencijalnih outliera u podacima.
    - **GPA** i **testni rezultati** imaju otprilike normalnu distribuciju, dok su druge karakteristike bile značajno asimetrične.
- **Scatter Plots**: Grafički prikazi su korišćeni za vizualizaciju odnosa između varijabli, pomažući da se identifikuju potencijalni korelativni trendovi.
- **Box Plotovi**: Korišćeni su za identifikaciju outliera, posebno u numeričkim kolonama. Kroz ovu analizu, outlieri nisu bili značajni za model, ali su se koristili za dalju analizu i čišćenje podataka.

3. Obrada Podataka

- **Kodiranje kategorijskih promenljivih**: Kategorijske varijable su kodirane pomoću `LabelEncoder`-a. Ovaj korak je bio neophodan da bi se kategorijske vrednosti prevele u numeričke vrednosti koje model može koristiti.
- **Normalizacija numeričkih podataka**: Sve numeričke karakteristike su normalizovane korišćenjem `StandardScaler`-a kako bi se osigurala ujednačenost u opsegu vrednosti i poboljšale performanse modela.

4. Naivni Bajesov Model

- **Modeliranje**: Naivni Bajesov model (`GaussianNB`) je korišćen za klasifikaciju. Model je treniran na trening skupu podataka.
- **Rezultati**:
  - **Tačnost**: 82,48% što ukazuje na to da model dobro predviđa većinu slučajeva.
  - **Izveštaj o klasifikaciji**:
    - **Preciznost za klasu "no"**: 0.93, što znači da je model vrlo precizan u predviđanju negativnih slučajeva.
    - **Recall za klasu "no"**: 0.87, što pokazuje da model otkriva veliki broj stvarnih "no" slučajeva.
    - **F1-score za klasu "yes"**: 0.40, što ukazuje na slabiji učinak modela za predikciju pozitivne klase.
  - **Confusion Matrix**: Ukazuje na veliki broj tačno predviđenih "no" slučajeva, ali vrlo malo "yes" slučajeva.

5. ROC-AUC Analiza

- **ROC-AUC**: Površina ispod ROC krive je 0.8094, što ukazuje na dobru sposobnost modela da razlikuje klase. Međutim, niska vrednost za klasu "yes" pokazuje da model nije dobro naučio predikciju manje zastupljene klase.
- **ROC Kriva**: Kriva ROC prikazuje odnos između tačno predviđenih pozitivnih i negativnih slučajeva za različite pragove klasifikacije.

6. SVM Model

- **Modeliranje**: Trenirali smo SVM model sa RBF kernelom koristeći najbolji set parametara.
- **Rezultati**:
  - **Tačnost**: 88% koja je visoka, ali maskira problem sa slabim predviđanjem klase "yes".
  - **Izveštaj o klasifikaciji**:
    - **Preciznost za klasu "no"**: 0.88, što pokazuje dobar učinak za predikciju klase "no".
    - **Preciznost za klasu "yes"**: 0.00, što ukazuje na potpunu nemogućnost modela da prepozna pozitivne slučajeve.
  - **Confusion Matrix**: Pokazuje da je model tačno predvideo većinu negativnih slučajeva, ali nije predvideo nijedan slučaj iz klase "yes".

7. Ključni Uvidi i Zapažanja

- **Klasa "no" dominira** u predikcijama, što ukazuje na ozbiljan problem sa nebalansiranim podacima.
- Naivni Bajesov model je uspeo da predvidi "no" slučajeve sa visokom tačnošću, ali nije imao dobar učinak za predikciju klase "yes".
- SVM model je takođe imao dobar učinak za klasu "no", ali nije bio u mogućnosti da prepozna klasu "yes", što je veliki nedostatak.
- **Class imbalance** je glavni razlog za ove slabosti u modelima, jer su podaci znatno neregularni sa više slučajeva klase "no" nego klase "yes".

8. Zaključak

Modeli kao što su Naivni Bajes i SVM dali su dobre rezultate za većinu slučajeva, ali nisu bili efikasni u predikciji manje zastupljenih pozitivnih slučajeva. Da bi se poboljšali rezultati, neophodno je raditi na balansiranju podataka, optimizaciji hiperparametara modela, i potencijalno primeni naprednijih tehnika poput SMOTE-a za generisanje novih uzoraka za klasu "yes".