Heart disease is the major cause of morbidity and mortality globally: it accounts for more deaths annually than any other cause. According to the WHO, an estimated 17.9 million people died from heart disease in 2016, representing 31% of all global deaths. Doctors and scientists alike have turned to machine learning (ML) techniques to develop screening tools and this is because of their superiority in pattern recognition and classification as compared to other traditional statistical approaches.

**Problem Statement:**

The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset provides the patients’ information. It includes over 3390 records and 17 attributes.

The project has been conducted in 5 steps:

1. Data Cleaning
2. Exploratory Data Analysis (EDA)
3. Data Transformation
4. Model Building and Evaluation
5. Hyperparameter Tuning

First, we did data cleaning by fill null values by KNN Imputer then we separated our data into categorical and continuous variable after that performed EDA. Before applying algorithm, we did some techniques such as, SMOTE for balancing dependent variable data, Standardization **to remove outliers by making mean 0 and standard deviation 1 after that we used PCA to reduce the dimensionality of data then** we apply classification algorithm such as, **Logistic Regression, KNeighbors Classifier, Decision Tree Classifier, Naive Bayes classifiers,** **Support vector machine, SGD Classifier.** we used GridsearchCV to do hyperparameter optimization and tweak our model to achieve the best possible outcome and were able to raise recall (false negative) performance matrix above other performance matrixes such as accuracy, F1, and so on.

**Results**

The support vector machine was the best performing model across all metrics. Its best parameters were a radial kernel. Its high AUC and Recall also show that the model has a low false negative rate and is thus sensitive to predict if one has a high risk of developing CHD, i.e., getting a heart attack within 10 years.

**1. Tanjul Gohar**

* Introduction
* Data handling
* EDA

Categorical Variables

* Correlation

Correlation of all categorical variables

* Correlation and multicollinearity of all variables
* Model performance

Logistic Regression

Support Vector Classifier

Naïve Bayes Classifier

* Metric comparison
* Conclusion

**2. Kanika Singh**

* Introduction
* Data handling
* EDA

Continuous Variables

* Correlation

Correlation of all continuous variables

* Correlation and multicollinearity of all variables
* Model performance

Decision Tree

KNN

SGD Classifier

* Metric comparison
* Conclusion

**GITHUB LINK-**

**https://github.com/Tanjul5/Cardiovascular-Risk-Prediction.git**

**DRIVELINK- https://drive.google.com/drive/folders/11PpXRaYwhhh8XlMbl94OoIbJCm9b59uX?usp=sharing**