Breast Cancer Diagnostic Classification – Projekat

Opis problema

Predviđanje maligniteta tumora (*benigni / maligni*) bi trebalo da bude ostvareno *na osnovu numeričkih karakteristika ćelijskih jedara* dobijenih iz digitalizovanih slika biopsija. Očekuje se da model omogući identifikaciju ključnih faktora povezanih sa razlikovanjem među malignim i benignim uzorcima, uz jasan fokus na analizu, vizuelizaciju i interpretaciju, koristeći isključivo *data.csv*.

Pregled kolona i njihovo značenje

Dataset (*data.csv*) sadrži sledeće podatke za svaki od 569 uzoraka:

- **diagnosis** — **ciljana promenljiva** (Benignni / Maligni tumor).
- **mean radius** — prosečna udaljenost od centra do tačaka na periferiji jezgra ćelije.
- **mean texture** — standardna devijacija vrednosti sivih tonova (tekstura).
- **mean perimeter** — obim jezgra.
- **mean area** — površina jezgra.
- **mean smoothness** — lokalna varijacija dužine radijusa.
- **mean compactness** — vrednost izražena formulom: (perimeter^2 / area) – 1.0.
- **mean concavity** — ozbiljnost konkavnih delova oboda jezgra.
- **mean concave points** — broj konkavnih delova na obodu jezgra.
- **mean symmetry** — stepen simetrije jezgra.

- **mean fractal dimension** — približna “dužina obale” u prostoru (fraktalna dimenzija).
- **radius se** — standardna greška za radijus.
- **texture se** — standardna greška za teksturu.
- **perimeter se** — standardna greška za obim.
- **area se** — standardna greška za površinu.
- **smoothness se** — standardna greška za glatkoću.
- **compactness se** — standardna greška za kompaktnost.
- **concavity se** — standardna greška za konkavnost.
- **concave points se** — standardna greška za broj konkavnih tačaka.
- **symmetry se** — standardna greška za simetriju.
- **fractal dimension se** — standardna greška za fraktalnu dimenziju.
- **worst radius** — najveća (ili najteža) vrednost radijusa među merama.
- **worst texture** — najveća vrednost teksture.
- **worst perimeter** — najveća vrednost obima.
- **worst area** — najveća vrednost površine.
- **worst smoothness** — najveća vrednost glatkoće.
- **worst compactness** — najveća vrednost kompaktnosti.
- **worst concavity** — najveća vrednost ozbiljnosti konkavnih delova.
- **worst concave points** — najveći broj konkavnih tačaka.
- **worst symmetry** — najveći stepen simetrije.

- **worst fractal dimension** — najveća vrednost fraktalne dimenzije.

Navedene karakteristike odražavaju statističku obradu parametara ćelijskog jezgra izrađenih za svaki prikupljeni uzorak (srednja vrednost, standardna greška, ekstremne vrednosti).

Očekivani izlazi (output):

- **Evaluacija klasifikacionog modela** kroz metrike: *tačnost, preciznost, odziv i F1-skor*.

- **Vizualna analiza najznačajnijih atributa** (*feature importance*) izražena grafički.

- **Grafički prikazi** distribucija i odnosa između ključnih karakteristika i tipa tumora (benigni/maligni).

- **Tumačenje rezultata** kroz identifikovane osobine koje najviše doprinose razlikovanju malignih i benignih slučajeva, uz preporuke za eventualna poboljšanja modela.

Zahtevi za procesiranje podataka:

- **Podaci moraju biti obrađeni** tako da odsustvo ili anomalije u vrednostima ne utiču negativno na model.

- **Korišćenje numeričkih atributa** (mean, se, worst vrednosti) mora biti opravdano u kontekstu njihove relevantnosti za diferencijaciju između klasa.

- **Bilo koja odluka o isključivanju ili grupisanju atributa** (npr. kombinovanje srednjih i worst vrednosti, redukcija dimenzija) mora biti jasno obrazložena.

Dokumentovanje rezultata:

- **Prikaz rezultata modela i vizuelni elementi** (*grafici, tabele*) moraju biti predstavljeni jasno, u formatu pogodnom za Word dokument.

- **Interpretacije i zaključci** moraju biti utemeljeni na analizi podataka, uz eventualne sugestije za poboljšanja modela ili dodatne eksperimente.