IDS Project Report

# Prediction of Heart Disease

Code Link: Github

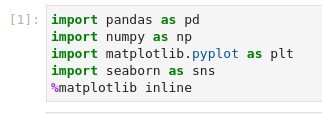
|  |  |
| --- | --- |
| Team: |  |
| Vani Agarwal | 18ucs098 |
| Abhay Singhal | 18ucs011 |
| Hritwik Singhal | 18ucs055 |
| Sameer Gupta | 18ucs008 |

## Project Goal:-

The goal of this project is to visualise the dataset containing factors like age, sex, cholestrol level, and from these factors predict if a person is suffering from heart disease or not.

Plan of Working

1. **Importing Libraries**:- We import libraries like matplotlib, pandas, numpy, seaborn, sklearn. Here, numpy is used for the mathematical part of working with the dataset and matrices. Then pandas is required for performing various kinds of slices and selections which are inbuilt in it.. Matplotlib and seaborn will be used for visualizations, sklearn has important libraries related to applying machine learning algorithms and check for its accuracies.



1. **Data Description:-** We import the dataset and give a description of it. Also, we have taken a preprocessed dataset, so we have not found any missing value or null values.
2. **Data Inferences:**- We shall visualise the dataset and get inferences from it. The graphs we shall plot shall help us to see various trends as on what factors does our target variable actually depend on and how other values relate among themselves.
3. After getting inferences, we shall think of **choosing the correct algorithm** to solve the above problem.
4. Finally we have applied our algorithm and see its **result** on the testing dataset and see how well our algorithm has performed and end this project by looking at their confusion matrix and other related things.

### Data Description

The dataset is taken from this link. The model we implemented is predictive and supervised. Output variable is binary (heart disease or not) and there are 14 input variables (you can see all of the variables below).

**Output variables**

1. Target (Heart disease)

1 = No

0 = Yes

**Input variables**

1. age (in years)
2. sex (sex)
   1. = female ,
   2. = male
3. cp: (chest pain type) Value 1: typical angina

Value 2: atypical angina

Value 3: non-anginal pain

Value 4: asymptomatic

1. trestbps (resting blood pressure (in mm Hg on admission to the hospital))
2. chol (serum cholesterol in mg/dl)
3. fbs (fasting blood sugar (> 120 mg/dl))

1 = true

* 1. = false

1. restecg (resting electrocardiographic results)

Value 0: normal,

Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV),

Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria)

1. thalach (maximum heart rate achieved)
2. exang (exang: exercise induced angina)

1 = yes,

* 1. = no

1. oldpeak (ST depression induced by exercise relative to rest)
2. slope (the slope of the peak exercise ST segment)

Value 1: upsloping

Value 2: flat

Value 3: downsloping

1. ca (number of major vessels (0-3) colored by fluoroscopy)
2. Thal (Thalassemia: A blood disorder)

3 = normal;

* 1. = fixed defect;
  2. = reversible defect

1. target (the predicted attribute)
   1. = heart disease ,
   2. = No heart disease

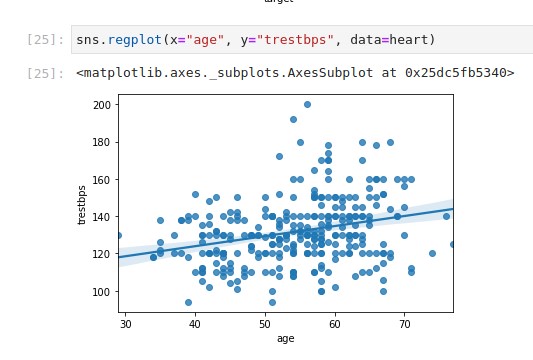
A sample of five rows.



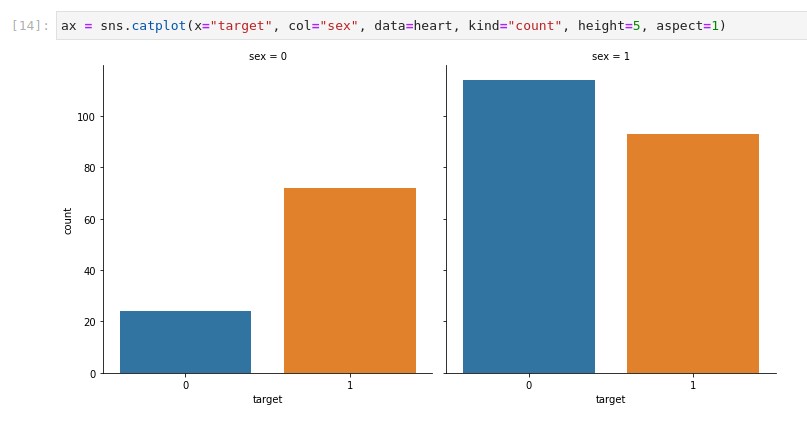
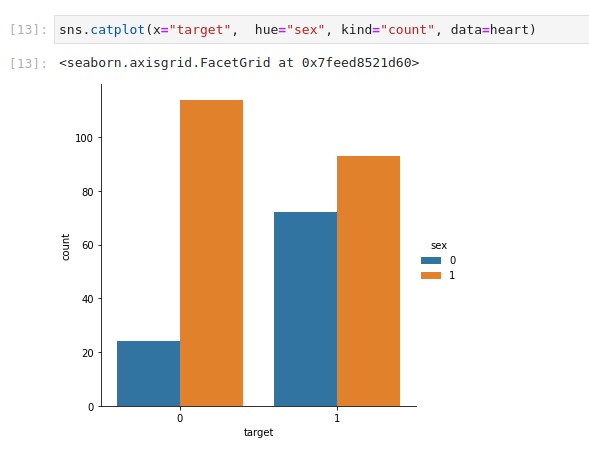
Inferences from Dataset:

We analyse our dataset and get some inferences from it:-

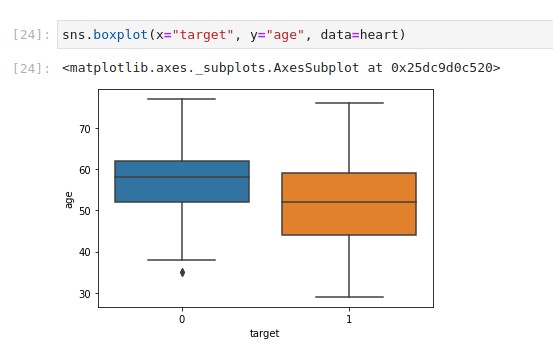
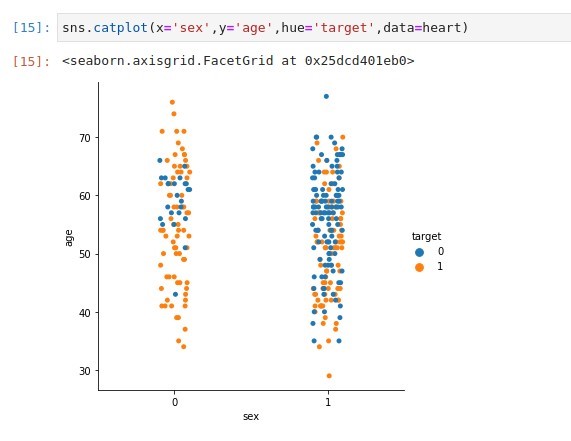
1. The patients with age range (40,70) have resting blood pressure(trestbps) between 120 to140 mm Hg.

.

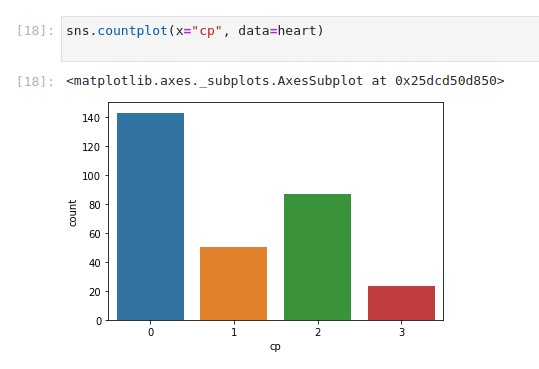
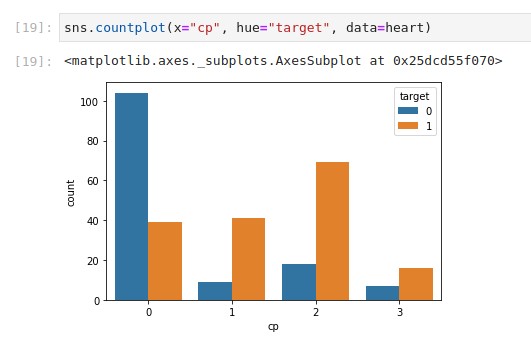
1. The ratio of heart disease / no heart disease is greater for males than females.



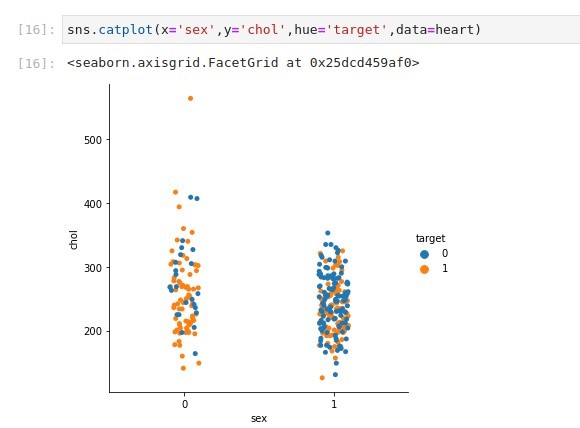
1. Heart disease occurs mostly to patients whose age is greater than 50 years.



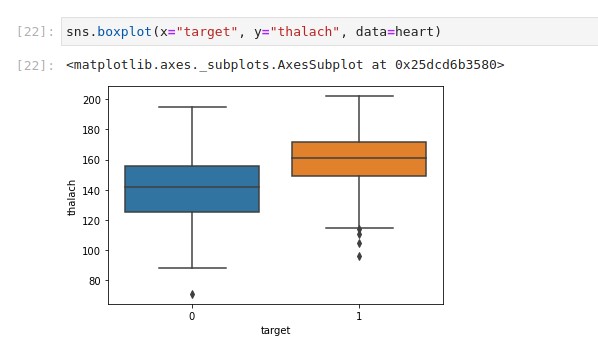
4 . The patients are reporting typical angina chest pain has a high risk of heart disease.



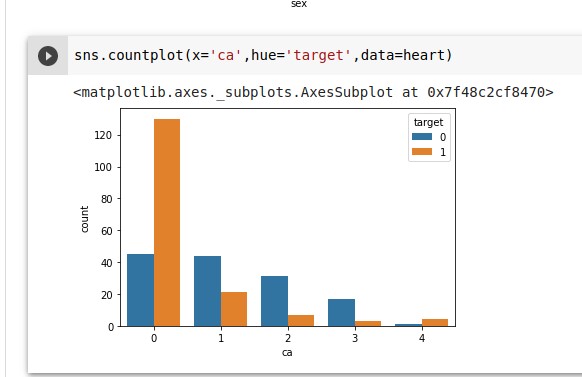
Patients with cholesterol higher than 180 have very high rates of heart attack. Further categorizing them on basis of sex, we see that males have much much higher heart attack rates than females in any segment.



1. The people suffering from heart disease (target = 0) have relatively higher heart rate (thalach) as compared to people who are not suffering from heart disease (target = 1).

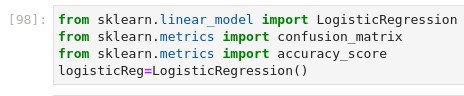
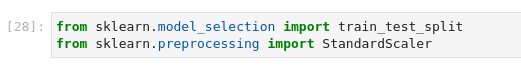


1. People with a higher number of major vessels with blockage have higher chances of suffering from heart disease.



## Algorithms

Importing Libraries



A screenshot of a computer

Description automatically generated

Testing DataSet



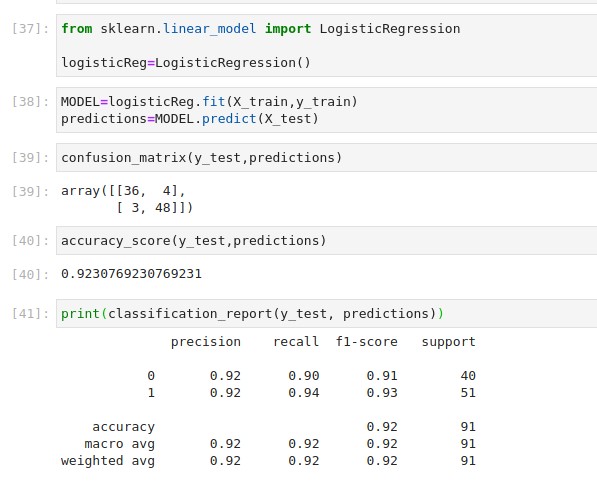
For X we have considered ‘age’, ‘oldpeak’, ‘trestbps’, ‘chol’, ‘thalach’ and y has a target, which is basically an output variable.

We have Compared the following classifiers:

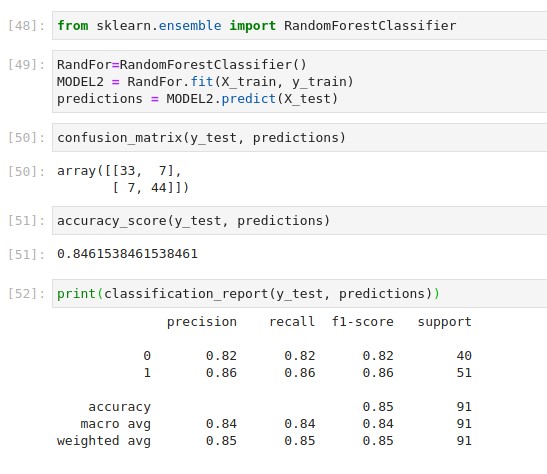
1. Logistic Regression
   1. Logistic regression works well for cases where the dataset is linearly separable.
   2. Logistic regression not only gives a measure of how relevant an independent variable is, but also tells us about the direction of the relationship.
2. Random Forest
   1. This algorithm can be used for both classifications and regression tasks. It provides higher accuracy through cross validation.
   2. This classifier will handle the missing values and maintain the accuracy of a large proportion of data
3. Decision tree
   1. Decision tree implicitly perform variable screening or feature selection. Non-linear relationships between parameters do not affect the tree performance.
   2. Another main reason was that decision trees are very easy to interpret.

Implementing the algorithms:

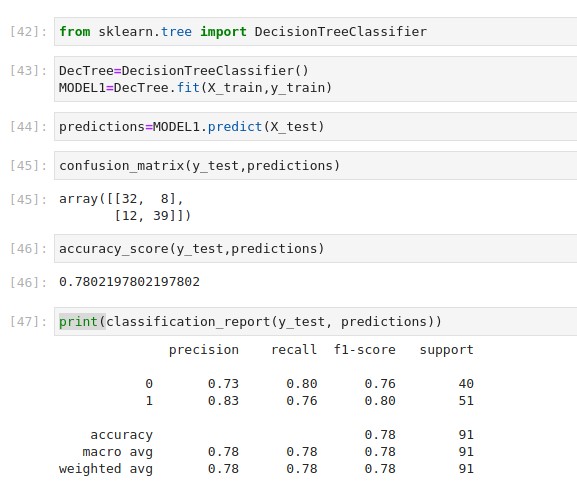
**Logistic Regression**:



**Random Forest:**



**Decision Tree:**

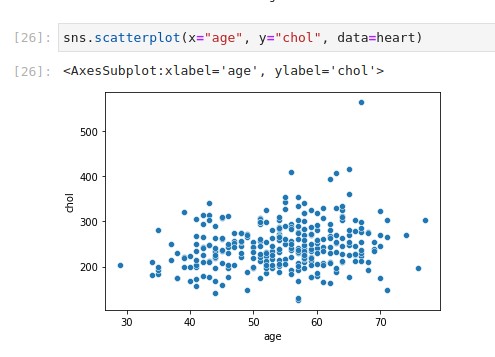
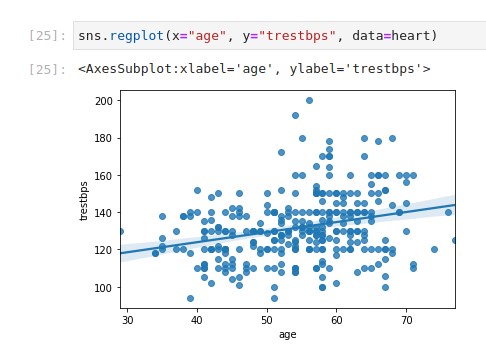


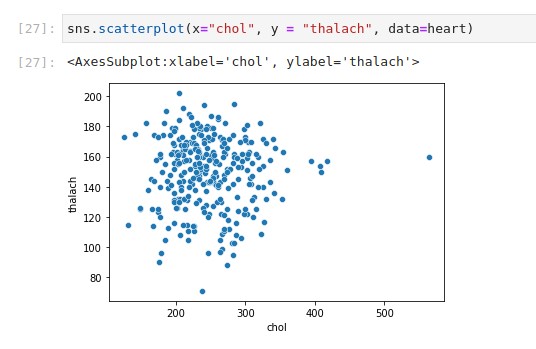
Detailed Results:-

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Decision Tree** | **Logistic Regression** | **Random Forest** |
| **Accuracy** | **0.78** | **0.92** | **0.85** |
| **Precision** | **0.78** | **0.92** | **0.84** |
| **Recall** | **0.78** | **0.92** | **0.84** |
| **f1 score** | **0.78** | **0.92** | **0.84** |

**Our Choice of algorithm: Logistic Regression**:

We wanted a classification algorithm which would give us a value of 0 or 1. Logistic regression is used to predict an outcome variable that is in numerical form, predictor variables that are either continuous or in categorical form. This was exactly our case as many of our predictor variables were numerical. Logistic regression deals with the problem by using a logarithmic transformation, called logit function on the outcome variable which allows us to model a nonlinear relation in a linear way. It expresses the linear regression equation in logarithmic terms. Also, from the graphs between target(output variable) and most correlated input variables, As there we can see that it has only one major cluster, we were expecting the best accuracy to be on Logistic Regression, we tried random forest and decision trees but the best accuracy was on Logistic Regression only.





## References:-

1. Stackoverflow.com 2. Researchgate.net

1. Careerfoundry.com
2. Official documentations of Numpy, Pandas, Seaborn, Matplotlib
3. Medium.org
4. Kaggle
5. archive.ics.uci.edu/ml/datasets/heart+disease