Tim_5

Branislav Stojković SV64/2020

Anja Petković SV22/2020

Predikcija stadijuma raka pluća

Tema projekta: Predikcija stadijuma raka pluća kod pacijenata korišćenjem mašinskog učenja.

Opis problema: Rak pluća je jedna od najsmrtonosnijih bolesti širom sveta. Precizna predikcija stadijuma raka za pacijente sa rakom pluća može značajno pomoći lekarima u ranom otkrivanju bolesti i donošenju odluka o tretmanu. Ovaj projekat ima za cilj da razvije model mašinskog učenja koji može predvideti stadijum raka pluća koristeći različite demografske, medicinske i tretmanske podatke.

Podaci: Za potrebe ovog projekta koristićemo sintetički dataset koji pruža detaljne informacije o pacijentima sa rakom pluća. Dataset obuhvata sledeće atribute:

- **Patient_ID**: Identifikacioni broj pacijenta
- **Age**: Godine starosti pacijenta
- **Gender**: Pol pacijenta
- **Smoking_History**: Istorija pušenja pacijenta (npr. Current Smoker, Never Smoked, Former Smoker)
- **Tumor_Size_mm**: Veličina tumora u milimetrima
- **Tumor_Location**: Lokacija tumora (npr. Lower Lobe)
- **Stage**: Stadijum raka (npr. Stage I, Stage II, Stage III)
- **Treatment**: Tretman koji pacijent prima (npr. Surgery, Radiation Therapy, Chemotherapy)
- **Survival_Months**: Broj meseci preživljavanja nakon dijagnoze
- **Ethnicity**: Etnička pripadnost pacijenta
- **Insurance_Type**: Tip osiguranja (npr. Medicare, Private, Other)
- **Family_History**: Porodična istorija raka (Yes/No)
- **Comorbidity_Diabetes**: Prisustvo dijabetesa (Yes/No)
- **Comorbidity_Hypertension**: Prisustvo hipertenzije (Yes/No)
- **Comorbidity_Heart_Disease**: Prisustvo srčanih oboljenja (Yes/No)
- **Comorbidity_Chronic_Lung_Disease**: Prisustvo hroničnih plućnih bolesti (Yes/No)
- **Comorbidity_Kidney_Disease**: Prisustvo bolesti bubrega (Yes/No)
- **Comorbidity_Autoimmune_Disease**: Prisustvo autoimunih bolesti (Yes/No)
- **Comorbidity_Other**: Prisustvo drugih komorbiditeta (Yes/No)
- **Performance_Status**: Performans status pacijenta (skala od 0 do 4)
- **Blood_Pressure_Systolic**: Sistolički krvni pritisak
- **Blood_Pressure_Diastolic**: Dijastolički krvni pritisak
- **Blood_Pressure_Pulse**: Puls
- **Hemoglobin_Level**: Nivo hemoglobina
- **White_Blood_Cell_Count**: Broj belih krvnih zrnaca
- **Platelet_Count**: Broj trombocita
- **Albumin_Level**: Nivo albumina

- **Alkaline_Phosphatase_Level**: Nivo alkalne fosfataze
- **Alanine_Aminotransferase_Level**: Nivo alanin aminotransferaze
- **Aspartate_Aminotransferase_Level**: Nivo aspartat aminotransferaze
- **Creatinine_Level**: Nivo kreatinina
- **LDH_Level**: Nivo laktat dehidrogenaze
- **Calcium_Level**: Nivo kalcijuma
- **Phosphorus_Level**: Nivo fosfora
- **Glucose_Level**: Nivo glukoze
- **Potassium_Level**: Nivo kalijuma
- **Sodium_Level**: Nivo natrijuma
- **Smoking_Pack_Years**: Broj godina pušenja (izraženo u pakovanjima godišnje)

Link do dataset-a: https://www.kaggle.com/datasets/rashadrmammadov/lung-cancer-prediction

Metodologija:

1. **Prikupljanje i analiza podataka:**
   - Preuzimanje dataset-a i pregled strukture podataka.
   - Obrada podataka: čišćenje podataka, rukovanje nedostajućim vrednostima, kodiranje kategorijalnih podataka i normalizacija numeričkih atributa.
   - Moguća redukcija dimenzionalnosti.
   - Podela inicijalnog skupa podataka na skup za trening i skup za test u razmeri 70:30.
2. **Model:**
   - Korišćenje ansambl modela (Bagging, Stacking, Voting) za izradu prediktivnog modela.
   - Optimizacija hiperparametara korišćenjem kros validacije.
3. **Evaluacija modela:**
   - Evaluacija performansi modela korišćenjem mikro F1 skora.
   - Evaluacija performansi modela korišćenjem makro F1 skora.
   - Evaluacija ponašanja modela (Precision, Recall i F1 score) po klasama.
4. **Implementacija i testiranje:**
   - Testiranje modela na nepoznatim podacima i procena njegove generalizacije.

Evaluacija: Evaluacija rešenja će se vršiti korišćenjem sledećih metrika:

- **Mikro F1 skor:** Harmonijska sredina preciznosti i odziva, uzimajući u obzir sve instance. Ova metrika pruža balans između preciznosti i odziva na mikro nivou, tj. na nivou svih instanci zajedno.
- **Makro F1 skor:** Harmonijska sredina preciznosti i odziva za svaku klasu posebno, pružajući balans između preciznosti i odziva na nivou svake klase.