# Cardiovascular Disease Prediction

## Abstract

Cardiovascular diseases (CVDs) are disorders of heart and blood vessels. This is the leading cause of deaths worldwide. Early detection and diagnosis can help the patients. Machine learning can be used to create the predictive model using cardiovascular diseases risk factors like cholesterol level, glucose level and blood pressure

The goal of this project was to use classification models to predict the presence or absence of a cardiovascular disease (CVD) using the patient examination results.

I employed different machine learning, and the results shows that the random forest model has achieve promising results for this problem.

## Design

Source of data is kaggle machine learning competitions. The machine learning algorithms used in this project are Random Forest, Decision Tree, MLP, KNN and Logistic Regression.

This dataset presents a three There are 3 types of input features:

- Objective: factual information;
- Examination: results of medical examination;
- Subjective: information given by the patient.

The results of comparison Between the models shows that Random Forest achieve high classification accuracy of 73 %, The model can be used in medical field for prediction of cardiovascular diseases

## Data

The dataset consists of 70000 records of patients data in 13 features, such as age, gender, systolic blood pressure, diastolic blood pressure, and etc. The outcome of this dataset is in binary form which 1 indicates having cardiovascular disease and 0 represents none.
Credits for Dataset : Svetlana Ulianova

# Algorithms

*Feature Engineering*

1. Imputation : luckly the dataset doesn’t have missing or duplicated values
2. Handling outliers records containing outliers are removed from the distribution

*Models*

Logistic regression, k-nearest neighbors, and random forest , Decision Tree, and MLP classifiers were used before settling on random forest as the model with strongest cross-validation performance.

*Model Evaluation and Selection*

The entire training dataset of 70000 records was split into 80/20 train vs. test, and all models scores were calculated with 10-fold cross validation .

**Final random forest after hyperparameter tuning using Grid Search CV scores:**

- Accuracy: 0.73
- F1: 0.72 macro avg, 0.73 weighted avg
- Precision: 0.73 macro avg, 0.73 weighted avg
- Recall: 0.73 macro avg, 0.73 weighted avg

# Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting and visualizations

# Communication

In addition to the slides and the jupyter note book Code submited, I will deliver a 5 minutes slide presentation in the final day.