**Introduction**

Machine Learning is a process of automating the decision making through experience with data and performing some task with the data.In amchine learning we need data and code for creating the machine learning algorithms.Machine Learning Algorithms can learn patterns from historical data and also can find patterns from the data. According to this property we can classify the machine learning algorithms in to three broad categories

1. Supervised Learning

In this method , the data should contain the label according to each record.The label feature is important for this method , then only we can train the supervised learning algorithms.For example, we can say teacher taking the class according to each subject,at that time teacher will give the labelled data to their students. And we can further classify in to two.That is Regression and Classification.In regression the target variable should be continous and in classification , the label should be categorical.The supervised algorithms are SVM,Decision Tree,Logistic Regression etc.

1. Unsupervised Learning

Here , we have no label feature and no prediction.But here we are doing the grouping or clustering the data accor ding to similar features. That is we are creating the clusters according to the theory of similarity.The cluster should have high intra class similaritry and low inter class similarity.We are calculating the similarity using eucledian or manhattan distance formulas. The popular unsupervised algorithms are K-means clustering,DBSCAN,Apriori Algorithms etc.

1. Reinforcement Learning

In reinforcement learning we have different components to perform the algorithm.First one is Agent, an agent perform the task to the environment, for each state the agent will get the reward according to the environment feedback.

Example : Qlearning

In this project , we are focusing on the supervised methods like Regression and Classification.

For regression we are using the Life Expectancy data from kaggle data set([Data](https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who)). We need to predict the life expectancy of different countries according to the 21 independent variables.

For classification, we are using Heart Disease dataset([Data](https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?resource=download)),we need to classify the patients who have the heart disease or not.For that we have 13 independent variables to classify the target variable.

**Regression Analysis with Life Expectancy Data**

**Business Understanding**

From the life expectancy data , we need to predict the life expectancy of the countries according to their characterstics of the country like Mortality rate, country status , GDP, Healthness etc. Business want to know which country has high expectancy in the future, according to this prediction the business can have decide the new ventures about their business. The aim of the analysis is to predict the life expectancy of the countries in the next year(2016) according to historical data.

**Data Understanding**

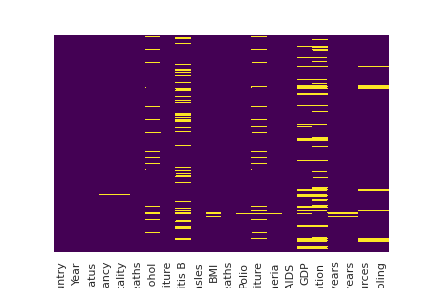
|  |  |
| --- | --- |
| Column Name | Description |
| Country | Country Names |
| Year | Years (15Years) |
| Status | Is developed or developing country? |
| Life Expectancy | Taregt Variable , life expectancy of the country.Life expectancy in age |
| Adult Mortality | Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population) |
| Infant Deaths | Number of Infant Deaths per 1000 population |
| Alcohol | Its recorded per capita , consumption in litres |
| Percentage expenditure | Percentage on health as a percentage of gross domestic product per capita(%) |
| Hepatitis B | Immunization coverage of one year old(%) |
| Measles | Number of reported cases per 1000 population |
| BMI(Body Mass Index) | Average BMI of entire population |
| Under-five deaths | Deaths under five year old per 1000 poulation |
| Polio | Polio immunization coverage under one year old(%) |
| Total Expenditure | General government total expenditure(%) |
| Diphtheria | Diphtheria tetanus toxoid and pertusis immunization coverage among 1 year old (%) |
| HIV/AIDS | Deaths per 1000 live births (0-4 year old) in % |
| GDP | GDP per Capita |
| Population | Population of the country |
| Thinness 1-19 years | Prevalence of thinness among children and adolescents for age 10 to 19(%) |
| Thinness 5-9 years | Prevalence of thinness among children for age 5 to 9 years (%) |
| Income composition of resources | Human development index in terms of income composition of resources |
| Schooling | Number of years of schooling(years |

Here we have 21 feature columns with 2939 records to predict the life expectancy of the country in future. Here we have 15years of data for each for 183 countries.That is range from 2000 to 2015.So for using this data for modelling we need to do the exploratory data analysis and we need to prepare final data.

**Data Preparation**

In Data preparation, we need to clean the data and prepare the data for model. For that we need do data exploration analysis as following:

1. Checking Missing Values and Finding Outliers
2. Correlation
3. Uni Variate Analysis
4. Bivariate analysis
5. Creating New feature from the Country Column
6. **Checking Missing Values and Finding Outliers**



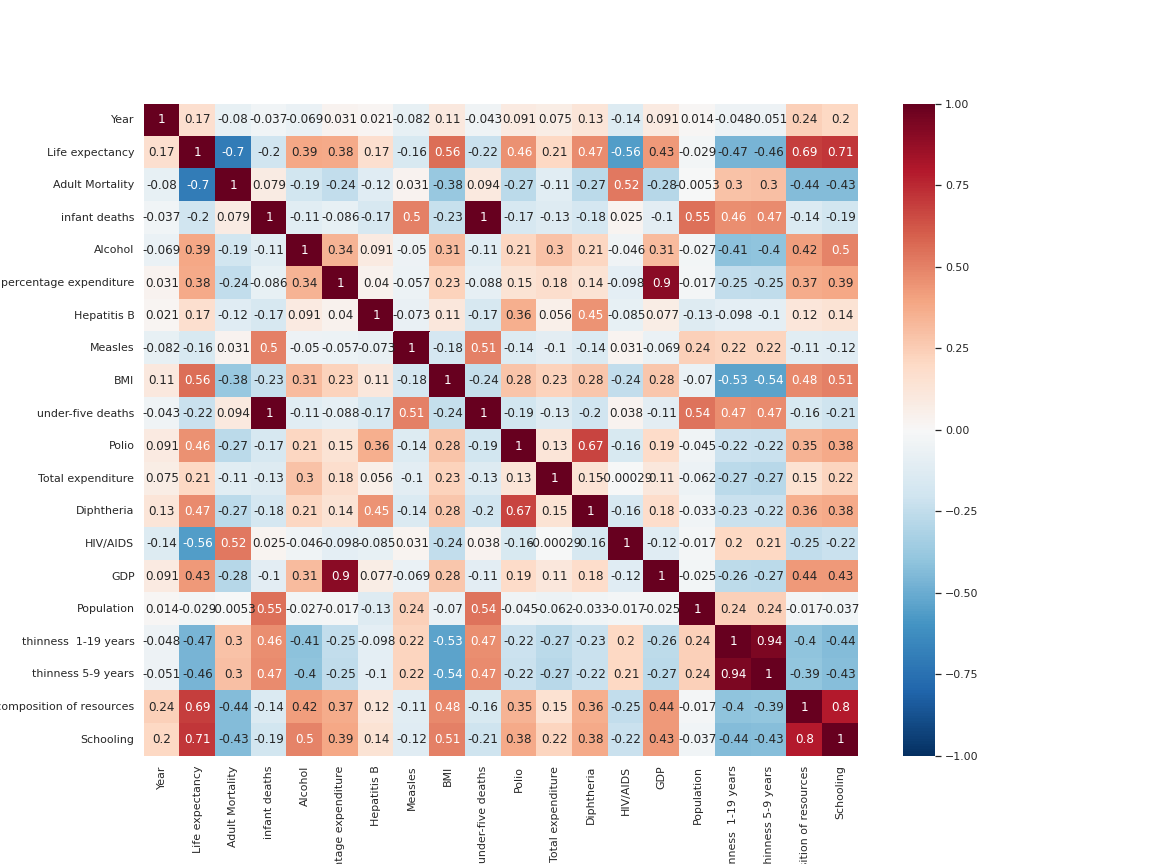
We have missing or null values in our feature and target columns.We have missing values in 14 columns. The column with missing values are the following:

'Life expectancy', 'Adult Mortality', 'Alcohol', 'Hepatitis B', 'BMI', 'Polio', 'Total expenditure', 'Diphtheria', 'GDP', 'Population', 'thinness 1-19 years', 'thinness 5-9 years', 'Income composition of resources', 'Schooling'

For treating the missing values we are using box plot and Inter Quartile range method. According to box plot we can visualize the data with outliers. Using IQR test, we can find the how much outliers we want to ignore from the data. And we filled the variable has null values according to their central tendancy measures like mean, median and mode.For discrete quantitative variable we used median of the data for filling the null values. And for the continous variable we are using the mean of the column. We have only 10 outliers with Life expectancy column. That is our target column. We have no Outliers in Alcohol variable , but we have the missing values. We used median value for the replacement of null values. And we can remove the records having null values, but we have only 2938 records, so we are not removing the data, it leads to data loss.And we are doing the efficient treatment with the data.

1. Correlation

After removing the missing values, we can find the correlation between dependant and independent variables using heatmap with correlation matrix.

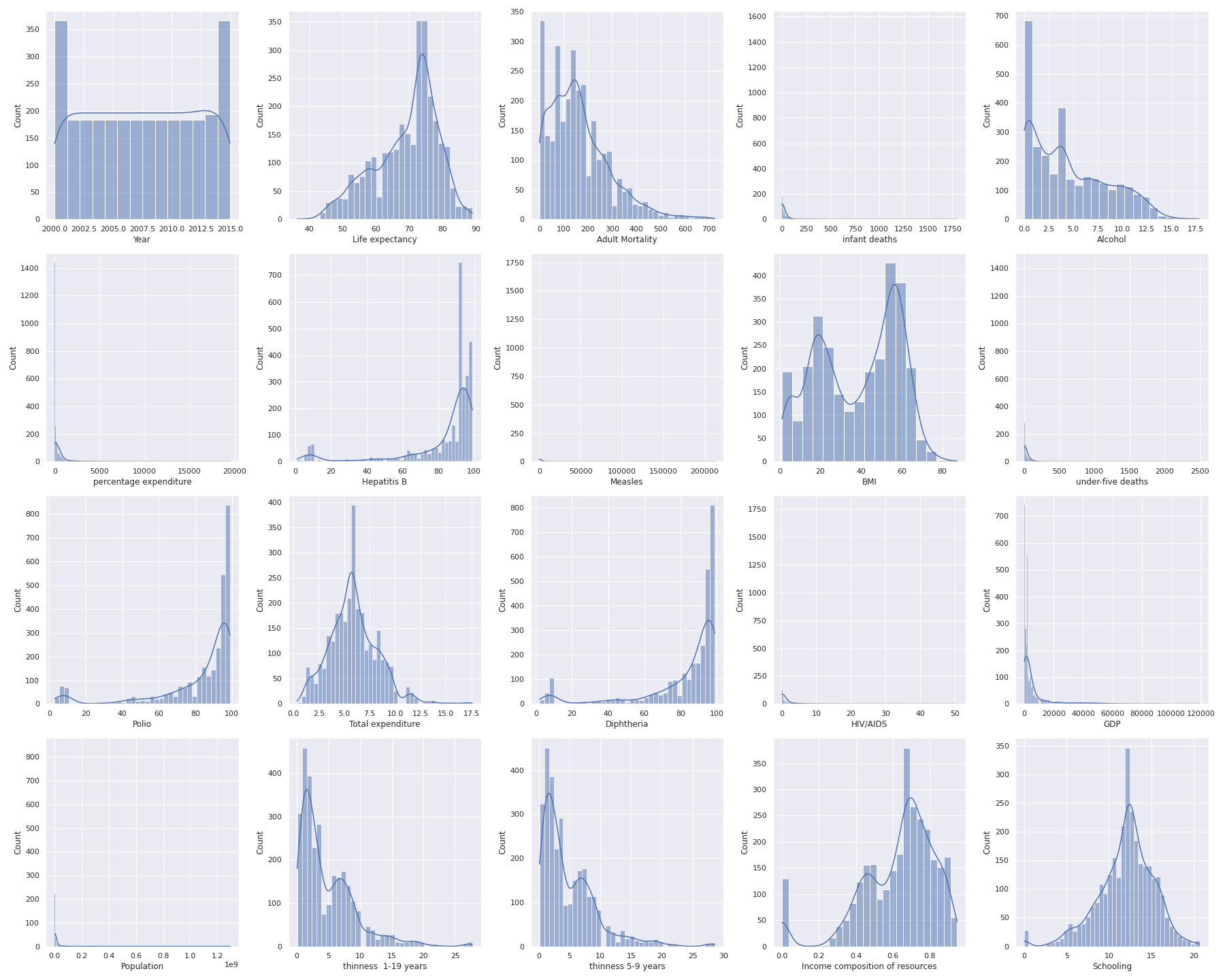


From the correlation plot we have the following insights about the data.

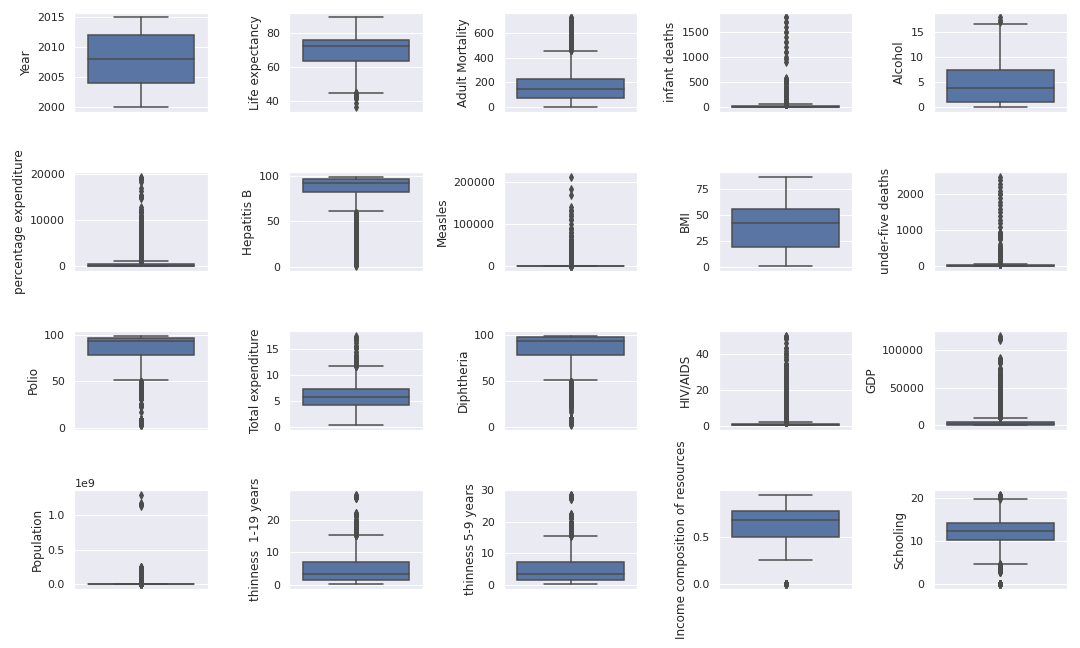
* We have Multi- collinearity in between under five-deaths an infant deaths independent variables. Also in between GDP and percentage of expenditure.So we can not use multi-collinear varibles. It will affect the performance of the model.So that, we used infant deaths and GDP variables for the model. And also we have selected schooling and thinness 1-19 years for the model to remove the multi-collinearity from the data.
* Also, we have correlated varibles like schooling,income composition of resources and BMI have more than 50% positive correlation with dependant variable.
* HIV and thinness 1-19 years have more negative correlation with dependant variable life expectancy.
* And we have no variable with no correlation with dependant variable, we have small percent positive or negative correlation.

1. **Univariate Analysis**

In univariate analysis, we are analyse the categorical and numerical features with plots, We have different types of univariate plots like histogram, box plot for numerical analysis and Count plot and Pie chart for categorical data analysis.

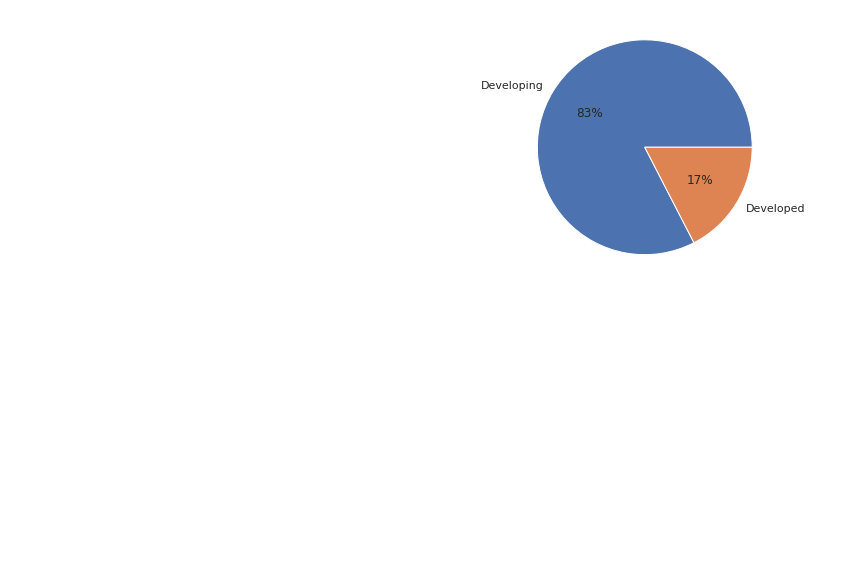


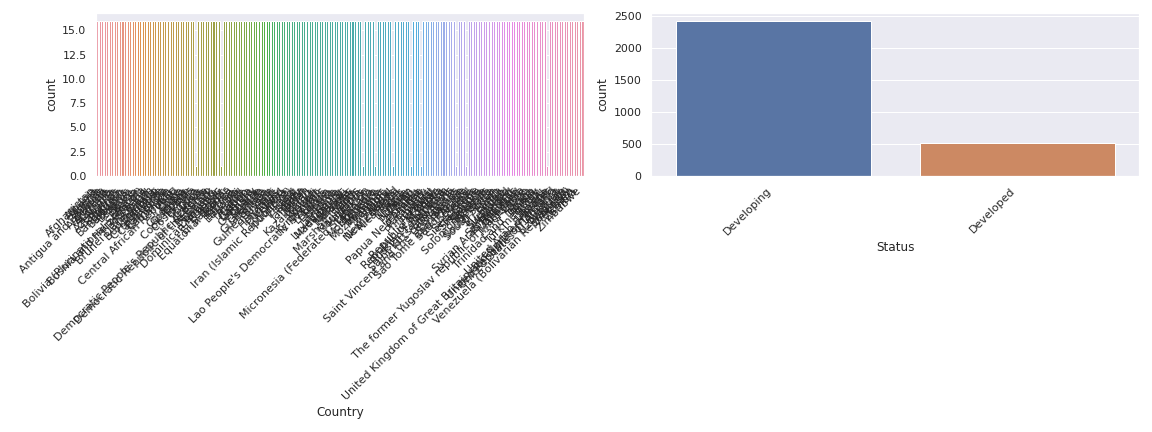
From the Above histograms, we have life expectancy has in Normal distribution.Also Schooling,Total expenditure have normal distributions.



From the above box plot, we can see the oultliers and what is the inter quartil range aswell as median of the numerical variables. And the varibale Population and GDP have high number of outliers and the target variable and Alcohol have only less number outliers.outliers, At last, BMI variable has no outliers.

Then, for categorical variables we have Pie chart and Count plot to explore the categorical data.And we have onle two categorical variables like country and Status. So, in status variable we have the data about country status like developed country or developing country.

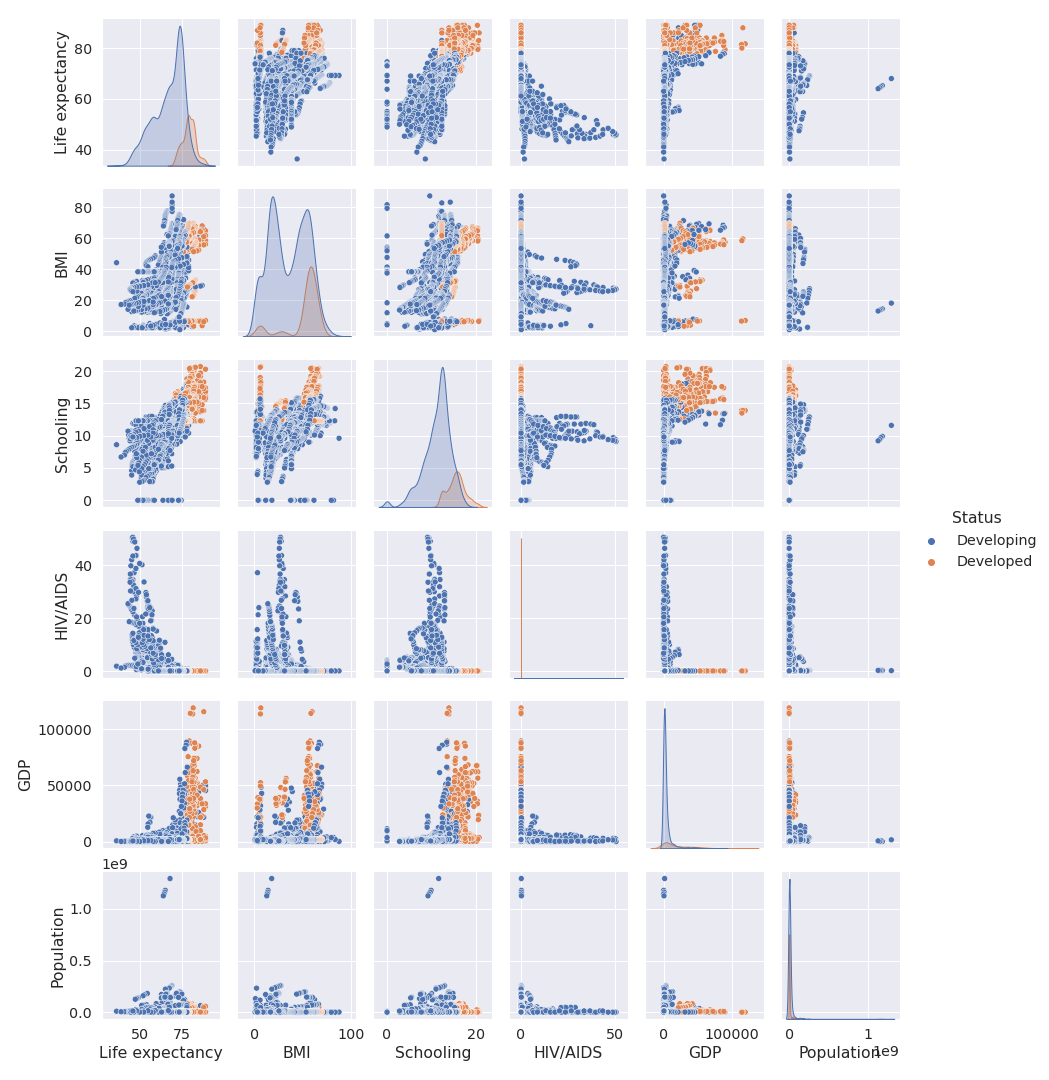




From the above plots we can say , The status variable has only two values Developing and Developed.In which have more count for developing. So we have more developing countries rather than developed.

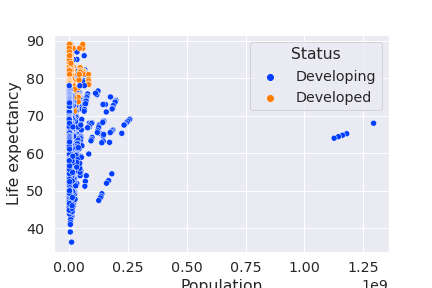
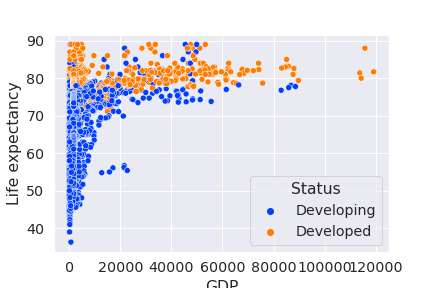
1. **Bivariate Analysis**

In Bivariate analysis, we are using two variables for the analysis.And we are finding the relationship between the two variables with the plots like scatter plot, pair plot,joint plot and line plots.After bivariate analysis we will have good understanding about our data and we can proceed with data pre-processing.



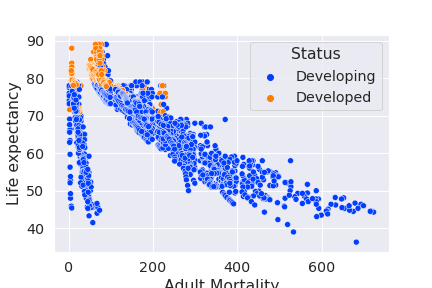
Firstly, we have pair plot for analysing the relation ship betwenn two variables with combined plot. The diagonal graphs are relationship between the variable itself with kde. From the graph we have negative correlation betwenn life expectancy and HIV variables. And Schooling and Life expectancy have positive relation betwenn the values. That means The independent variables like HIV, schooling have more imprtance to predict the life expectancy.

And we have scatter plot to determine the relationship with two any two numerical variables from the data.



Here , we are showing the relation ship between Population and GDP with life expectancy variable. From the first graph, developed countries has high expectancy when we have high GDP as well as low GDP compared to developing countries

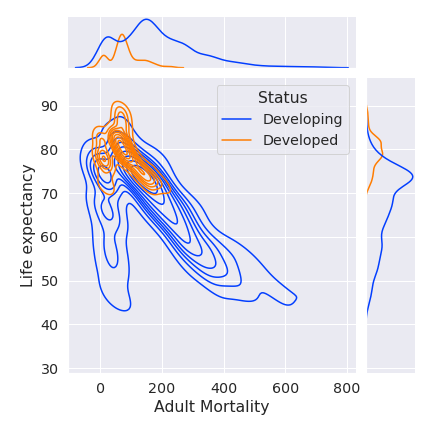
In second graph, Population has no measurable relation ship with life expectancy.And also we have good relationship between diphtheria, Polio with target variable.



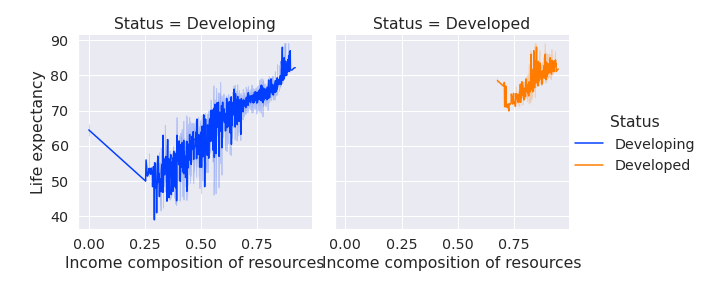
From this graph, we have high negative correlation between Adult mortality and Life expectancy.This feature also important for the model to predict the life expectancy.



The above graph shown the multi-collinearity between GDP and Percentage of expenditure, So we removed percentage of expenditure column.



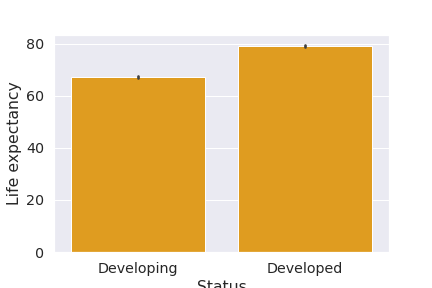
Also we have joint plot to show the relationship between the variables. And we have combined plots in a single graph,so we can analyze in better way.



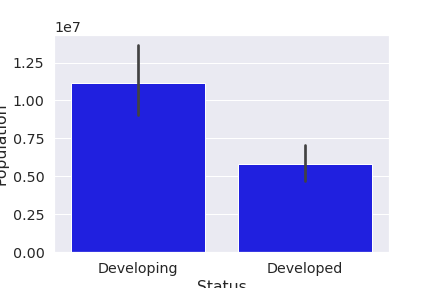
Here, we are plotting line plot for showing the relationship between income composition of resources and life expectancy.We have good relationship with them , and we sowed separate graph for according to the status column.

Also, we have bivariate analysis plots for categorical data like bar chart and point plot. In this analysis, we are using one categorical variable and one numerical variable, because we have only one categorical variable from the data that is Status variable.

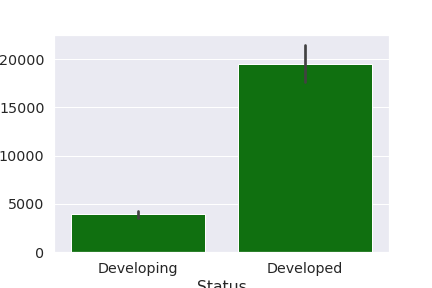
Here we are showing some graphs between Status and the numerical variable relationship.



Here, we have high expectency in developed countries, But not a big difference.



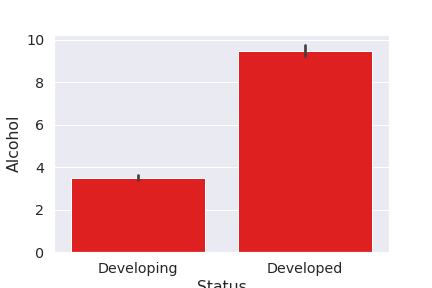
In developing countries have more population, because we have only less number of developed countries.



Here, we are plotting between GDP and Status, So the developed countries have high gdp value compared to developing countries, that is a normal thing.

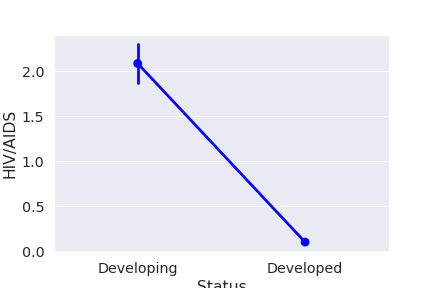


Polio is distributed in almost same way for developing and developed countries.

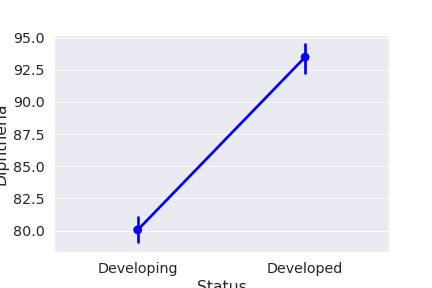


Alcohol usage more in developed countries, because they have high income or revenue with people.

And we have Point plots for describing the categorical variables.



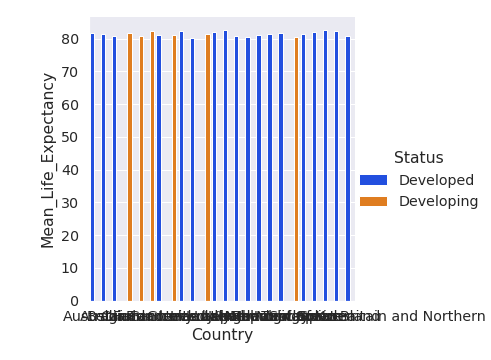
This shows Developed countries have almost zero number of HIV cases.



Here, more diphtheria in developed countries and less in developing countries.

1. **Creating New feature from the Country Column**

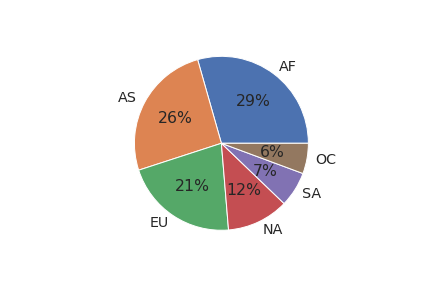
We have minimum 16 records for each countries in the data, Because of we take 16 year data of each country for the analysis. So that we need to get the average life expectancy of the countries by countries and status variable. And we took countries have average life expectancy greater than 80 and we got 24 countries have more than 80% life expectancy. In 24 countries , eighteen countries are developed countries. And only six countries are developing countries.



We completed the data pre-processing and analysis, so now we need to prepare final data for the model.

Firstly we are removing the Year column , because of no correlation and no use for model to predict the expectancy, because it just a number.And also we removed multi colliniarity features like 'under-five deaths','percentage expenditure' from the data.

And we used Binning to bin the Country column to create new feature called Continent,The country column also have duplicates , if we are trying to reduce the unique values in categorical column that will be better for the model. So we created Continent according to the country name. We used python package called ‘pycountry\_convert’ pacjkage to find the continents of each country. And we have 6 unique values for continents.we have high number of countries from africa.that is 864 countries.

AS - Asia

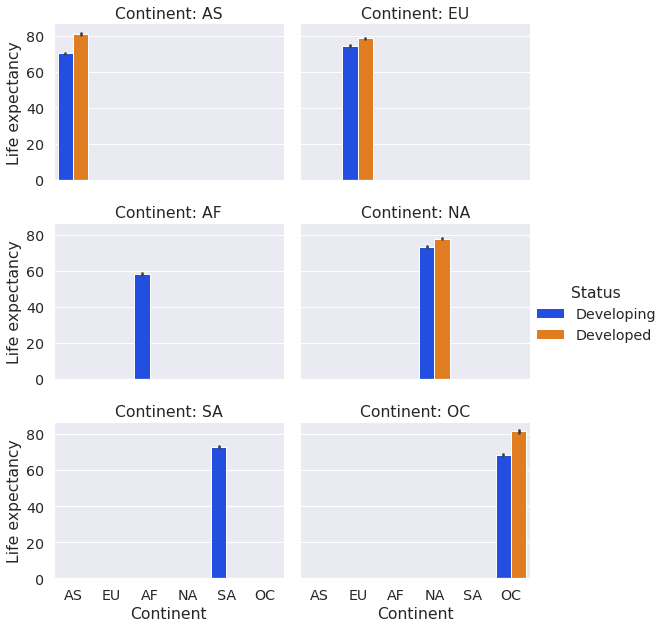
EU - Europe

AF- Africa

NA - North America

SA - South America

OC- Australia



From the barchart, we can say, we only have developing countries from africa and south america in our data.In other continent, we have high number of developed countries compared to developing countries.

After that, we removed country column and created continent column for better understanding for the model.

Now , we have 19 features in our data. And we need to convert categorical string values in to numerical values using Label encoder method, We converted status and continent column in to numeric columns.



Then, we need to normalize the numeric continous variables using Min-Max Scaler method.Then we will get the values in between 0 and 1. Also we have another approach standardization , it will deviate the data according to z-score. Here, we are using normalization to purify the data.

Finally, I prepared the data for the model, now we have cleaned and pre-processed data for them, the normalized data only use for regression models like linear regression. For tree based regression algorithms do not use normalized data.

**Modelling**

**Split the Data**

So, we reached the important and the crucial step in machine learning. Firstly we need to split the data into train and validation set. For that we are using sklearn package classes. So we have total 2938 records. So, we decided to use 80% data for training and 20% for validating the model. So we have 2358 records for training and 588 records for testing or validating the model.

**Train the model**

For training the regression model we have different algorithms in machine learning. Here we are using Linear regression, Decision tree , renadom forest and XGBoost regressor.

In Linear Regression, we are using multiple linear regression.Because we have multiple features here to predict the dependant variable. We predict the values with those independent variables. For training the linear regression model, we will be using the normalized data. We have 81% variance score from linear regression model. And the cross validation score for the data is 78%. That means our model will good predictions according to our input values.

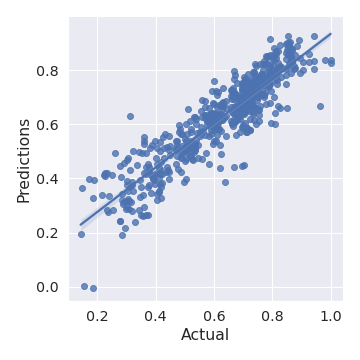
For random Forest Regressor, we are using the original data with out normalization. Because Random forest is an ensemble method which means it uses multiple decision tree to predict the out come value . So that it is called Random Forest. We are using 100 trees in random forest for the training. In random forest we have 89% cross validation score.

XGBoost algorithm is the state of the art model for regression. It Is also an ensemble boosting method.It calculates the error in each iteration and improve in the next iteration. It will boost the model. So, we got 90% cross validation score from the data.

Decision Tree is an inverted natural tree. It is an algorithm which can use for classification and regression problems. According to our data we got 76% cross validation score. It is a tree based algorithm which have nodes and leaves to take decisions in each levels.

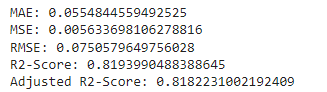
**Evaluation**

**Linear Regression**

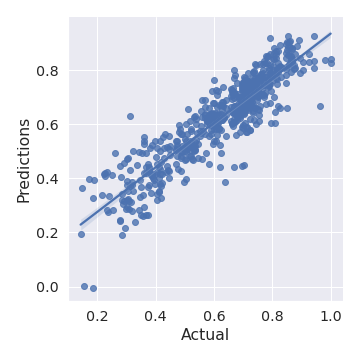


From the above graph we are getting the good predictions according to the actual value. So our input features are good to predict the target variable.

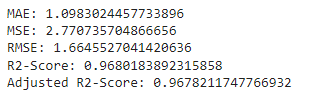
Below we have the value for different metices to evaluate the model. Here MAE, MSE, and RMSE are the errors, so we need to minimize the eeror of the regression model. The R2-score is the coefficient of determination metric for finding the variability of the model. That is if we have the feature significant to the target variable the R2-score will be higher. So we want maximum R2-score, the value should be in between 0 and 1. The Adjusted R2-score is the modified version of R2-score, the value of the adjusted R2-score will be less than or equal to the Original R2-score.



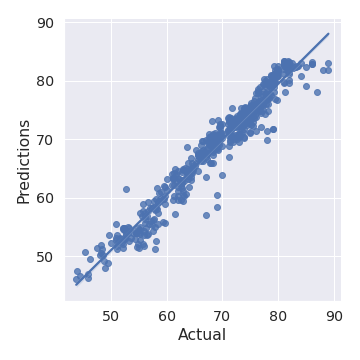
**Random Forest**

****

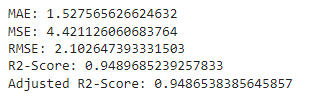
The plot is a same as regression but the values of the evaluation metrices have big change. That is MAE, MSE and RMSE values increased and also increased the R2-score values. So we can not say the model is predicting correct values all the time.



**XGBoost Regressor**

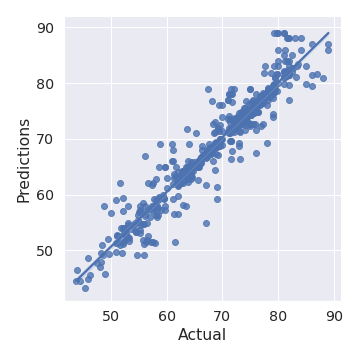


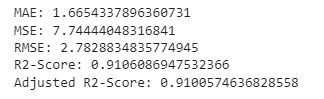
From the graph we cans say the predicted values are more nearer to the actual values compared to random forest and linear regression



The R2-score increased to 94%. and RMSE is 2.10. This model should be reliable compared to the predicted values and the evaluation metrices.

**Decision Tree Regressor**

****



Here, we have 91% variability and RMSE as 2.7.It is also a good model for our data to predict the target variable.The graph also looks fine according to our test data

**Conclusion**

We trained four models with our 2938 records. So we got good results from all models except Random forest algorithm. It is an overfitted model to our data, the basic problem with random forest we need more data to train.The XGBoost model and decision are work fine and predicts almost same value of the actual one. If we have more data we would have better result from the models.

So , the important features for the models are Schooling and Diphtheria features more importance to predict life expectancy of the country.

**Classification- Heart Disease Dataset**

**Business Understanding**

According to Heart disease dataset, we need to predict the Patient hasheart disease or not. We have some demographic and som medical features to classify the patients who has heart disease or not. For having heart disease we have many reasons and its not a sudden disease, its due to life style habits and other reasons. Using our dataset we can analyse the reasons for coming heart disease according to each patient records.

**Data Understanding**

For Analysis, we have dataset which has 1024 records and 13 feature variables.

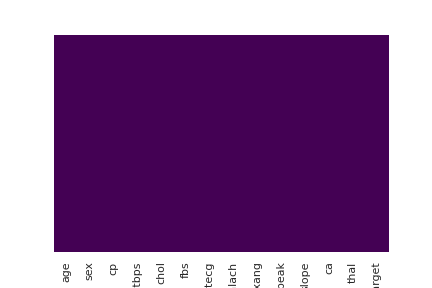
|  |  |
| --- | --- |
| Column | Description |
| Age | Age Of the person in years |
| Sex | Sexof the person, 1=Male, 0=Female |
| Cp | Chest pain Type(4 values) |
| Trestbps | Resting blood pressure in mmHg on admission to the hospital |
| Chol | Serum cholestoral in mg/dl |
| Fbs | Fasting Blood sugar and gt 120 mg/dl if >120 = true =1 otherwise 0 |
| Restecg | Resting electro cardiographic results(0,1,2) |
| 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) coloured by flouroscopy |
| Thal | Defect  1= Normal,2=fixed defect,3= reversible defect |
| Target | 1 or 0, The patient has cancer = 1,or not =0 |

**Data Preparation**

In Data preparation, we need to clean the data and prepare the data for model. For that we need do data exploration analysis as following:

1. Checking Missing Values and Finding Outliers
2. Univariate Analysis
3. Bivariate analysis
4. Correlation
5. Binning

1. **Checking Missing Values and Finding Outliers**



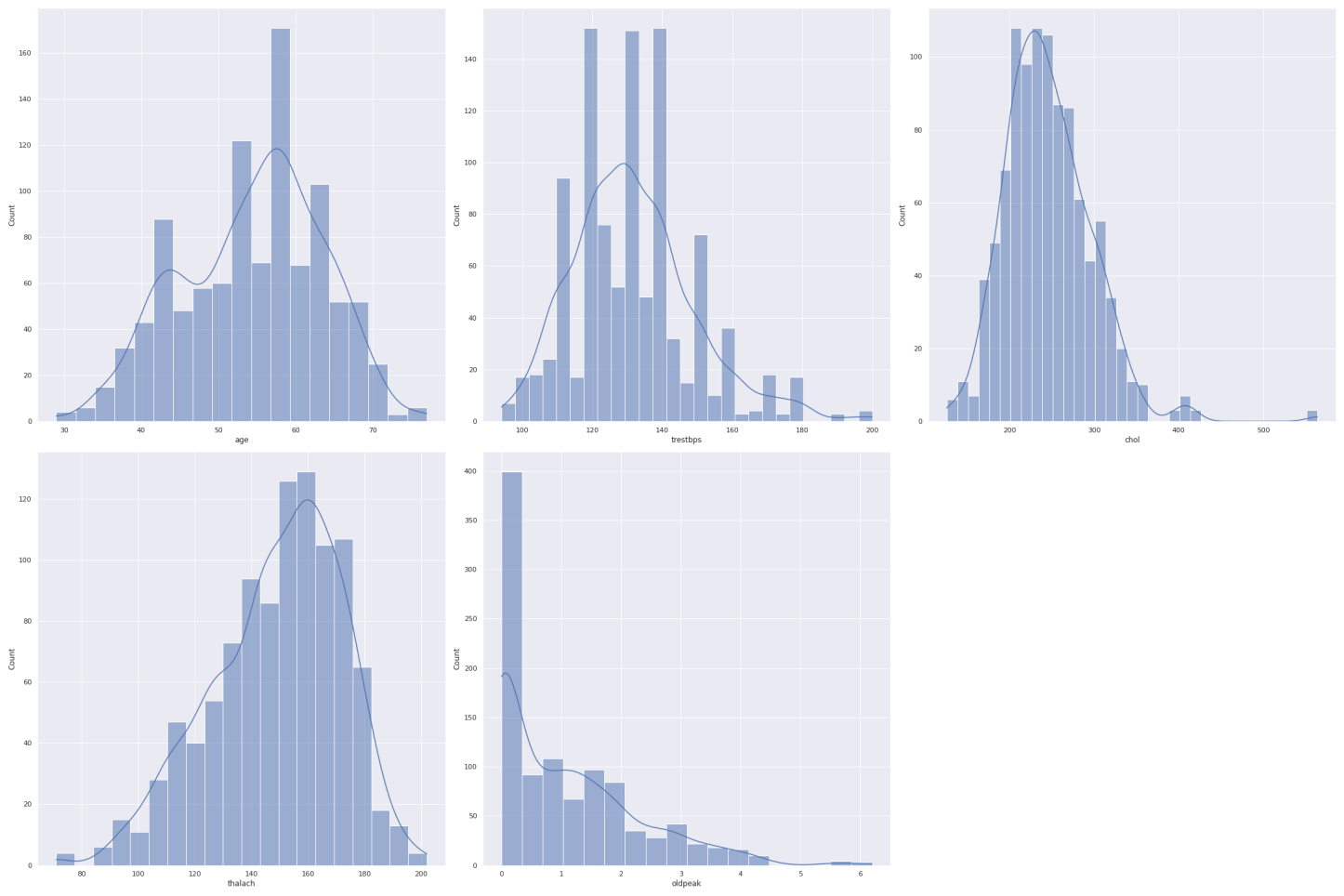
According to the graph, we have no missing values in our dataset.

For outlier detection we are using inter quartile range and plotting the box plot with each variable. The variable age has no outliers. Resting blood pressure has 30 outliers. But we are not ignoring them, because we have only 1000 records here. We only remove the variable which ahs greater than 100 outliers.Cholestoral has 16 outliers. Thalach, that is maximum heart rate achieved variable has 4 outliers. For outlier detection we only consider quantitative features. For Oldpeak has only 7 outliers and we are considering those values.

1. **Univariate Analysis**

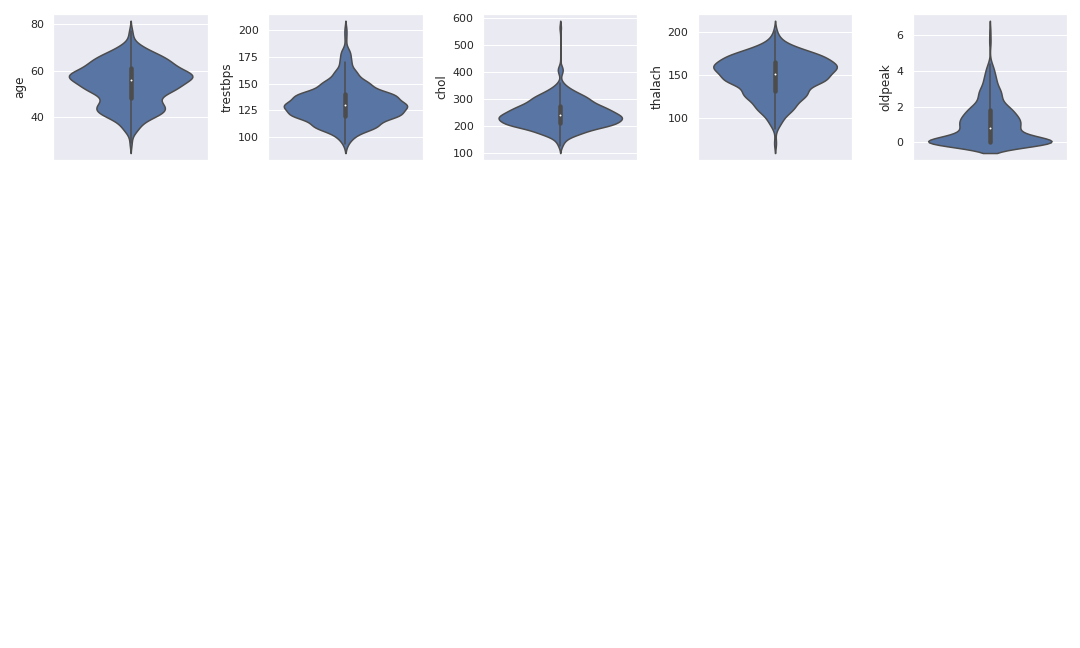
First of all we analysing the numeric variables.

1. Age
2. Trestbps
3. Chol
4. Thalach
5. Oldpeak



From the above graph ,the age,trestbps,chol and thalch have normal distribution and the oldpeak has right skewed distribution.

We can also use Violin plots for univariate analysis, it alos plot the outliers.



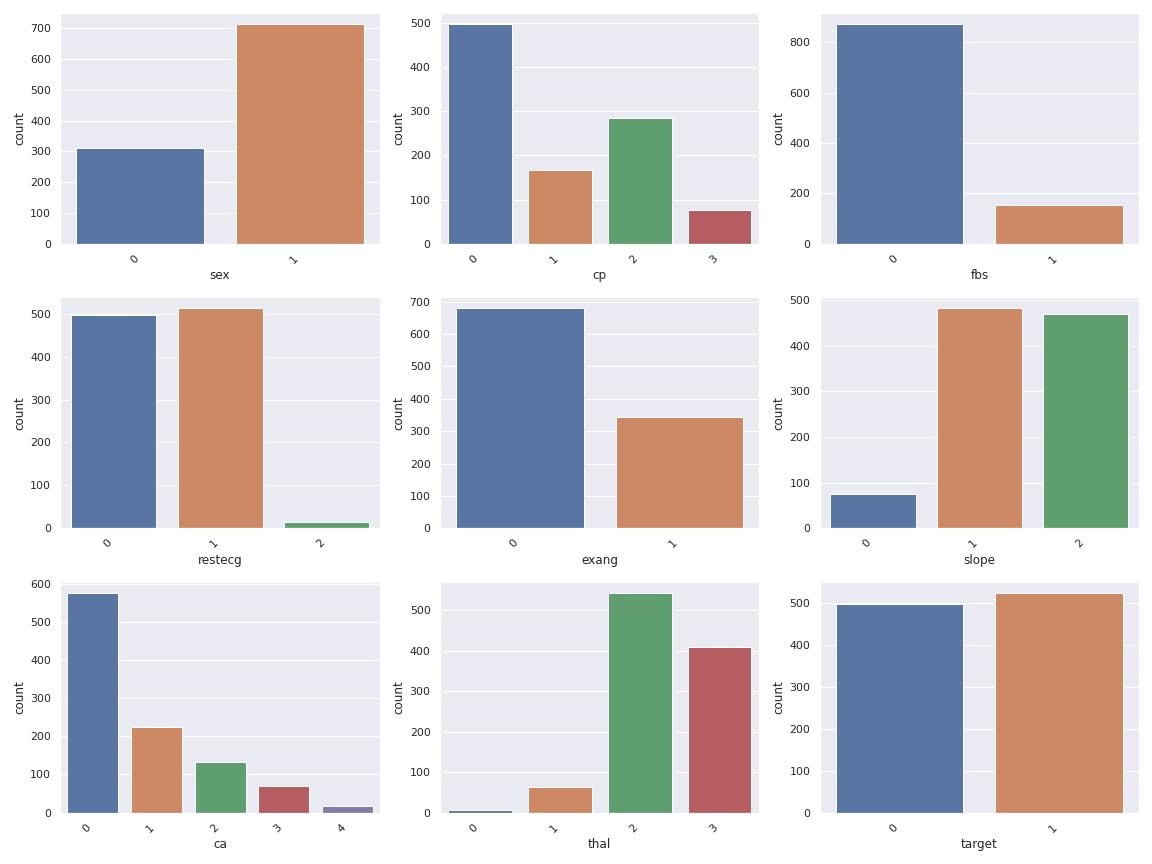
From the violin plot , we can plot the values range in the dataset.

For categorical univariate analysis, we have count plot and Pie charts. In this data we have more number of categorical features compared to numerical features.

Categorical features:

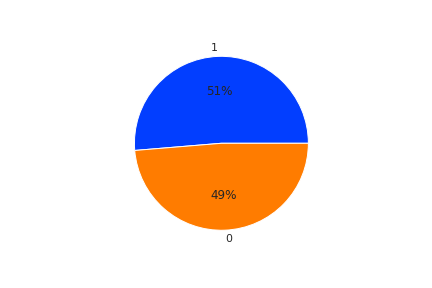
1. Sex
2. Cp
3. Fbs
4. Restecg
5. Exang
6. Slope
7. Ca
8. Thal
9. Target

And we can plot the combioned graph for all categorical variables.

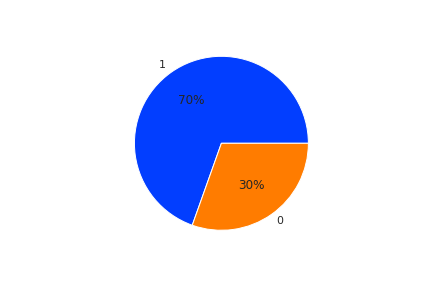


We have more male patients in our dataset compared to females.Chest pain has more values as 0. More patients have fasting blood sugar less than 120 mg. Maximum heart rate achieved only few people and exercise induced angina has less compared to non induced. We have more number of fixed defect patients in defect column.That is Thal.

And Our target variable is balanced between 1 and 0s.



We have 50% customer has heart disease and 49% customer has not heart disease.



We have 70% male patients and 30% female patients.

1. **Bivariate Analysis**

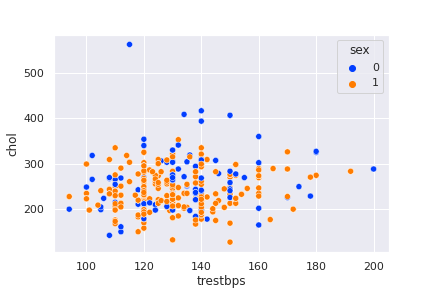
In bivariate analysis, we can analyse the relationship between two variables. Wehave the pair plot, scatter plot and line plot for analyse the numerical variables and bar chart and point plot for categorical analysis.



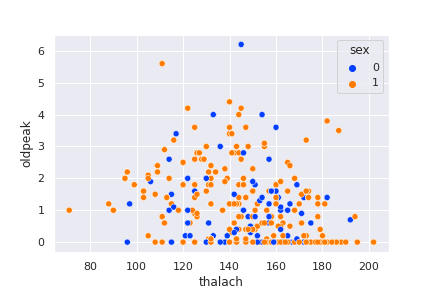
Here we have the pair plot to find the relationship between all numerical variables.Unfortunately we have no variable related directly each other.That its it scattered in different way.

Also we can plot some scatter plot between the numeric variables.

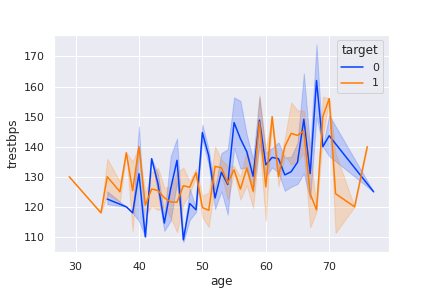
1. Trestbps vs chol



1. Thalach vs oldpeak



We have also line plot, we plotted the line plot between age and trest bps with target variable.

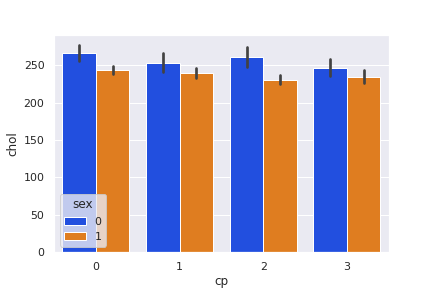


From the graph, we can see , if age greater than 50, we have the patients having heart disease and not.But after age 70 increase the heart patients and decrease the not heart diseases.

Also , we can find the relationship between categorical and numerical variable. But we have no relationship between categorical variables .That is true statement.

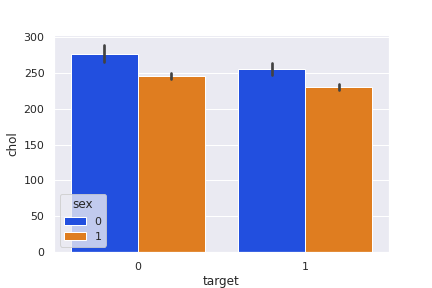
We have the bar chart for finding the relationship between categorical and numerical variables.

* Chest Pain and cholestarol

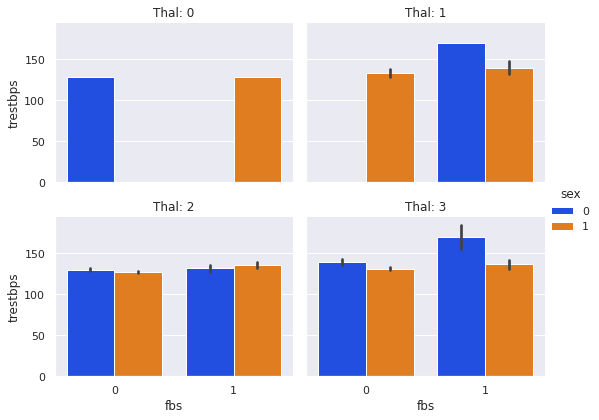


Cp is a categorical feature which has 4 categories. And for every categories female has more cholestrol .

* Chol and Target variable

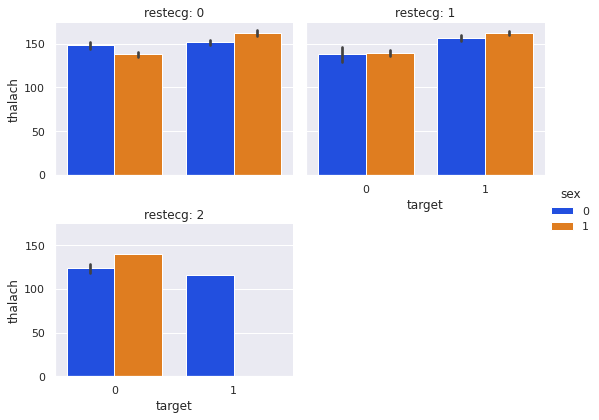


* Fbs vs trestbps with defect variable



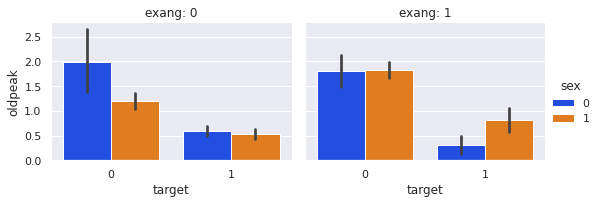
This is a combined plot between trestbps, fbs and defect columns. From the graph, the women have reversible defect more than men having blood pressure greater than 120 mg and also in the case of n ormal defect.

* Target vs Thalach variable with restecg



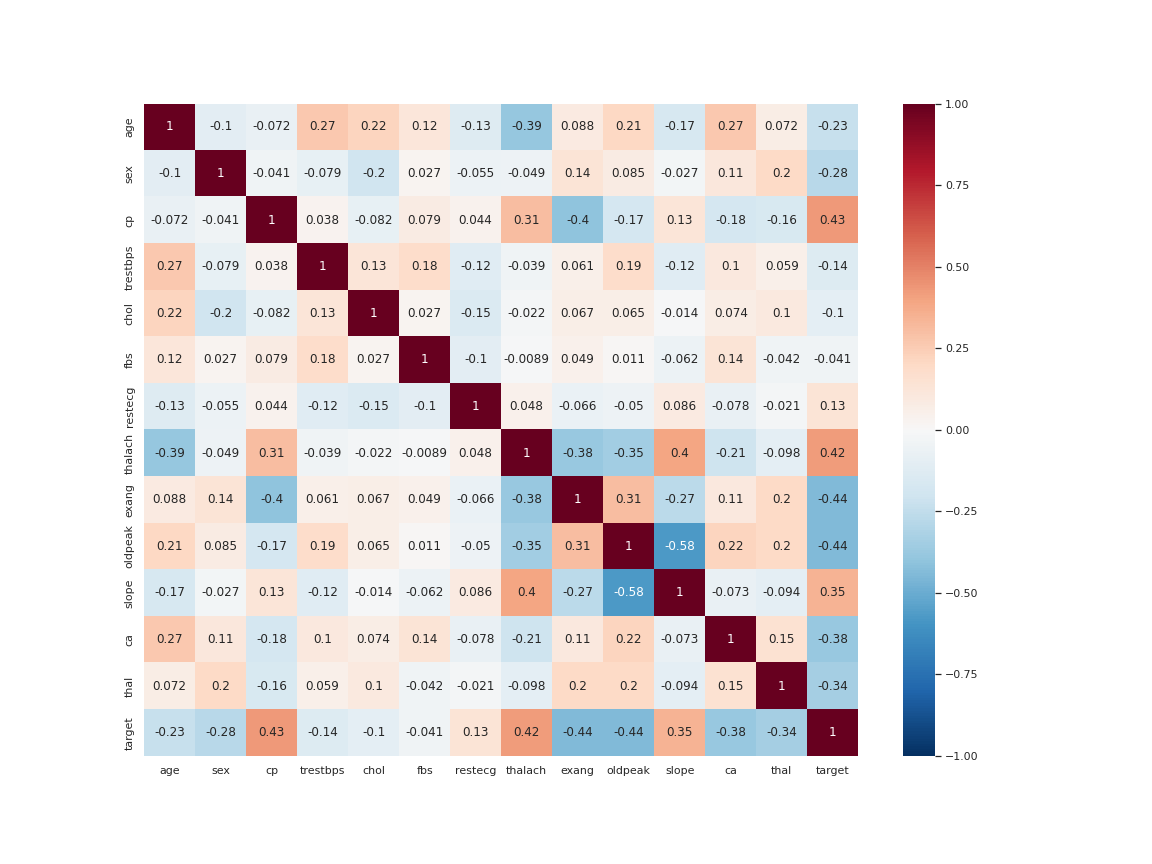
Maximum heart rate achieved patients have the chance for have the heart disease when resting electrogradiac result 0 and 1.If electro gradiac value is 2, then men has no heart disease.

* Target vs Oldpeak with exang variable



When increasing the oldpeak decrease the chance of heart disease in both men and women when the exercise induced angina in 0 or 1.

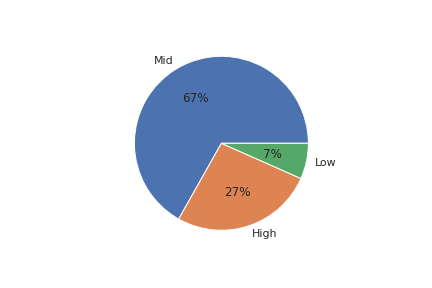
1. **Correlation**



From the correlation plot, we have cp, thalach and slope have positively correlation with target variable. And exang, oldpeak, thal and ca have negative correlation with target variable. So these six variables or features are more important to the out come variable to predict the patient has Heart disease or not.

1. **Binning**

It is a technique for converting numerical variables into categorical bins according to some range of data. Here we are binning the age variable according the range like if we have the age in 0-40,40-60,60-100 and it convert into low, Mid ,high bins in another column.



After binning , we want to encode the variable in to numbers.We converted string values low,mid and high into 1,2,3 respectively.

Then we want to apply normalization to the continous variable, we applied min-max scaler for normalization.We completed all the pre-procesing steps with data.Now we can move to Model development. And we only use age binned feature for model training.

**Modelling**

Now, we are training the model. We are training 4 different models.

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. XGBoost

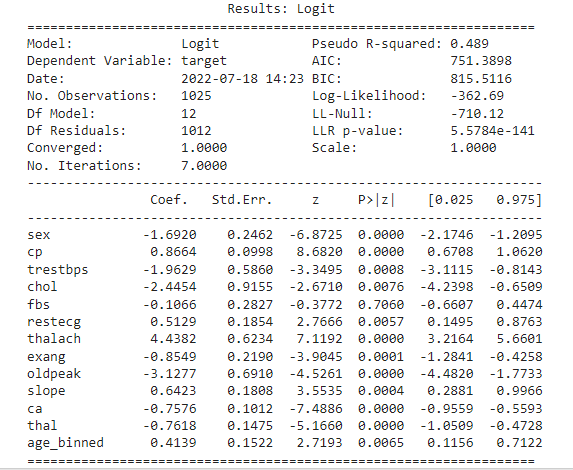
**Split The Data**

We need to create train and validation data for model training. Here we have only 1024 records and 13 features. We are splitting 90% for train the model and 10% for validating the prediction from the model. Also we selecting randomly for each run.So we have 922 records for train the model and 103 for test the model.

**Train The Model**

1. **Logistic regression**

It is a classification algorithm and it is uses sigmoid function to classify the outcome variable.Commonly, we are using for Binaryclassification.

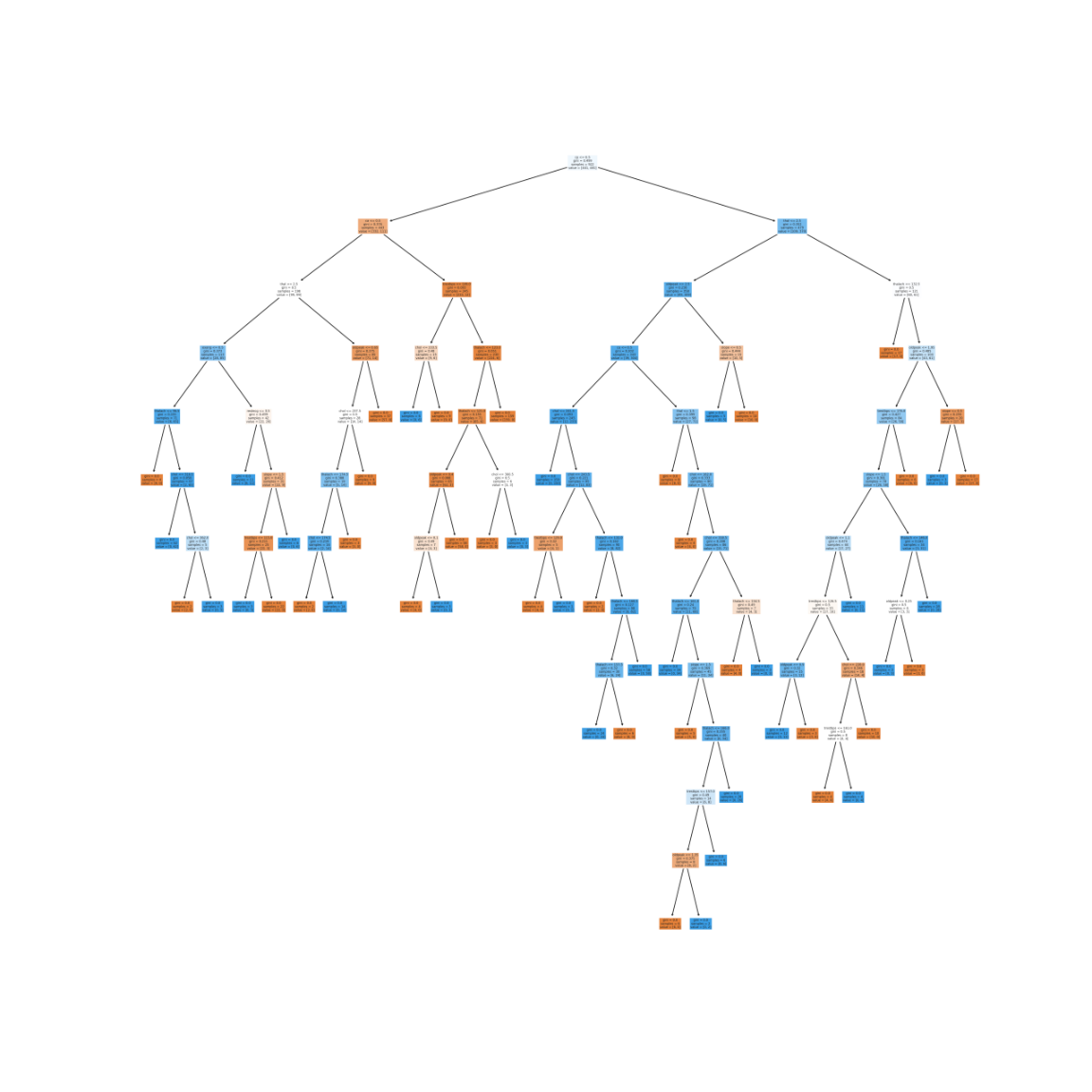


This is the summary of the model.From this, cp ,thalach and slope variables have more importance to predict the outcome variable.Because that coefficient are higher.

1. **Decision Tree**

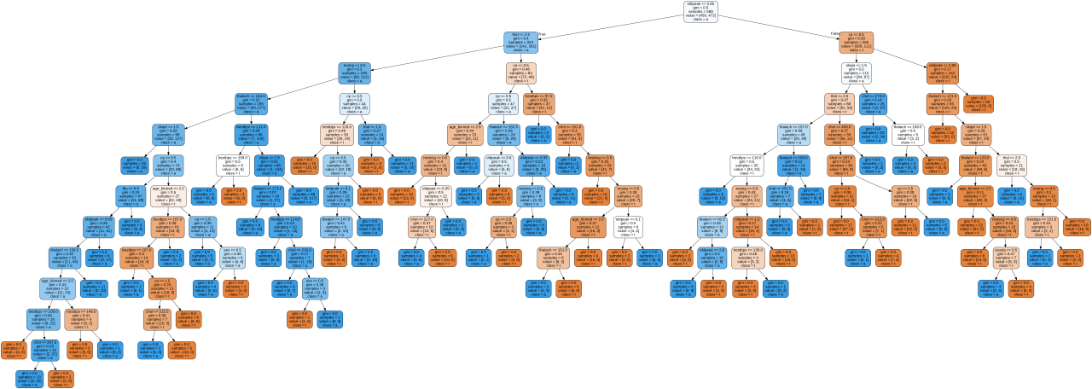
It is a tree based algorithm in machine learning we can use for both classification and regression tasks. We are using un normalized data for tree and ensemble methods. In decision tree , we can pass some parameters like criterion as gini here,And all other parameters are default by the model. Here , cp, thal and ca have more importance to the model.

Decision Tree Plot:



1. **Random Forest**

Random Forest is an ensemble bagging method for classification. It uses multiple decision tree to train the model for better result. We are choosing the default values for the parameters and trained the model.Here, cp ,ca and thalach have more importance here.



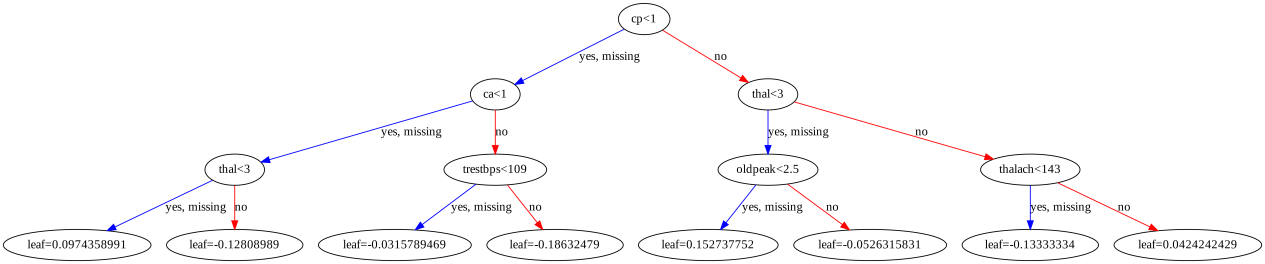
It is a single decision tree from the random forest.

1. **XGBoost**

It is also an ensemble method for classification. It is a boosting ensemble method. It boosts the model according to the result of each model in the XGBoost pool.

Here,thalach ,cp and ca have more importance for the XGBoost model.

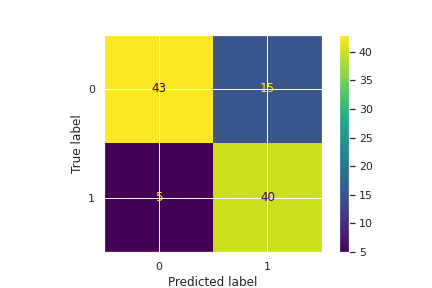
XGBoost plot tree:



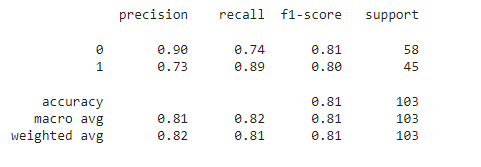
**Evaluation**

For evaluating the classification model we need a matrix called confusion matrix. According to confusin matrix we are calculating the accuracy,precision,recall and F1-score of the model.We need to compare all of this measure according to our business usecases. Here we need to use Recall as the measure and also use F1-score.

For Logistic Regression We have the Following Result obtained.



Classification report:

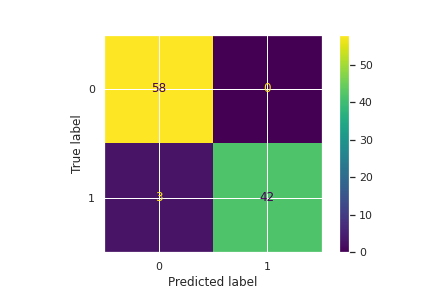


In Logistic Regression Model, we got 81% accuracy and 80% F1-score.Here we need more recall value compared to precision.we have recall as 89%. So It is a good predictor for this dataset.

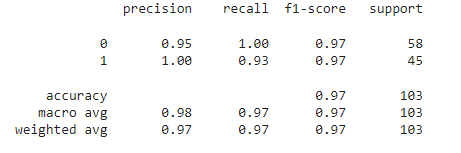
We have 103 samples for test data, Out of 103, 83 records are predicted correctly and only 20 predicted incorrectly. False negative is less.

So our model good for predicting the patient which has heart disease.Because we have more Recall.

For Decision Tree we have the following values of our evaluation metrics.

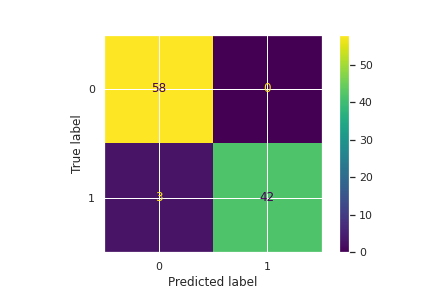


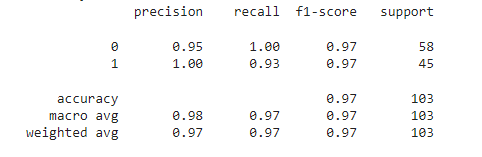
Classification Report:



Decision Tree Gives the better result Compared to Logistic Regression. Because increased Recall as 93% and accuracy as 97%.Accuracy and F1-score as same.

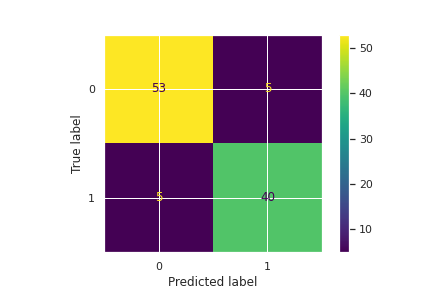
In Random Forest we got same result as decision tree. And also same confusion matrix.



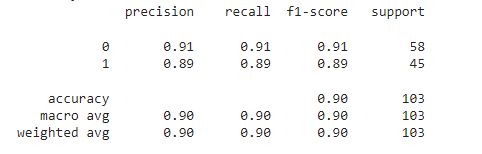


So we have the same result but small difference in AUC score.

XGBoost is A Extreme gradient Boosting method, We will get very good result compared to other classification algorithms.



Classification report:

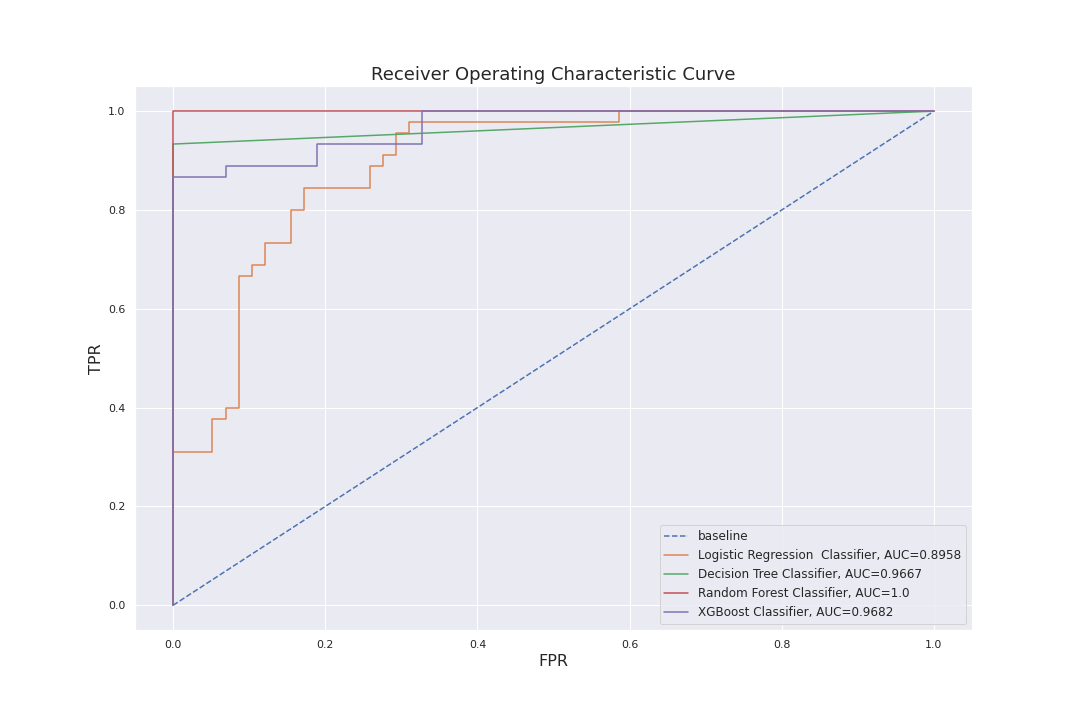


We got 90% accuracy and F1-score. And also we got same precision and recall. It is more converged version of model.Because Here both precision and recall have same value.

**ROC-AUC Curve**

In classification , we can create multiple models with same data.But we need a better classifier among them to classify correctly. We need to Minimize the False Negatives and False Positives. Using ROC-AUC curve we can determine the best classifier.

ROC-AUC curve is the curve in between True positive rate and false positive rate values. From the below graph, we have dotted line in centre of the graph.That means AUC=0.5.So our curve should be above the ldotted line.Otherwise our classifier have no ability to classify the data.



Here we have AUC=1.0 for Random Forest classifer, we can not say models are 100% accurate. So Random Forest is overfitted to our data. But decision tree and XGBoost have Robust Roc curve and AUC value as 96% with our data, So the better model is XGBoost according to the evaluation metrics values. And Logistic Regression has 89% AUC value.

**Conclusion**

From the model results, we can say XGBoost has Robust result according to our data and with ROC-AUC curve. And also cp,ca and Thalch are the most important features to predict the heart disease variable.

In this analysis, we used a small dataset, so that the Random Forest model Overfitted to our data. And we can not say these are the perfect result and our model is best.Because the final decision comes only after discuss the metrics with business team.

========================