Prvi deo projektnog zadatka – analiza baze podataka

30000 Spotify Songs

Informacije o bazi podataka

- **Korišćena baza podataka:** 30000 Spotify Songs
- **Link do baze podataka:** https://www.kaggle.com/datasets/joebeachcapital/30000-spotify-songs?select=spotify_songs.csv

Tim studenata

- IN 34/2020 – Nikola Kerleta
- IN 47/2020 – Srđan Petrović

Analiza

1. **Definisati u 2-3 rečenice problem koji će se u projektu rešavati. (primer: Rešavaće se problem detekcije karcinoma dojke na osnovu analiza iz krvi. U pitanju je klasifikacioni problem sa 2 klase.)**

   - Ovaj projekat će se baviti problemom praćenja trendova u muzici. Svodi se na predikciju popularnosti neke pesme po određenim karakteristikama koje je opisuju.

2. **Koliko ima uzoraka u bazi?**

   - Ima ukupno 32833 uzoraka.

3. **Jednom rečenicom objasniti šta predstavlja jedan uzorak u konkretnoj bazi.**

   - Jedan uzorak u bazi pretstavlja opšte informacije o pesmi dobijene sa aplikacije *"Spotify",* njene karakteristike i informacije o njnoj uspešnosti.

4. **Koliko ima obeležja u bazi?**

   - Ima ukupno 23 obeležja.

5. **Navesti sva obeležja (jasnim imenom na srpskom ili opisno, nebitan je naziv u samoj bazi).**

   - Id pesme (type:string)
   - Ime pesme (type:string)
   - Izvođač (type:string)
   - Mera popularnosti date pesme u tom momentu (type:int)
   - Dužina pesme data u ms (type:int)
   - Ime playliste iz koje je povučena pesma (type:string)
   - Id playliste iz koje je povučena pesma (type:string)
   - Žanr playliste (type:string)
   - Podžanr playliste (type:string)
   - Id albuma date pesme (type:string)
   - Ime albuma (type:string)
   - Datum kada je album objavljen (type:string)
   - Mera koliko je pesma ritmična (type:float)
   - Mera koliko je pesma energična (type:float)
   - Tonalitet pesme (type:int)
   - Glasnoća pesme (type:float)
   - Broj koji predstavlja da li je pesma u duru ili molu (type:int)
   - Mera koliko pesma zvuči kao priča (type:float)
   - Mera kvaliteta akustike u pesmi (type:float)
   - Koliko ima čistog instrumentala u pesmi (type:float)
   - Mera prisustva publike u pesmi (type:float)
   - Mera toga koliko je pesma pozitivna (type:float)
   - Tempo pesme dat u BPM (type:float)

6. **Koliko ima numeričkih obeležja?**

   - Ima ukupno 9 numeričkih obeležja.

7. **Ako ima kategoričkih obeležja, navesti koje od njih ima najmanji broj kategorija i koje su, i navesti ono koje ima najveći broj kategorija i koliko ih je**.

   - Najmanji broj kategorija ima kategoričko obeležje mode koje nam govori da li je u duru ili u molu (0 ili 1). A najveći broj kategorija ima obeležje track_name koje ima 23449 kategorija.

8. **Ako se rešava regresioni problem: navesti opseg, sr.vr. i medijanu obeležja koje će se predvidjati. Ako se rešava klasifikacioni problem: navesti procentualno koliko ima uzoraka u svakoj od klasa.**

   - Minimalna vrednost: 0.00
   - Maksimalna vrednost: 100.00
   - Srednja vrednost: 42.477081
   - Medijana: 45.00

9. **Da li postoje obeležja u bazi koja smatraš da treba izbaciti iz baze? Koja su to i zašto smatraš da ih treba izbaciti?**

   - track_id (Id pesme) - Nije nam potreban za bilo kakvu analizu i unikatan je za svaki primerak
   - track_name (Ime pesme) - Ima previše unikatnih vrednosti pa nam nije od koristi u daljoj analizi
   - track_artist (Izvođač pesme) - Bio bi preveliki bias prema određenim izvođačima. To je problem jer - želimo da predviđamo uspeh i nepoznatih izvođača.
   - track_album_id i track_album_name (Id albuma i ime albuma) tokođe nisu objektivni pokazatelji popularnosti neke pesme.
   - playlist_name i playlist_id (Ime playliste i id playliste iz koje je uzeta pesma) - nisu nam potrebni za dalju analizu.

10. **Da li u bazi ima nedostajućih vrednosti? Ako ima, navesti za svako od obeležja koliko vrednosti mu procentualno nedostaje?**

    - U bazi postoje nedostajuće vrednosti za ukupno 3 obeležja i to *track_name*, *track_artist)* i *track_album_name*.
      - track_name: 0.000152%
      - track_artist: 0.000152%
      - track_album_name: 0.000152%

11. **Da li ima nevalidnih vrednosti u bazi? Ako ima, navesti za svako od obeležja koje su vrednosti nevalidne i zašto se smatraju nevalidnim.**

    - U bazi postoji jedna nevalidna vrednost. Jedan uzorak predstavlja reklamu koja traje 4 sekunde (4000 ms) što je netipično za regularne pesme.

- Obeležja u kojima taj uzorak sadrži nevalidne vrednosti su:
    - danceability: 0
    - speechiness: 0
    - acousticness: 0
    - liveness: 0
    - valence: 0
    - tempo: 0
    - duration_ms: 4000

12. Ako ima nedostajućih i/ili nevalidnih vrednosti u bazi, za svako od obeležja navesti kako će problem biti rešen.

- Za obeležja koja imaju nedostajuće vrednosti, izbacićemo kolone koja sadrže ta obeležja, jer smatramo da su nam ona nebitna za obuku modela.
- Što se tiče nevalidnih vrednosti, uzorak koji sadrži te vrednosti će biti izbačen iz skupa podataka.

13. Kada je završeno izbacivanje, dopuna, i drugo, navesti koliko je u sređenoj bazi ostalo uzoraka, a koliko obeležja.

- Ostaje nam 32832 uzoraka i 16 obeležja.

14. Da li neka od obeležja sadrže autlajere? Navesti koja obeležja ih sadrže.

- Obeležja koja sadrže autlajere su:
    - danceability
    - energy
    - loudness
    - speechiness
    - liveness
    - tempo
    - duration

15. Da li postoje parovi obeležja korelisani više od 0.7? Navesti takve parove obeležja.

- Ne postoji nijedan par obeležja korelisan više od 0.7. Međutim, postoji jedan par blizu 0.7, i to par loudness–energy (Glasnoća–Energičnost) koji su korelisani vrednošću 0.68.

16. Ako se rešava regresioni problem: utvrditi koliko je odstupanje raspodele varijable koja se predviđa od normalne raspodele dobijene korišćenjem uzoračke sr.vr. i standardne devijacije (asimetričnost i spljoštenost)?

- Asimetričnost: -0.23334
- Spljoštenost: -0.93272