HEART DISEASE PREDICTION USING MACHINE LEARNING

INTRODUCTION

Heart disease is an umbrella term used to describe a range of conditions such as blood vessel diseases, coronary artery diseases, heart defects and many more. The term is often used interchangeably with 'Cardiovascular Disease', though cardiovascular disease (CAD) is a specific class of disease that includes heart attacks, chest pains, angina, stroke etc. Heart disease continues to be the leading cause of death globally, according to the Annual Heart Disease and Strokes Statistics Update from the American Heart Association. For the purpose of providing appropriate results and making effective decisions on data, some advanced data mining techniques are used.

ABSTRACT

Health care industries collect huge amounts of data using which patterns can be established to gather more information. Here, we use data collected from a hospital in Cleveland from the UCI repository. Machine Learning proves to be effective in making decisions and predictions from large quantities of data. I will be applying and comparing various approaches of Machine Learning to help discover whether one suffers from heart disease or not, using the mentioned dataset. It enables relationships between medical factors related to heart disease and mathematical patterns, to be established.

DATA COLLECTED

The dataset consists of information pertaining to 303 individuals. There are 14 columns, each serving as a different parameter used for prediction.

1. **Age**: displays the age of the individual.
2. **Sex**: displays the gender of the individual under the following format:
   1 = male
   0 = female
3. **Chest-pain type**: displays the type of chest-pain experienced by the individual under the following format:
   1 = typical angina
   2 = atypical angina
   3 = non-anginal pain
   4 = asymptomatic
4. **Resting Blood Pressure**: displays the resting blood pressure value of an individual in mmHg (unit)
5. **Serum Cholesterol**: displays the serum cholesterol in mg/dl (unit)
6. **Fasting Blood Sugar**: compares the fasting blood sugar value of an individual with 120mg/dl.
   If fasting blood sugar > 120mg/dl then: 1 (true)
   else: 0 (false)
7. **Resting ECG**: displays resting electrocardiographic results
   0 = normal
   1 = having ST-T wave abnormality
   2 = left ventricular hypertrophy
8. **Max heart rate achieved**: displays the max heart rate achieved by an individual.

9. **Exercise induced angina**:
   1 = yes
   0 = no

10. **ST depression induced by exercise relative to rest**: displays the value which is an integer or float.

11. **Peak exercise ST segment**:
   1 = upsloping
   2 = flat
   3 = downsloping

12. **Number of major vessels (0–3) coloured by fluoroscopy**: displays the value as integer or float.

13. **Thalassemia:** displays the thalassemia:
   3 = normal
   6 = fixed defect
   7 = reversible defect

14. **Diagnosis of heart disease**: displays whether the individual is suffering from heart disease or not:
   0 = absence
   1, 2, 3, 4 = present.

Index	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
1	67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
2	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
3	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
4	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
5	56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
6	62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
7	57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
8	63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
9	53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
10	57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
11	56	0	2	140	294	0	2	153	0	1.3	2	0	3	0
12	56	1	3	130	256	1	2	142	1	0.6	2	1	6	2
13	44	1	2	120	263	0	0	173	0	0	1	0	7	0
14	52	1	3	172	199	1	0	162	0	0.5	1	0	7	0
15	57	1	3	150	168	0	0	174	0	1.6	1	0	3	0
16	48	1	2	110	229	0	0	168	0	1	3	0	7	1
17	54	1	4	140	239	0	0	160	0	1.2	1	0	3	0
18	48	0	3	130	275	0	0	139	0	0.2	1	0	3	0

The mentioned parameters were taken into consideration due to the following:

1. **Age**: Age is the most important risk factor in developing cardiovascular or heart diseases, with approximately triple the risk with the passing of each decade of life. Coronary fatty streaks may begin forming in adolescence. It is estimated that 82 percent of people who die of coronary heart disease are 65 and older. Additionally, the risk of stroke doubles every decade after age 55.

2. **Sex**: Men are at greater risk of heart disease than pre-menopausal women. Post menopause, it has been argued that a woman's risk is similar to a man's, although more recent data from the WHO and UN disputes this. If a female has diabetes, she is more likely to develop heart disease than a male with diabetes.

3. **Angina (Chest Pain)**: Angina is chest pain or discomfort caused when the heart muscle does not receive sufficient oxygen-rich blood. One might experience the sensation of pressure or squeezing in

the chest. The discomfort may also occur in one's shoulders, arms, neck, jaw, or back. Angina pain may even feel like indigestion.

4. **Resting Blood Pressure**: Over time, high blood pressure can damage arteries that feed the heart. High blood pressure that occurs with other conditions, such as obesity, high cholesterol, or diabetes, increases the risk even more.
5. **Serum Cholesterol**: A high level of low-density lipoprotein (LDL) cholesterol (the "bad" cholesterol) is most likely to narrow arteries. A high level of triglycerides, a type of blood fat related to one's diet, also increases the risk of a heart attack. However, a high level of high-density lipoprotein (HDL) cholesterol (the "good" cholesterol) lowers one's risk of a heart attack.
6. **Fasting Blood Sugar**: Not producing enough of a hormone secreted by the pancreas (insulin) or not responding to insulin properly causes the body's blood sugar levels to rise, increasing one's risk of a heart attack.
7. **Resting ECG**: For people at low risk of cardiovascular disease, the USPSTF concludes with moderate certainty that the potential harms of screening with resting or exercise ECG equal or exceed the potential benefits. For people at intermediate to high risk, current evidence is insufficient to assess the balance of benefits and harms of screening.
8. **Max heart rate achieved**: In a study, it was found that the increase in cardiovascular risk, associated with the acceleration of heart rate, was comparable to the increase in risk observed with high blood pressure. It has been shown that an increase in heart rate by 10 beats per minute was associated with an increase in the risk of cardiac death by at least 20%, and this increase in the risk is similar to the one observed with an increase in systolic blood pressure by 10 mm Hg.
9. **Exercise induced angina**: The pain or discomfort associated with angina usually feels tight, varies from mild to severe and can feel like gripping or squeezing. Angina is usually felt in the centre of one's chest but may spread to either or both of the shoulders, or the back, neck, jaw, or arms. It can even be felt in one's hands.

   Types of Angina:

   a. Stable Angina / Angina Pectoris

   b. Unstable Angina

   c. Variant (Prinz metal) Angina

   d. Microvascular Angina.
10. **Peak exercise ST segment**: A treadmill ECG stress test is considered abnormal when there is a horizontal or down-sloping ST-segment depression ≥ 1 mm at 60–80ms after the J point. Exercise ECGs with up-sloping ST-segment depressions are typically reported as an 'equivocal' test. In general, the occurrence of horizontal or down-sloping ST-segment depression at a lower workload (calculated in METs) or heart rate indicates a worse prognosis and higher likelihood of multi-vessel disease. The duration of ST-segment depression is also important, as prolonged recovery after peak stress is consistent with a positive treadmill ECG stress test. Another finding that is highly indicative of significant CAD is the occurrence of ST-segment elevation > 1 mm (often suggesting transmural ischemia); these patients are frequently referred urgently for coronary angiography.

MACHINE LEARNING APPROACH

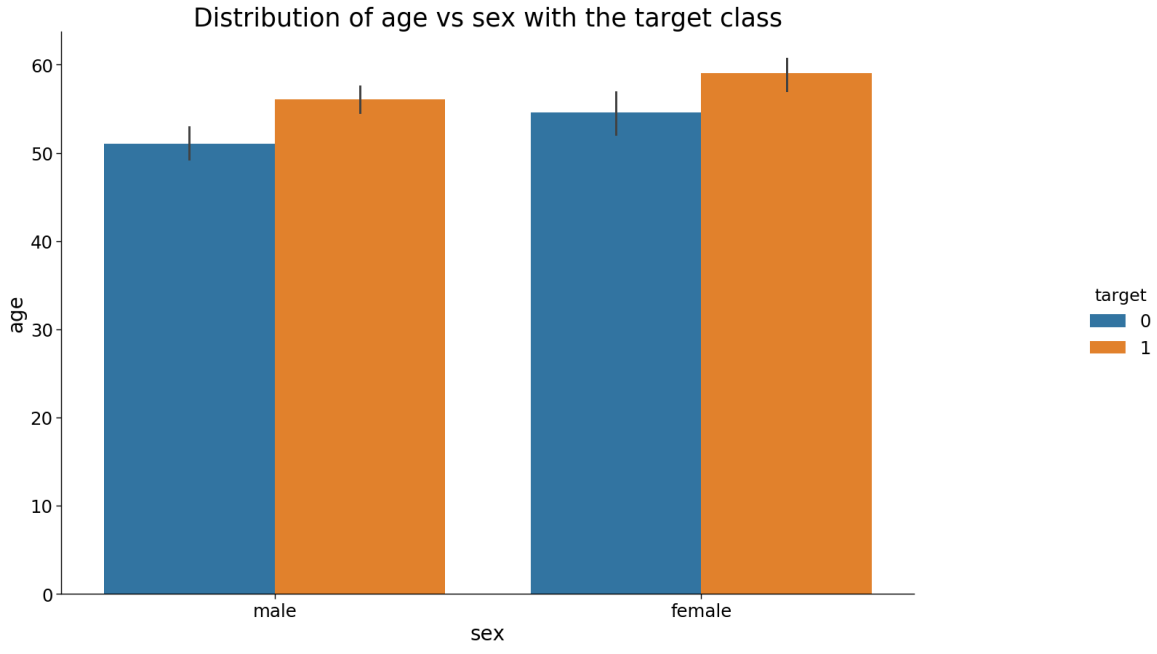
The following algorithms were applied in this project-

- SVM
- Naive Bayes
- Logistic Regression
- Decision Tree
- Random Forest

DATA ANALYSIS

We plot the data acquired to simplify it by visualising. ‘0’ indicates that the individual investigated was not suffering from heart diseases; ‘1’ indicates that the individual is diagnosed with a certain heart related disease. We see that most people who are suffering are 58 years old, followed by those who are 57. A majority of individuals above the age of 50 are diagnosed with heart diseases.

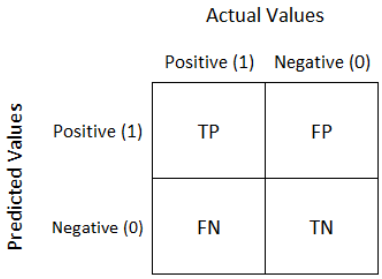
Next, we plot the age and gender of the individuals.



It is evident that females diagnosed with the disease are much older than males diagnosed with the same.

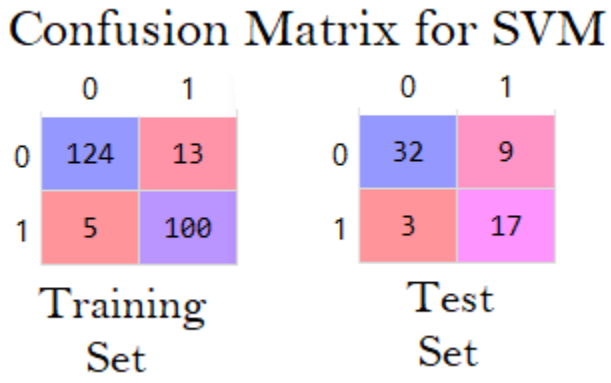
TRAINING DATASET

The confusion matrix was used as the evaluation metric.



PREDICTIONS PROCURED

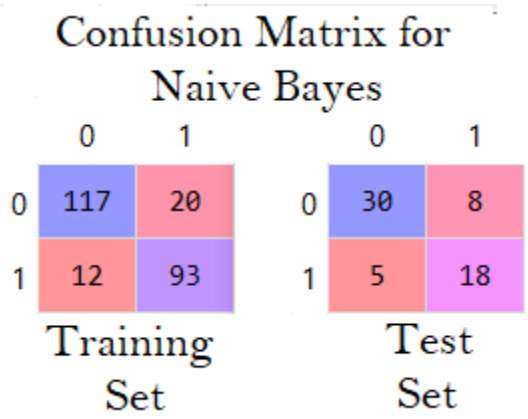
- **SVM**



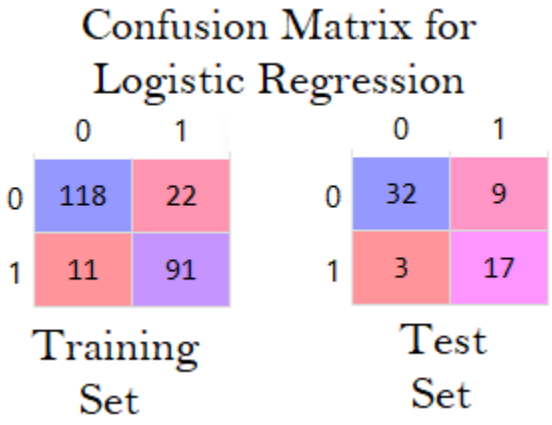
Accuracy for SVM for training set = ((124+100)/(5+13+124+100))*100 = 92.51%
Accuracy for SVM for test set = 80.32%

Similarly, we see the matrices developed for each classifier.

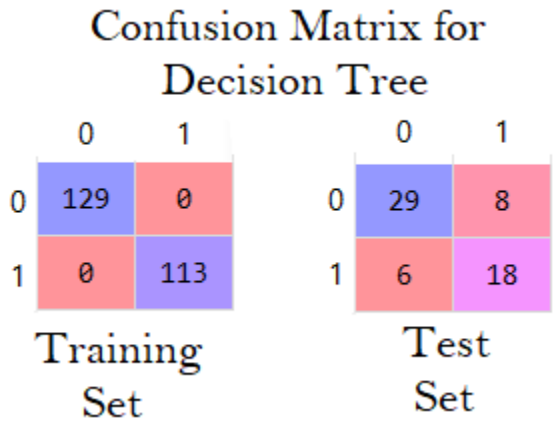
- **NAÏVE BAYES**



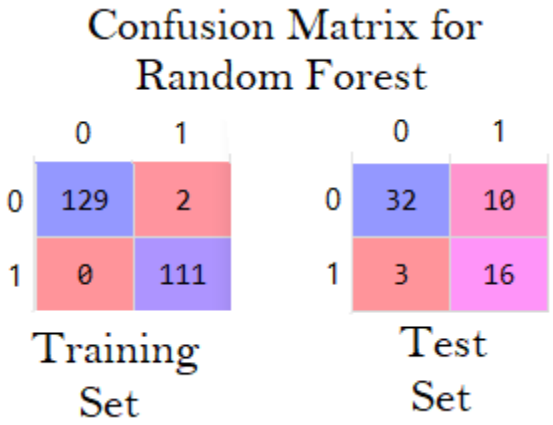
- **LOGISTIC REGRESSION**



- **DECISION TREE**



- **RANDOM FOREST**



Summarising the system of prediction and comparing using Machine Learning, I have listed the accuracies of all the classifiers.

```
Accuracy for training set for svm = 0.9256198347107438
Accuracy for test set for svm = 0.8032786885245902
```

```
Accuracy for training set for Naive Bayes = 0.8677685950413223
Accuracy for test set for Naive Bayes = 0.7868852459016393
```

```
Accuracy for training set for Logistic Regression = 0.8636363636363636
Accuracy for test set for Logistic Regression = 0.8032786885245902
```

```
Accuracy for training set for Decision Tree = 1.0
Accuracy for test set for Decision Tree = 0.8032786885245902
```

```
Accuracy for training set for Random Forest = 0.9917355371900827
Accuracy for test set for Random Forest = 0.7704918032786885
```

The highest accuracy recorded for the test set was achieved by Logistic Regression and SVM, equal to 80.32%. The highest accuracy for the training set is 100%, achieved by Decision Tree.
The algorithms were implemented using default parameters.

BIBLIOGRAPHY

1. https://archive.ics.uci.edu/ml/datasets/Heart+Disease
2. https://archive.ics.uci.edu/ml/index.php
3. https://en.wikipedia.org/wiki/Cardiovascular_disease#:~:text=Cardiovascular%20disease%20(CVD)%20is%20a,known%20as%20a%20heart%20attack)
4. https://www.healio.com/news/cardiology/20210127/aha-heart-disease-remains-leading-cause-of-death-worldwide-trends-discouraging#:~:text=In%202019%2C%20the%20latest%20year,leading%20cause%20of%20death%20worldwide
5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5863635/#:~:text=The%20EHDPS%20predicts%20the%20likelihood,and%20patterns%2C%20to%20be%20established
6. https://www.ijert.org/heart-disease-prediction-using-machine-learning
7. https://www.kaggle.com/ronitf/heart-disease-uci
8. https://www.kaggle.com/ronitf/predicting-heart-disease
9. https://www.kaggle.com/cdabakoglu/heart-disease-classifications-machine-learning