Napovedovanje raka dojk (breast cancer)

Miha Prajs, 89231246

Opis projekta: Določanje ali je tumor maligen oz. rakav ali benigen z uporabo algoritma k-NN.

Dataset: Za projekt bom uporabil Cancer dataset iz Kaggle (https://www.kaggle.com/datasets/erdemtaha/cancer-data, dostop 16. 1. 2025). Dataset vsebuje 32 atributov, kjer vsak atribut predstavlja specifično značilnost ali informacijo o 569 pacientih in njihovem stanju. Ključne atributi dataseta so:

- *id*: Unikaten identifikator vsakega pacienta.
- *diagnosis*: Vrsta raka, ki je lahko *"M"* (maligen) ali *"B"* (benigen).
- *radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean*: Povprečne vrednosti vizualnih značilnosti raka, ki zajemajo različne mere, kot so oblika, tekstura, gladkost, gostota in konkavnost tumorskih celic.
- *Dodatne značilnosti*: Dataset vključuje tudi več kategorijskih značilnosti, ki pacientom dodeljujejo numerične vrednosti, ter tabelarne povzetke povprečnih vrednosti vizualnih značilnosti za različne razpone.

Vsak vzorec vsebuje unikaten ID pacienta, diagnozo raka ter povprečne vrednosti vizualnih značilnosti tumorskih celic. Dataset je primeren za učenje in testiranje modelov, ki se uporabljajo za diagnozo raka. Analiza podatkov lahko prispeva k izboljšanju metod odkrivanja in razumevanja raka.

Algoritmi in pristopi: Pripravo podatkov bom izvedel z uporabo R-a, algoritem k-NN pa bom implemential v Javi. R koda bo v svoji R-skripti, ki bo nato pognana v Javi z uporabo Oracle-ove knjižnice FastR (https://github.com/oracle/fastr). Za primerjavo bom k-NN algoritem implementiral še z R-jem.

Cilji: Doseči visoko zanesljivost in uporabnost napovedi pri določanju raka na dojkah.