ANALIZA VELIKIH PODATAKA

školska 2024/2025 godina

Instrukcije za ispitni projekat

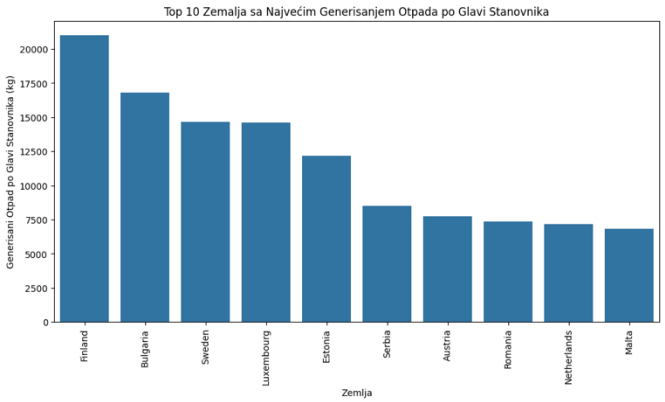
Opšti zahtevi za izradu projekta

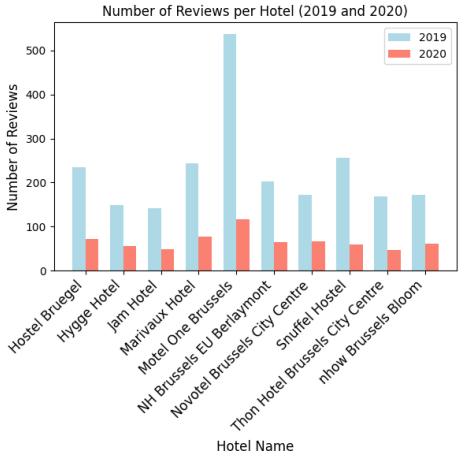
Svi projekti se izrađuju u **Google Colab-u** i neophodno je da imaju **kombinaciju teorijskog objašnjenja i praktične implementacije**. Sav kod se radi u Pythonu i treba da ima **detaljne komentare**, a sve sekcije rada moraju biti jasno odvojene i obeležene. Rad treba da bude razumljiv kao tehnički izveštaj, ali i kao vodič za drugu osobu koja prvi put čita i pokreće kod. Na dan pismenog dela ispita prezentujete svoj Colab, a pre toga možete poslati link do istog.

Obavezne sekcije rada:

1. **Uvod**
    - Predstaviti temu rada i njen značaj u savremenom kontekstu.
    - Objasniti čemu služi model ili analiza koju radimo.
    - Navesti koji su potencijalni benefiti primene ovog modela.
2. **Istraživački deo (Literatura / Previous Findings)**
    - Navesti ranija istraživanja, primere iz industrije, poznate pristupe rešavanju sličnih problema.
    - Kratko uporediti šta se radilo ranije i šta će vaš rad obuhvatiti.
    - Obavezno pravilno označiti reference i u tekstu i na samom kraju rada u posebnoj sekciji (posle zaključka). Ispratite **APA style.**
3. **Metodologija**
    - Objasniti kako su podaci prikupljeni ili kreirani.
    - Odraditi obavezno i EDA **(Exploratory Data Analysis),** proći kroz redove i kolone, grupisati adekvatne pojmove, prikazati vizualno kakvi su to podaci.

- Na primer, EDA može ovakve informacije da sadrži, čisto da imamo uvid o raspodeli na kakvim podacima treba da se implementira ML pristup.





- Zatim, predstaviti kako su podaci očišćeni (npr. uklanjanje null vrednosti, obrada teksta, transformacije).
- Kako su pripremljeni za model (skaliranje, encoding, vektorizacija).
- Koji **ML model** je korišćen i kako je treniran.
- Da li je i kako model dodatno optimizovan (npr. hyperparameter tuning).

4. **Rezultati i Diskusija**
   - Prikaz i komentarisanje metrika uspešnosti (accuracy, precision, recall, MAE, F1-Score, ROC-AUC curve, itd).
   - Vizualizacija rezultata kada je primenljivo.
   - Diskusija o tome šta rezultati znače – zašto je model bio uspešan ili nije, koje su prepreke. **Napomena: Potrudite se preko 85% accuracy da ostvarite.**
   - Koliko puta je model treniran, šta se menjalo, kako se ponašao.
5. **Zaključak**
   - Sumirati rad.
   - Dati preporuke za budući rad (dodatni podaci, bolji modeli, unapređenja).
   - Objasniti kako model može imati praktičnu primenu u realnom svetu (industrija, društvo, ekonomija...).

Lista tema

Za svaku temu možete da uzmete dataset sa Kaggle platofme koji se najviše uklapa u opis, ili da pronađete na nekom drugom sajtu adekvatne skupove podataka. Bitno je da su otvoreni i javno dostupni, da ne uzimate neke privatne podatke, ili ako sami sakupljate da imate bar nekoliko hiljada redova. Što se mene tiče možete slobodno odraditi i neki svoj web scraping.

---

1. Predikcija uspeha restorana na osnovu online ocena i recenzija

- **Kolone koje treba dataset da sadrži:** restaurant_name, location, review_text, rating, price_range, cuisine_type, review_date
- **Tehnike:** NLP obrada teksta, regresija, sentiment analiza, vizualizacija trendova
- **Cilj:** Predikcija ocene restorana na osnovu sadržaja recenzije i lokacije

---

2. Otkrivanje anomalija u finansijskim transakcijama

- **Kolone:** transaction_id, user_id, amount, timestamp, location, device_type, is_fraud
- **Tehnike:** Detekcija anomalija, klasifikacija, analiza vremena i lokacije
- **Cilj:** Identifikacija sumnjivih transakcija

---

3. Predikcija otkaza zaposlenih u firmi

- **Kolone:** employee_id, department, salary, tenure, promotion_last_5yrs, left
- **Tehnike:** Klasifikacija (Random Forest, Logistic Regression), analiza značaja faktora, korelacija
- **Cilj:** Predviđanje da li će zaposleni napustiti firmu

---

4. Predikcija cene nekretnina na osnovu karakteristika

- **Kolone:** location, area, number_of_rooms, year_built, price
- **Tehnike:** Višestruka regresija, skaliranje podataka, vizualizacija geografskih obrazaca
- **Cilj:** Precizna procena tržišne vrednosti

5. Analiza hotelskih recenzija i vremenskih obrazaca ponašanja korisnika

- **Kolone:** hotel_name, location, review_date, review_text, rating, reviewer_nationality
- **Tehnike:** Grupisanje po vremenskim periodima, analiza sentimenta, korelacije
- **Cilj:** Prikaz sezonskih trendova i obrazaca zadovoljstva gostiju

---

6. Analiza uspešnosti proizvoda na osnovu korisničkih ocena

- **Kolone:** product_id, category, rating, review_text, review_length, return_rate
- **Tehnike:** NLP, klasifikacija, regresija, vizualizacija ocena
- **Cilj:** Uočavanje koji proizvodi imaju visoku ocenu i nisku stopu vraćanja

---

7. Predikcija bolesti na osnovu simptoma (medicinski skup)

- **Kolone:** patient_id, age, gender, symptoms, diagnosis
- **Tehnike:** NLP na simptomima, klasifikacija (Decision Tree, Naive Bayes), balansiranje podataka
- **Cilj:** Predikcija potencijalne dijagnoze na osnovu simptoma

---

8. Detekcija lažnih recenzija na sajtovima za online kupovinu

- **Kolone:** user_id, product_id, review_text, review_date, review_length, is_fake
- **Tehnike:** NLP + klasifikacija, otkrivanje šablona, korelacija sa brojem recenzija istog korisnika
- **Cilj:** Uklanjanje sumnjivih recenzija sa platformi

---

9. Predikcija ponašanja korisnika na e-commerce platformi (npr. Temu ili AliExpress)

- **Kolone:** user_id, product_id, category, price, time_spent_on_page, discount_applied, purchase_made
- **Tehnike:** Višestruka regresija, klasifikacija, analiza korisničkih navika
- **Cilj:** Predikcija verovatnoće kupovine

10. Predikcija uspeha oglasa na društvenim mrežama

- **Kolone:** ad_id, ad_text, platform, audience_size, clicks, conversions
- **Tehnike:** NLP, regresija, analiza performansi kampanja
- **Cilj:** Predikcija koliko će oglas biti uspešan (CTR, ROI)

---

11. Detekcija govora mržnje u komentarima na društvenim mrežama

- **Kolone:** comment_id, comment_text, user_age, user_location, label (0/1), date_posted
- **Tehnike:** NLP obrada (lemmatizacija, vektorizacija), klasifikacija (Random Forest), evaluacija F1
- **Cilj:** Automatsko filtriranje govora mržnje i neprimerenih sadržaja

---

12. Predikcija prolaznosti ispita na osnovu aktivnosti učenja

- **Kolone:** student_id, hours_studied, assignments_completed, attendance, final_score
- **Tehnike:** Regresionu modeli, analiza uticaja faktora, vizualizacija
- **Cilj:** Pomoć u proceni rizika neuspeha

---

13. Analiza saobraćajnih nesreća i predikcija rizika po lokaciji

- **Kolone:** accident_id, location, time, weather, vehicle_type, casualties, is_fatal
- **Tehnike:** Analiza prostora i vremena, klasifikacija, mapa rizika
- **Cilj:** Otkrivanje rizičnih oblasti za unapređenje bezbednosti

---

14. Predikcija trajanja filma u bioskopima na osnovu početnih ocena i recenzija

- **Kolone:** movie_id, genre, opening_rating, review_volume, weeks_in_theaters
- **Tehnike:** Regresija, analiza korelacija, vizualizacija popularnosti
- **Cilj:** Predikcija koliko će film biti prisutan u bioskopima