A HEART DISEASE PREDICTOR USING MACHINE LEARNING

TEAM - 6

1. **M.Dharani (Malla Reddy Engineering College For Women)**
2. **VAIBHAVA LAKSHMI (Malla Reddy Engineering College For Women)**
3. **Vedhavrath (Malla Reddy Institute of Engineering and Technology)**
4. **Sai Bharadwaj (MLR Institute of Technology)**

Table of Contents:

Abstract

Heart disease commonly occurring disease and is the major cause of sudden death nowadays. This disease attacks the persons instantly. Most of the people do not aware of the symptoms of heart disease. Timely attention and proper diagnosis of heart disease will reduce the mortality rate. Medical data mining is to explore hidden pattern from the data sets. Supervised algorithms are used for the early prediction of heart disease. Nearest Neighbor (KNN) is the widely used lazy classification algorithm. KNN is the most popular, effective and efficient algorithm used for pattern recognition. Medical data sets contain a large number of features. The Performance of the classifier will be reduced if the data sets contain noisy features. Feature subset selection is proposed to solve this problem. Feature selection will improve accuracy and reduces the running time. Particle Swarm Optimization (PSO) is an Evolutionary Computation (EC) technique used for feature selection. PSO are computationally inexpensive and converges quickly. This paper investigates to apply KNN and PSO for prediction of heart disease. Experimental results show that the algorithm performs very well with 100% accuracy with PSO as feature selection.

Introduction:

Coronary Heart Disease (CHD) is obstruction of the coronary arteries with symptoms such as angina, chest pain, and heart attacks. Arteries supply blood to heart muscle. CHD is a leading cause of death in many countries. In India there are roughly 3 crore heart patients and 2 lakh open heart surgeries are performed every year. CHD is a leading cause of mortality claiming nearly 17.3 million people every year. The reason for this is smoking, high levels of cholesterol, diabetes. Early prediction of heart disease is essential to reduce the mortality rate. Data mining provides a user-oriented approach to extract novel and uncovered patterns in the data set. Data mining is to extract useful knowledge within medical data for medical diagnosis. Data mining is widely applied in the medical domain. Medical data mining is used to infer diagnostic rules and help physicians to make diagnosis process more accurate. K-nearest neighbour is the most widely used lazy classification algorithm as it reduces misclassification error. Feature Subset Selection (FSS) is widely used in data mining and machine learning. FSS is a dimensionality reduction technique use to enhance accuracy. Particle swarm optimization is an effective EC technique used as feature selection. PSO converges quickly and is computationally inexpensive.

Table of graphs or figures:

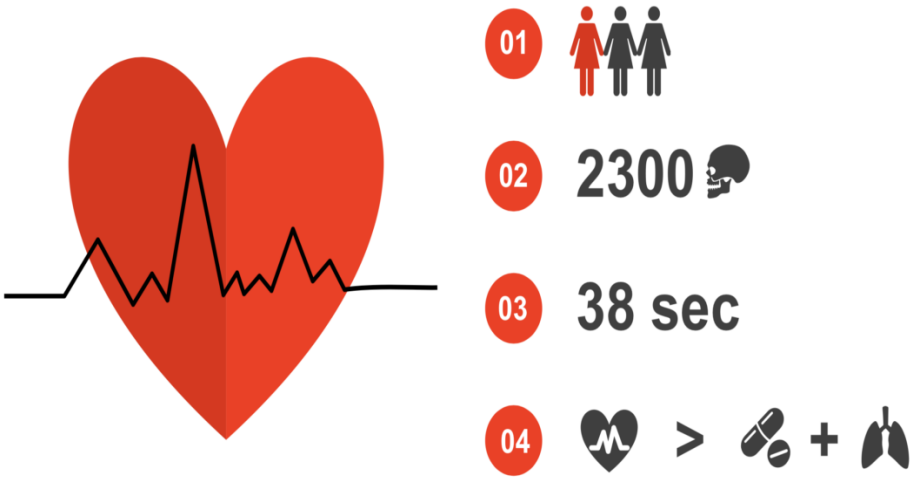


Fig 1: STATS REGARDING HEART DISEASES

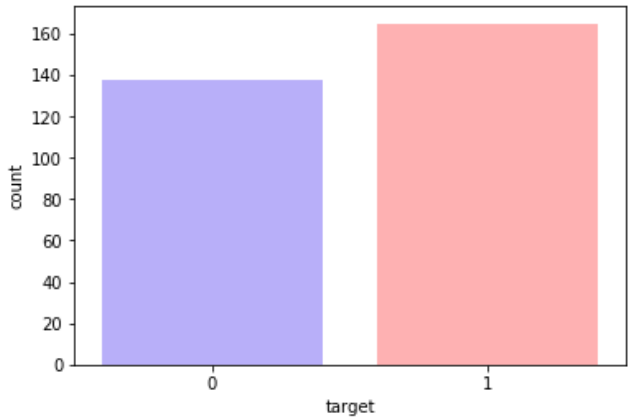


Fig 2: BAR PLOT FOR COUNT OF PEOPLE DISEASED AND NOT DISEASED

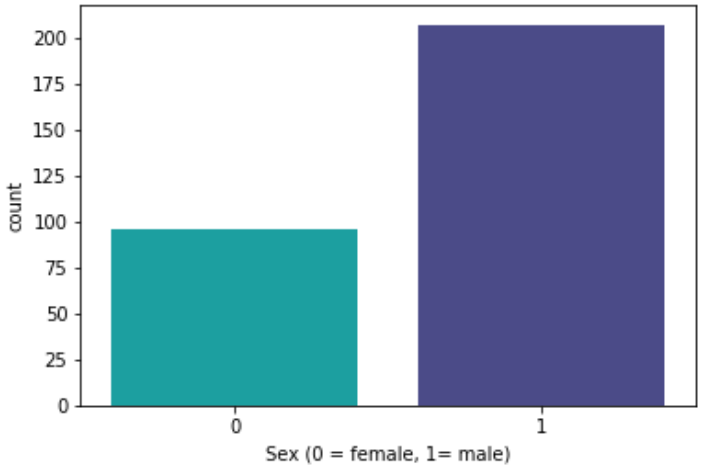


Fig 3: BAR PLOT FOR COUNT OF MALE AND FEMALE

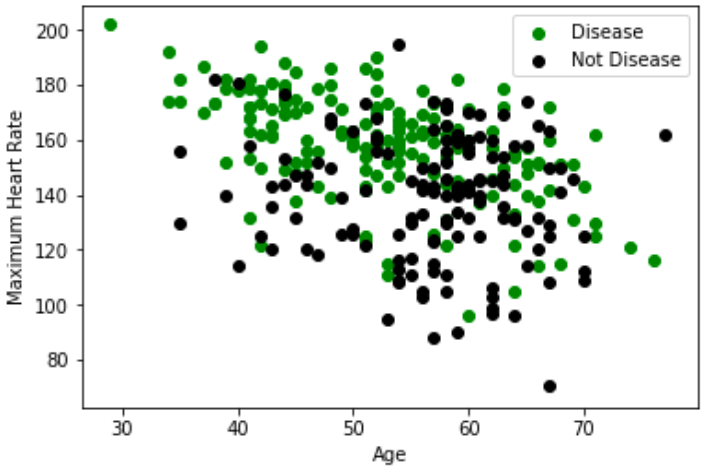


Fig 4: SCATTER PLOT BETWEEN AGE AND MAXIMUM HEART RATE

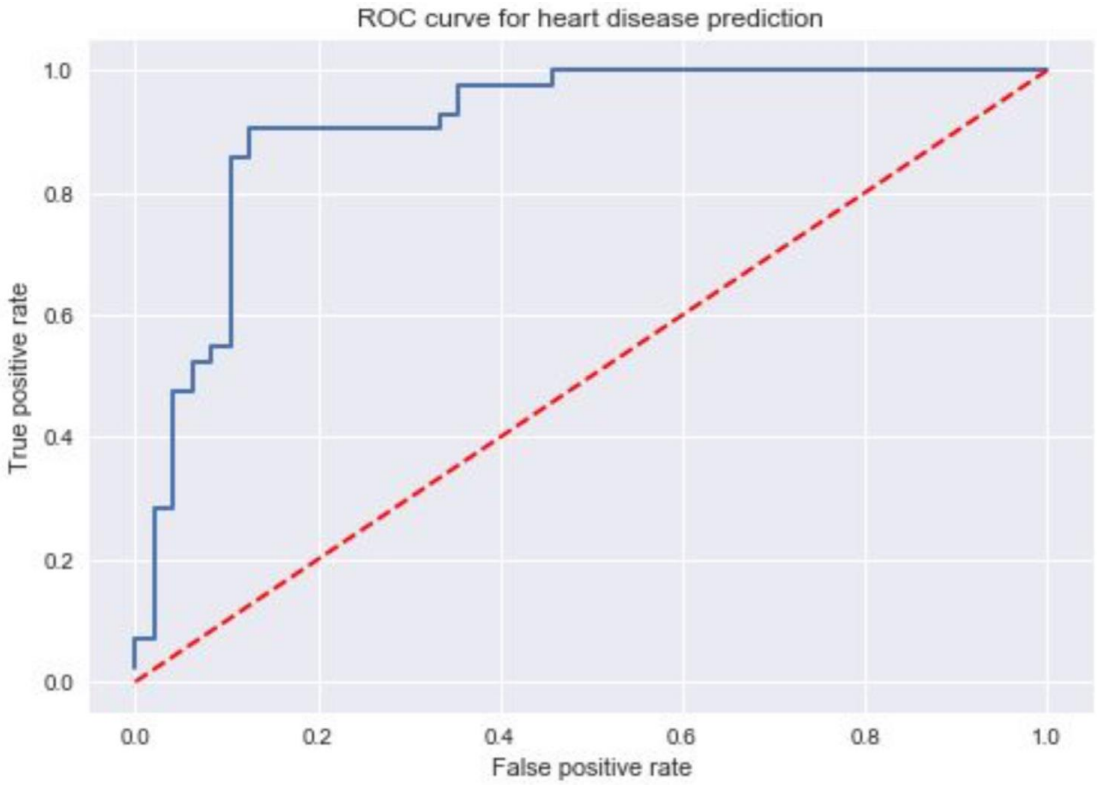


Fig 5: ROC CURVE FORHEART DISEASE PREDICTION

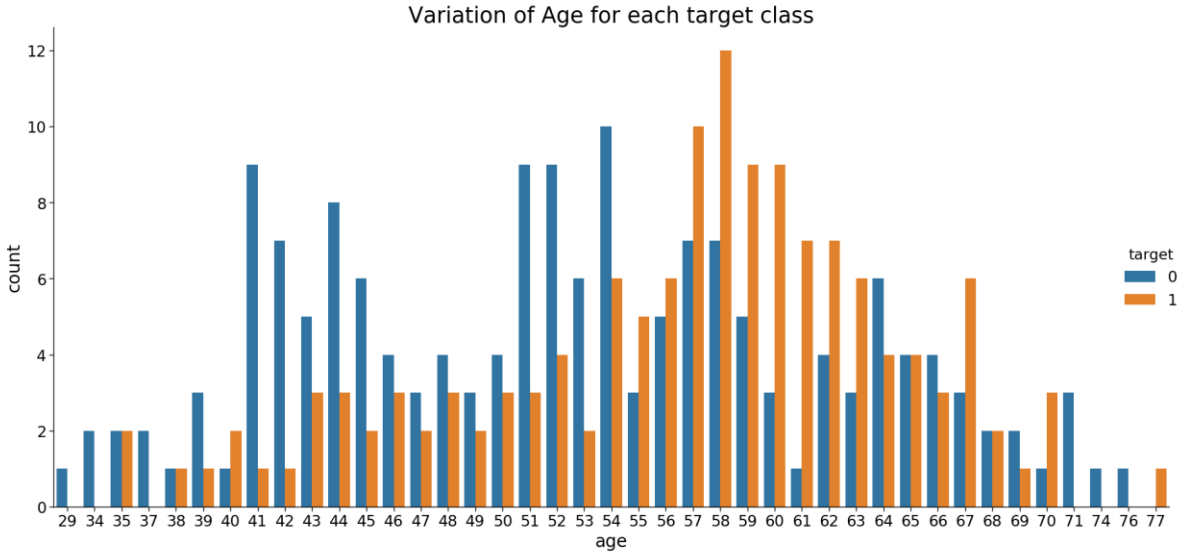


Fig 6: Heart Disease frequency for various ages

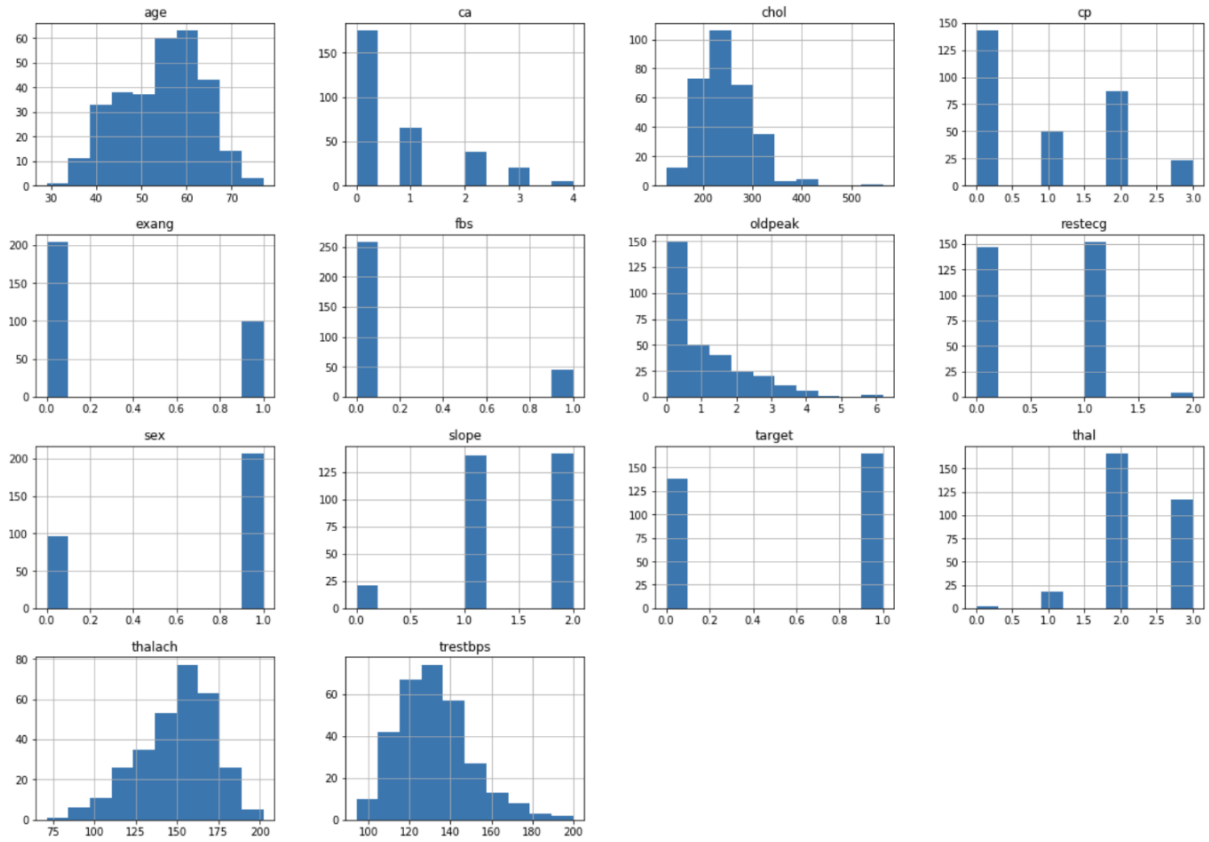


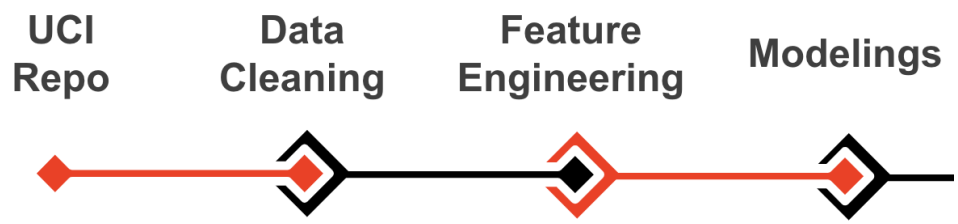
Fig 7: Bar plots for the given data

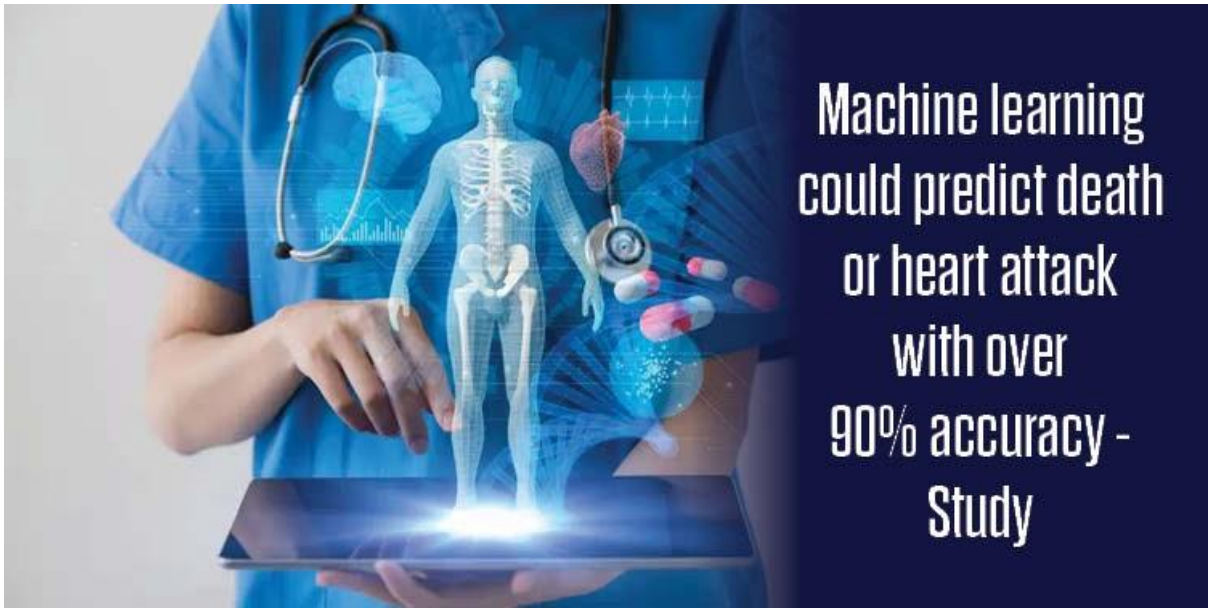
Project Design:

The problem that we are trying to solve is to predict/classify whether a person has heart disease or not, based on some of the personal info and medical test results.

We used some libraries like numpy, pandas, matplotlib, seaborn, sklearn for this project.

PIPE LINE:





Data:

The data that we had worked with is from UCI repo, The dataset contains 303 entries and 14 features. (The dataset is a little bit old, it is from 1988.)

The dataset includes features such as *age, gender, resting blood pressure, chest pain type and etc*. Some of these test can only be performed in a clinic, and patient won't be able to perform those tests at home by themselves. So my target audience will be doctors or nurses.

Dataset contains following features:

age — age in years
sex — (1 = male; 0 = female)
cp — chest pain type
trestbps — resting blood pressure (in mm Hg on admission to the hospital)
chol — serum cholestoral in mg/dl
fbs — (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg — resting electrocardiographic results
thalach — maximum heart rate achieved
exang — exercise induced angina (1 = yes; 0 = no)
oldpeak — ST depression induced by exercise relative to rest
slope — the slope of the peak exercise ST segment
ca — number of major vessels (0–3) colored by flourosopy
thal — 3 = normal; 6 = fixed defect; 7 = reversable defect
target — have disease or not (1=yes, 0=no)

Cleaning and EDA:

The target variable is the last column from the original dataset which is named as num. It is a categorical feature labeled as 0, 1, 2, 3, 4. While 0 means the patient has no heart disease, 1, 2, 3, 4 mean that the patient has some kind of heart disease. Since our problem is to predict whether the person has heart disease or not, we grouped 1, 2, 3, 4 all together as one group to label the person has heart disease.

Modeling and Feature Engineering:

Next, we decided to do some feature engineering to try to find out which variables contribute significantly to our models. We used chi-square test from sklearn to find out significance level of each variable to my model, and we selected all the variables that has a score is positive and high.

At last we developed four models for predicting heart disease:

1. Support Vector Machines (SVM)
2. K – nearest neighbor classifier (K-NN)
3. SVM with PCA
4. K-NN with PCA

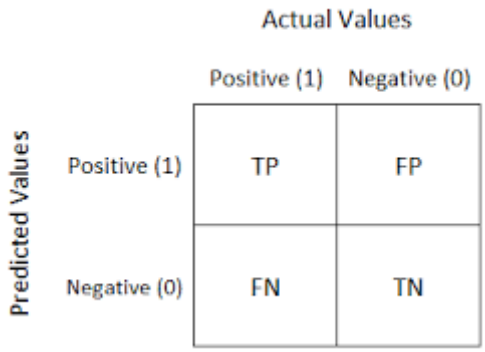
Confusion matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm.

The number of correct and incorrect predictions are summarized with count values and broken down by each class.

The confusion matrix shows the ways in which your classification model is confused when it makes predictions.

This also helps us to calculate Accuracy, Recall, Precision, F-Measure.



$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

We calculated accuracy and plotted the confusion matrix using seaborn module or plot_confusion_matrix from sklearn module .

Codes for all models:

1. **KNN MODEL**
   https://tinyurl.com/ybkzy8mg
2. **SVM MODEL**
   https://tinyurl.com/ya7jswn9
3. **SVM WITH PCA**
   https://tinyurl.com/ycwz89sl
4. **KNN WITH PCA**
   https://tinyurl.com/y98njagb

Conclusion:

Future Improvement

For the next step, if we have a chance to extend my project, there are a few things we would like to try:

- **More data**: We would like to work with a larger dataset and a more recent one.
- **Domain knowledge**: Since we are approaching the problem only with data science knowledge, we would like to get more domain knowledge about heart disease from doctors. Because we think that might help us discover some of the hidden elements for our model. For example, the relationship of the cost of medical tests and how each test contribute to our model (sensitivity). This might be helpful for some patients who doesn't want to perform some of the expensive tests or medical tests that have higher risks.

Inference

Our approach uses KNN,SVM,SVM with PCA,KNN with PCA as a classifier to reduce the misclassification rate. The project involved analysis of the heart disease patient dataset with proper data processing. Then, 4 models were trained and tested with maximum scores as follows:

1. KNN MODEL: 86%
2. SVM MODEL: 89%
3. SVM WITH PCA MODEL: 85%
4. KNN WITH PCA MODEL: 91%