HEART DISEASE PREDICTION

Arjun Kumar , Dharmil Shah , Anjana Kethineni , Sakshi Parikh , Anila Chintalapati , Pranitha Lolla
Advisor: Dr. James Foulds
IS Spring 2020, University of Maryland Baltimore County Baltimore, Maryland 21250

Overview

Project Goal:

The goal of this project is predict the presence of cardiovascular disease from the patient data such as

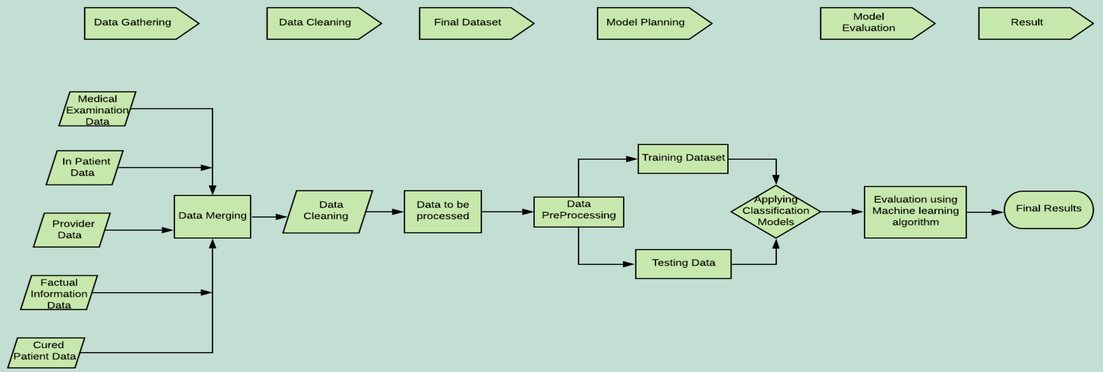
Objective: factual information;
Examination: results of medical examination;
Subjective: information given by the patient.

Dataset description:

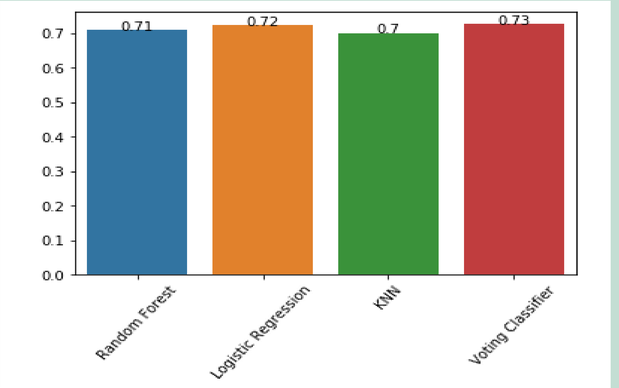
These are 11 features that are explained as per requirement:

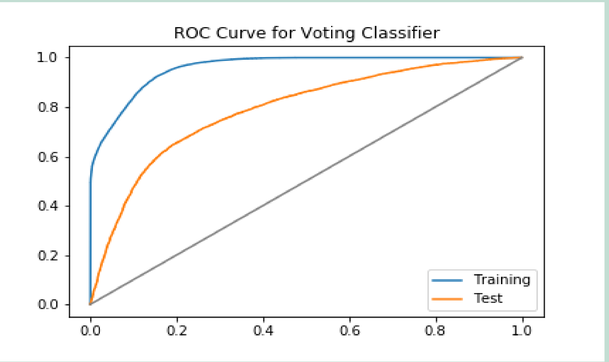
1. Age
2. Height
3. Weight
4. Gender.
5. Systolic blood pressure.(120- normal, 140- high B.P)
6. Diastolic blood pressure.(<80- low, >90- high B.P)
7. Cholesterol.
8. Glucose.
9. Smoking- smokes or not.
10. Alcohol- drinks or not.
11. Physical Activity.

Process Overflow



ACCURACY RESULTS





Data Preprocessing

Feature Selection:

- The original data set had many attributes. Only the attributes that contribute most are taken.

Handling Missing Values

- In this project we dropped the null values in the dataset.

Splitting Data

- Data is split into test and train sets on a random basis.

Encoding data

- The nominal attributes like cholesterol, glucose, alcohol, activity, smoking are taken as binary, or integer values(eg:1: normal, 2: above normal, 3: well above normal).
- The dataset has age in days and it is converted into years.

Challenges

- **Merging the data:** While merging it was challenging to analyze the appropriate features that to be selected.
- **Feature Reduction:** In order to retain useful feature, lot of calculation was required to merge multiple feature into fewer feature.
- **Convert Data to Years:** The initial dataset has age in days and challenging part is to convert the days into years for 70000 values

METHODS

Logistic Regression Classifier

Features are trained using logistic regression and evaluated. The purpose is to check linear behavior between dependent and Independent variables. Also this adds explicability to the predictions

Random Forest Classifier

The RFC handles the large data set with higher dimensionality. It can handle thousands of input variables and identify most significant variables. It also checks for non linearity between variables.

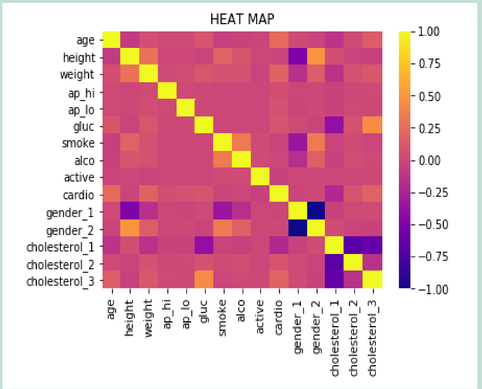
K-Nearest Neighbors(KNN)

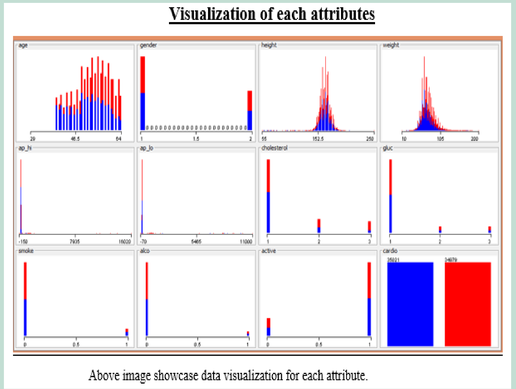
A supervised machine learning algorithm (as opposed to an unsupervised machine learning algorithm) is one that relies on labeled input data to learn a function that produces an appropriate output when given new unlabeled data

Voting Classifier

In soft voting, the output class is the prediction based on the average of probability given to that class. Suppose given some input to three models, the prediction probability for class A = (0.30, 0.47, 0.53) and B = (0.20, 0.32, 0.40). So the average for class A is 0.4333 and B is 0.3067, the winner is clearly class A because it had the highest probability averaged by each classifier.

DATA ANALYSIS





Conclusion

Models	Accuracy
Random Forest	71 %
KNN	70 %
Logistic Regression	72 %
Voting Classifier	73%

From the evaluation result we can conclude that Voting Classifier provides better accuracy of 73% which is almost 2% higher than other models such as Logistic Regression, KNN and Random Forest Classifier