Introduction

Heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease, heart rhythm problems (arrhythmias) and heart defects you're born with (congenital heart defects), among others.

The term "heart disease" is often used interchangeably with the term "cardiovascular disease". Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart's muscle, valves or rhythm, also are considered forms of heart disease.

Heart diseases have become a major concern to deal with as studies so that the number of deaths due to heart diseases have increased significantly over the past few decades in India, in fact it has become the leading cause of death in India even world. A study shows that from 1990 to 2021 the death rate due to heart diseases have increased around 40% from (150 to 220)/100000 deaths in India.Thus preventing Heart diseases has become more than necessary. Good data-driven systems for predicting heart diseases can improve the entire research and prevention process, making sure that more people can live healthy lives. This is where Machine Learning comes into play. Machine Learning helps in predicting the Heart diseases, and the predictions made are quite accurate.According to a news article, heart disease proves to be the leading cause of death for both women and men.

The article states the following :Machine learning (ML) proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry.

The Data

The dataset used in this article is the Cleveland Heart Disease dataset taken from the UCI repository.

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

Dataset

The dataset consists of 303 individuals data. There are 14 columns in the dataset, which are described below.

1. ***Age***: displays the age of the individual.
2. ***Sex***: displays the gender of the individual using the following format :
   1 = male
   0 = female
3. ***Chest-pain type***: displays the type of chest-pain experienced by the individual using the following format :
   1 = typical angina
   2 = atypical angina
   3 = non — anginal pain
   4 = asymptotic
4. ***Resting Blood Pressure***: displays the resting blood pressure value of an individual in mmHg (unit)
5. ***Serum Cholestrol***: 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 hyperthrophy
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) colored by flourosopy*** : displays the value as integer or float.
13. ***Thal*** : 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.

Why these parameters:

In the actual dataset, we had 76 features but for our study, we chose only the above 14 because :

1. **Age**: Age is the most important risk factor in developing cardiovascular or heart diseases,

with approximately a tripling of risk with each decade of life. Coronary fatty streaks can begin to form in adolescence. It is estimated that 82 percent of people who die of coronary heart disease are 65 and older. Simultaneously, the risk of stroke doubles every decade after age 55.

2. **Sex**: Men are at greater risk of heart disease than pre-menopausal women. Once past 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 your heart muscle doesn't get enough oxygen-rich blood. It may feel like pressure or squeezing in your chest. The discomfort also can occur in your 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 your heart. High blood pressure that occurs with other conditions, such as obesity, high cholesterol or diabetes, increases your 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 your diet, also ups your risk of a heart attack. However, a high level of high-density lipoprotein (HDL) cholesterol (the "good" cholesterol) lowers your risk of a heart attack.
6. **Fasting Blood Sugar**: Not producing enough of a hormone secreted by your pancreas (insulin) or not responding to insulin properly causes your body's blood sugar levels to rise, increasing your 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**: 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, gripping or squeezing, and can vary from mild to severe.
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 60–80 ms after the J point. Exercise ECGs with up-sloping ST-segment depressions are typically reported as an 'equivocal' test..

The Approach

In this project I use to classification models for classification :

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

Data Analysis

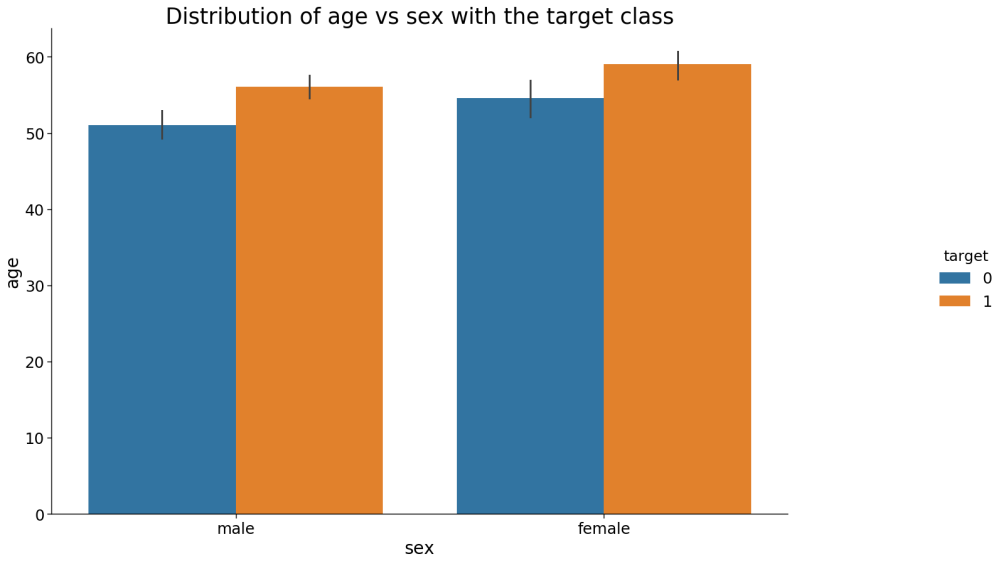
Let us look at the people's age who are suffering from the disease or not.

Here, target = 1 implies that the person is suffering from heart disease and target = 0 implies the person is not suffering.

We see that most people who are suffering are of the age of 58, followed by 57.

Majorly, people belonging to the age group 50+ are suffering from the disease.

Next, let us look at the distribution of age and gender for each target class.



Data Pre-Processing

The dataset contains 14 columns and 303 rows.

Let us check the null values

Finding the correlation among the attributes

In [9]:

```
plt.figure(figsize=(20,10))
sns.heatmap(df.corr(), annot=True, cmap='terrain')
```

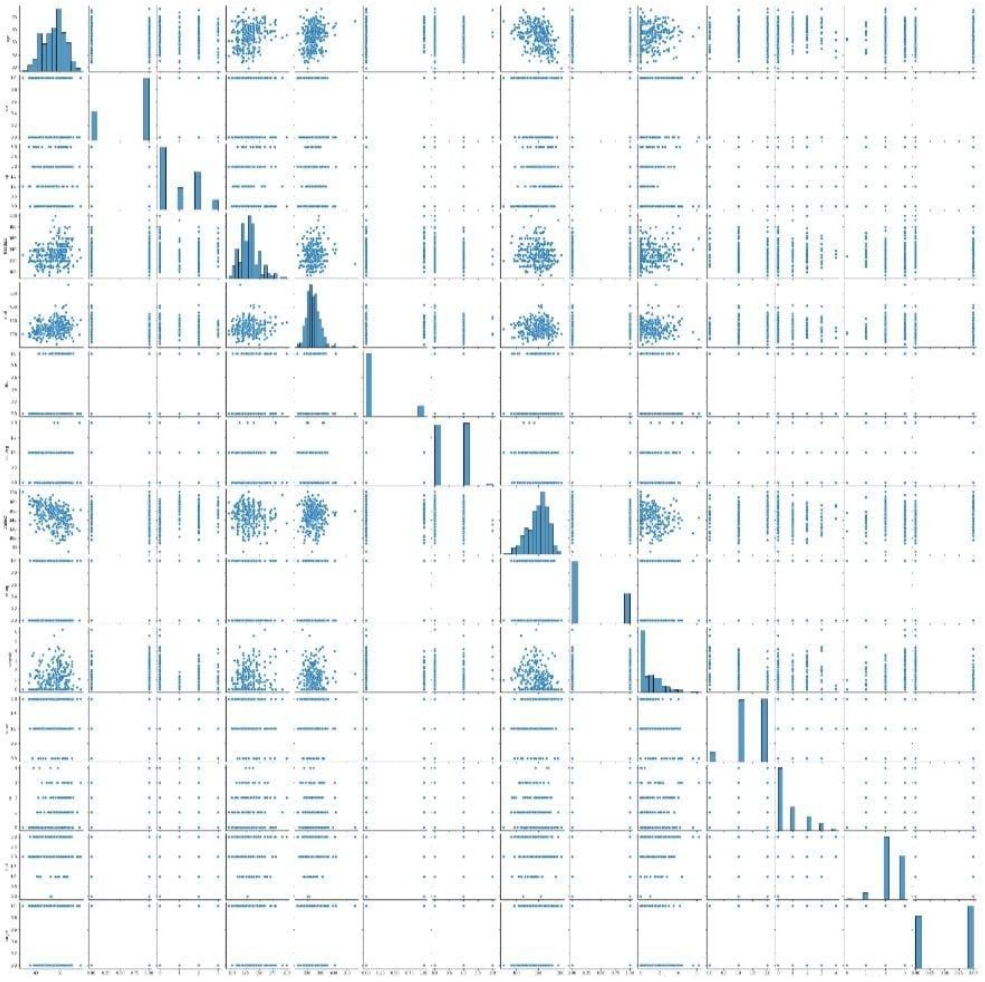
Out[9]:

<AxesSubplot:>

```
sns.pairplot(data=df)
```

Out[10]:

```
<seaborn.axisgrid.PairGrid at
0x210a25bd160>
```

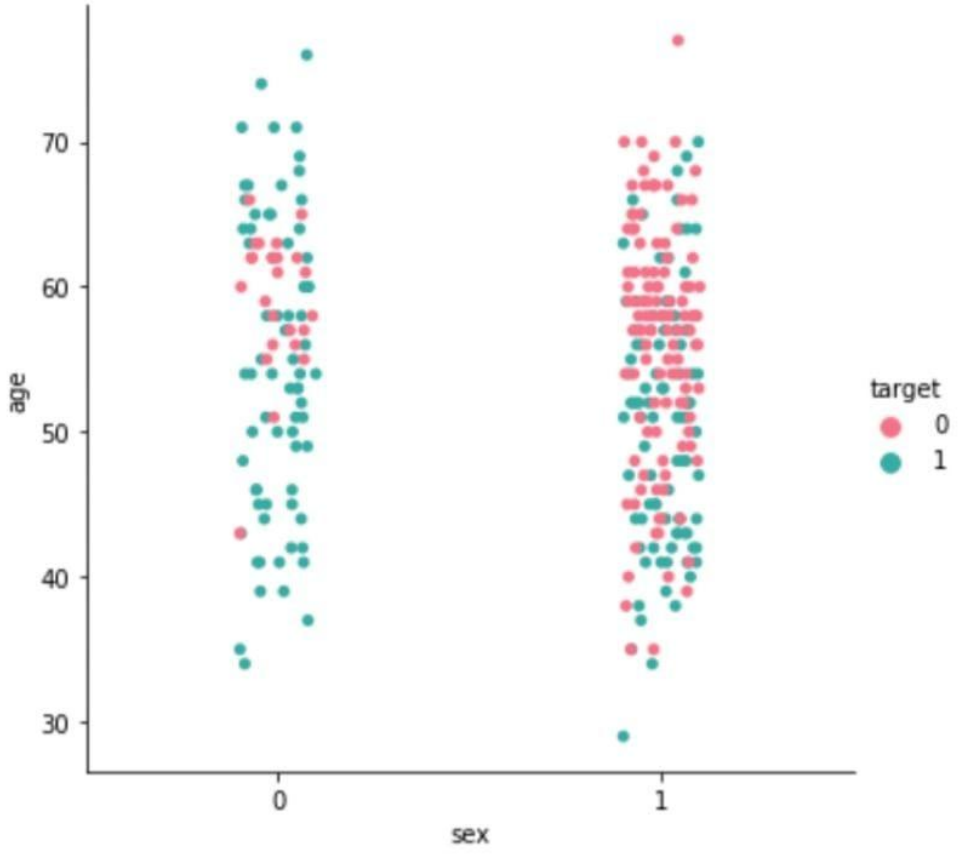


In [13]:

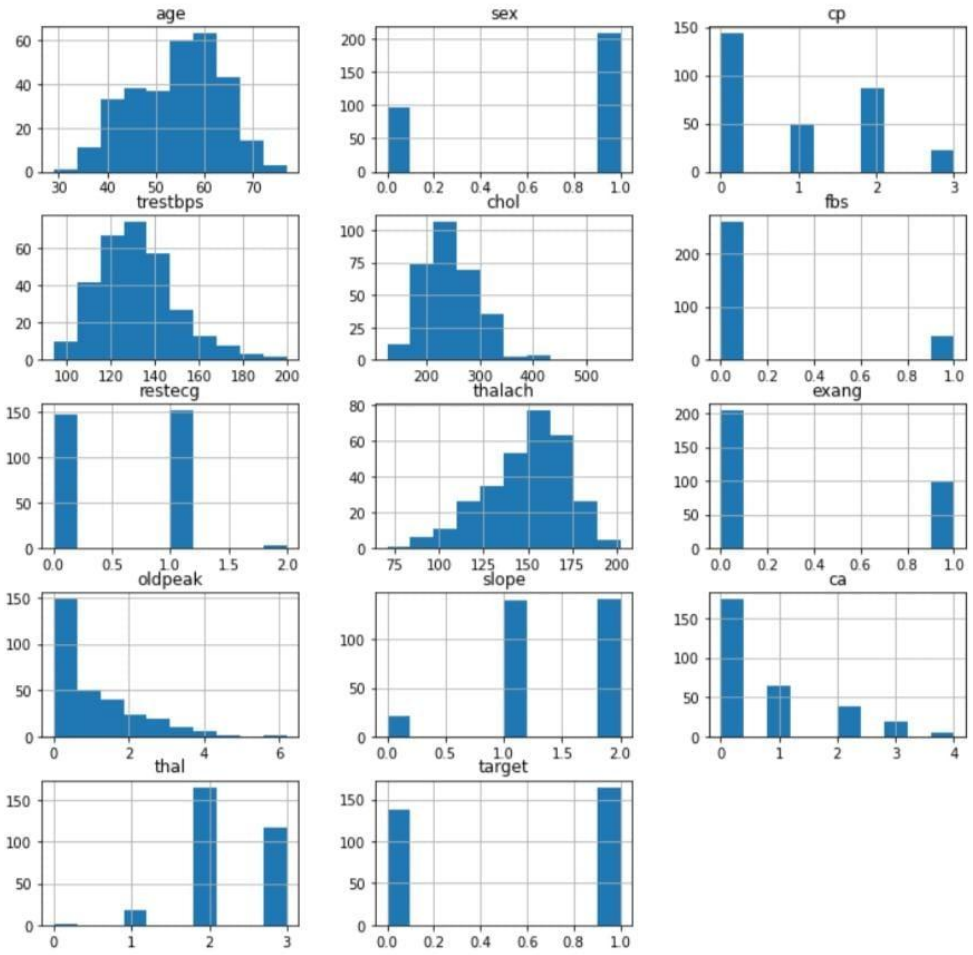
```
sns.catplot(data=df, x='sex',
y='age',  hue='target', palett
e='husl')
```

Out[13]:

```
<seaborn.axisgrid.FacetGrid at
0x210ac169520>
```



```
df.hist(figsize=(12,12), layout=(5,3));
```



```
Out[2]:
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          4
thal        2
target      0
dtype: int64
```

null values in each column of the data

We see that there are only 6 cells with null values with 4 belonging to attribute *ca* and 2 to *thal*.

As the null values are very less we can either drop them or impute them. I have imputed the mean in place of the null values however one can also delete these rows entirely.

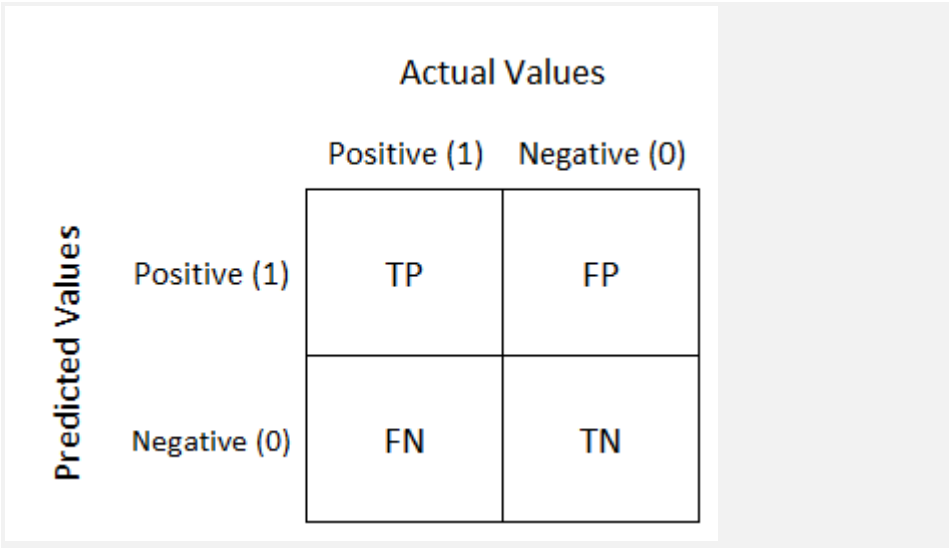
Now let us divide the data in the test and train set.

In this project, I have divided the data into an 80: 20 ratio. That is, the training size is 80% and testing size is 20% of the whole data.

Training

All the models discussed above are applied to get the results.

The evaluation metric used is the confusion matrix.

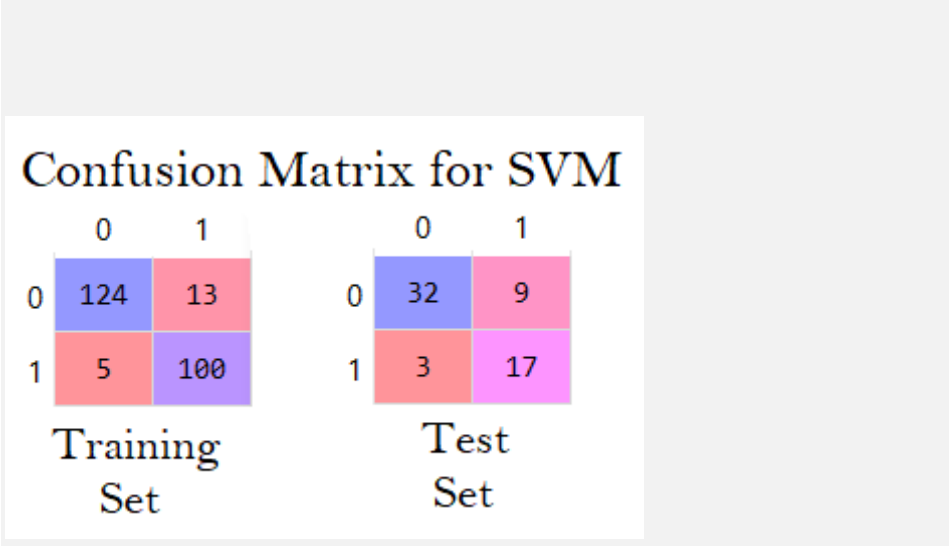


confusion matrix

The confusion matrix displays the correctly predicted as well as incorrectly predicted values by a classifier.

The sum of TP and TN, from the confusion matrix, is the number of correctly classified entries by the classifier.

SVM

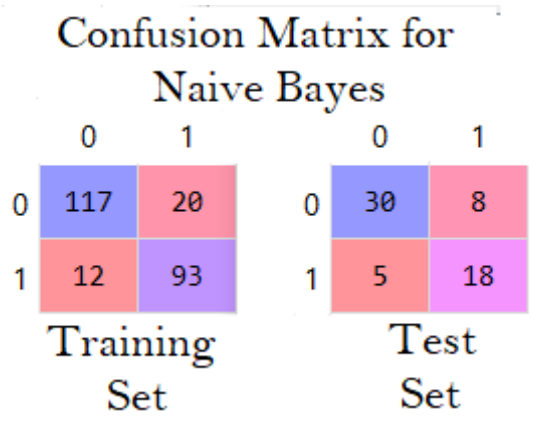


Accuracy for SVM for training set = ((124+100)/(5+13+124+100))*100 = 92.51%

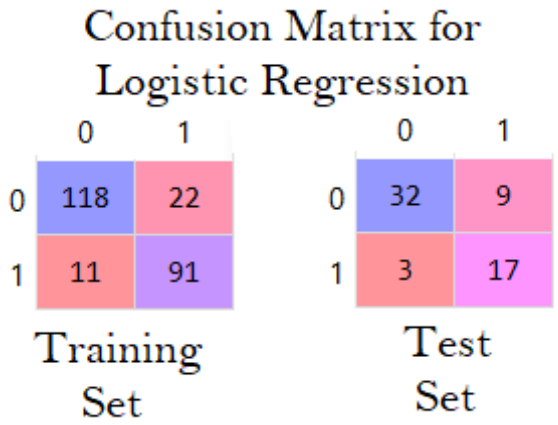
Accuracy for SVM for test set = 80.32%

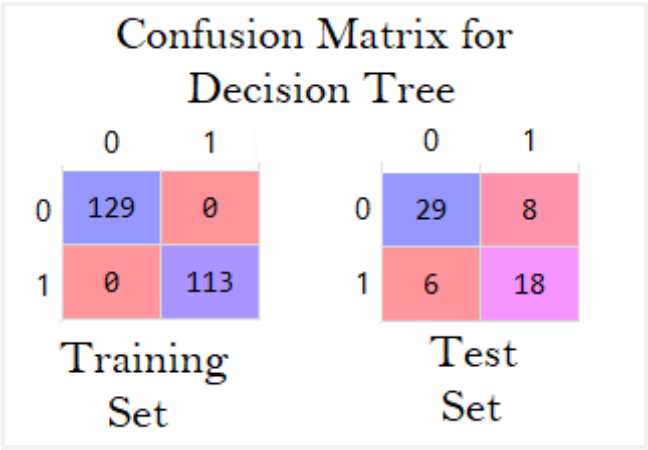
Similarly let us look at all the confusion matrices for each classifier.

Naive Bayes

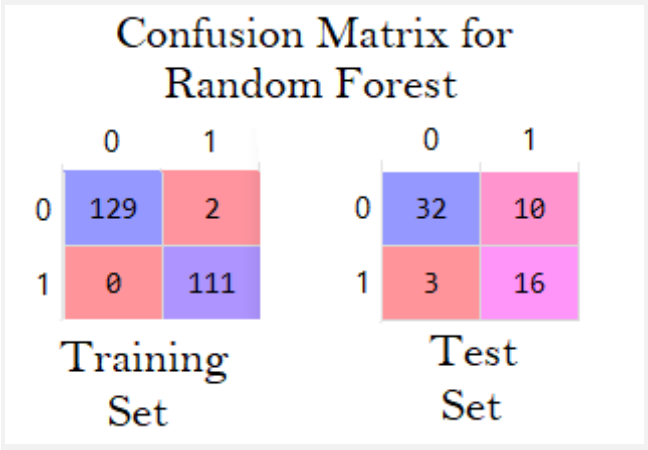


Logistic Regression





Random Forest



In [64]:

```
print('KNN :', accuracy_score
(y_test, prediction6))
print('lr :', accuracy_score(y
_test, prediction1))
print('dtc :', accuracy_score
(y_test, prediction2))
print('rfc :', accuracy_score
(y_test, prediction3))
print('NB: ', accuracy_score(y
_test, prediction4))
print('SVC :', accuracy_score
(y_test, prediction5))
```

```
KNN : 0.8351648351648352
lr : 0.9230769230769231
dtc : 0.7582417582417582
rfc : 0.8461538461538461
NB:  0.9010989010989011
SVC : 0.8791208791208791
```

Best accuracy is given by Logistic Regression : 92

followed by NB and Decision tree : 90

We see that the highest accuracy for the test set is achieved by Logistic Regression and SVM which is equal to 80.32%.

The highest accuracy for the training set is 100% achieved by Decision Tree.

The algorithms are implemented with the default parameters only.

Conclusion

Heart Disease is one of the major concerns for society today.

It is difficult to manually determine the odds of getting heart disease based on risk factors. However, machine learning techniques are useful to predict the output from existing data.