**Предвидување на срцева болест со**

**машинско учење и визуелизација преку**

**Spring Boot (MVC) апликација**

[**Video**](https://www.dropbox.com/scl/fi/z1zqxn94rgzzv21vqg3np/chrome_wZ2WX4DVSM.mp4?rlkey=1lc3rubs15x0dozneumk7ragb&st=f86wamkq&dl=0)

1. **Опис**

* Овој проект цели кон предвидување на присуството на срцева болест кај пациенти со соодветни испитувања и податоци преку употреба на неколку класификациски алгоритми. Резултатите од класификаторите како и податоците за пациентите се претставени со помош на Spring Booot MVC апликација, што дозволува интерактивен приказ да се соочат предвидувањата и релевантните податоци пред и после тренирање на моделите.

1. **Информации за податоците**

* Податочното множество кое се користи може да го видите [тука](https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data). Содржи записи за пациенти со нивните атрибути, опишани подолу.
  1. Атрибути

 Age: Возраст на пациентот.

 Sex: Пол.

 CP (Chest Pain Type): Вид на болка во градите.

 Trestbps (Resting Blood Pressure): Крвен притисок (во mm Hg).

 Chol (Serum Cholesterol): Холестерол во mg/dl.

 Fbs (Fasting Blood Sugar): Шеќер во крвта> 120 mg/dl (1 = true, 0 = false).

 Restecg (Resting Electrocardiographic Results): Електрокардиографски резултати

 Thalach (Maximum Heart Rate Achieved): Максималното срцево отчукување.

 Exang (Exercise Induced Angina): Болка во градите при активност

 Oldpeak: Падот на ST сегмент на електрокардиограм.

 Slope: Наклонот на ST сегментот при максимален физички напор.

 Ca: Број на главни крвни садови (0-3) обоени преку флуороскопија..

 Thal: Таласемија (Thal): 3 = нормална, 6 = фиксен дефект, 7 = реверзибилен дефект..

 Target (Heart Disease): Цел (Срцево заболување): 0 = нема заболување, 1 = заболување.

**3. Модели на машинско учење**

Проектот ги користи следниве класификатори за машинско учење за да го предвиди присуството на срцеви заболувања:

3.1 K-Најблиски Соседи (KNN)

* За моделот: KNN работи со наоѓање на *k* најблиски податочни точки и класифицирање на пациентот врз основа на мноозинството на тие точки.
* Хиперпараметри: Оптималната вредност на *k* беше одредена користејќи cross validation.

3.2 Логистичка регресија

* За моделот: Логистичката регресија ја моделира веројатноста за срцево заболување користејќи логистичка функција. Таа дава веројатности кои се претвораат во бинарни предвидувања.
* Перформанси: Логистичката регресија често се користи како базичен модел за бинарни класификациски проблеми.

3.3 Random Forest

* За моделот: Random Forest е ансамбл метод кој користи повеќе одлучувачки дрва. Крајната класификација се базира на мнозински глас на поединечните дрва.
* Предности: Random Forest може да ги улови нелинеарните односи и ефективно да ја искористи важноста на карактеристиките.

3.4 SVM

* За моделот: SVM ја наоѓа хиперрамнината која најдобро ги раздвојува класите (срцево заболување или не) во високодимензионален простор.
* Кернел: Користен е линеарен или RBF кернел во зависност од сложеноста на податоците.

3.5 Наивен Баесов Класификатор

* За моделот: Наивниот Баесов класификатор е базиран на Баесовата теорема и ја претпоставува независноста помеѓу карактеристиките. Иако оваа претпоставка е „наивна“ и не секогаш точно ја одразува реалноста, овој модел често дава добри резултати во класификациски проблеми.
* Предности: Добро се справува со големи податочни сетови и често е ефикасен.

3.6 Евалуација на моделите

Секој класификатор се оценува користејќи ги следниве метрики:

* Точност (Accuracy): Пропорцијата на точни предвидувања.
* Прецизност (Precision): Колку избраните елементи се релевантни.
* Повик (Recall/Сензитивност): Колку релевантни елементи се избрани.
* F1 скор: Хармонична средина на прецизност и повик.
* ROC-AUC Скор: Мери колку добро моделот ги раздвојува класите.

**4. Визуелизации и интеграција со Spring Boot**

Spring Boot MVC апликација е развиена за прикажување на резултатите од моделот и визуелизации на податоците. Корисниците можат интерактивно да ги гледаат резултатите од класификациите и метриките за перформанси.

4.1 Визуелизациски функции

* Визуелизации пред тренирање:
  + Countplot со различни атрибути и таргет
  + Фреквентност на атрибутите и таргетот
  + Опис на податоците
  + Scatter plots
* Визуелизации по тренирање:
  + Матрица на конфузија (Confusion Matrix): Визуeлизира вистински наспроти предвидени класификации.
  + ROC Крива: За бинарна класификација, прикажува компромис помеѓу вистинската позитивна и лажната позитивна стапка.
  + Извештај за класификација (Classification Report): Прецизност, повик, F1-скор и метрики за точност, прикажани графички како столбчест график.

4.2 Spring Boot апликација

Spring Boot апликацијата обезбедува кориснички интерфејс со следниве карактеристики:

* Почетна страница:
  + - Приказ на сите модели и нивно филтрирање по име, со копчиња за визуелизација на секој модел посебно како и на целото податочно множество
* Back-end (Java + Spring Boot)
* Front-end (Thymeleaf + HTML)

.

**Автори:**

* Самоил Јаќимовски (211036)
* Бранко Георгиев (213077)