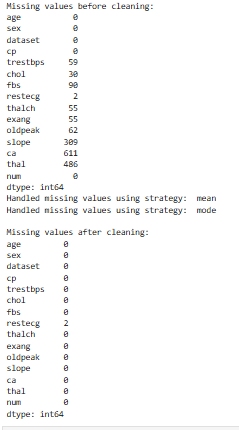
# Summary Of Task3

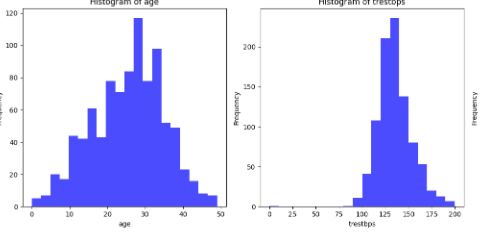
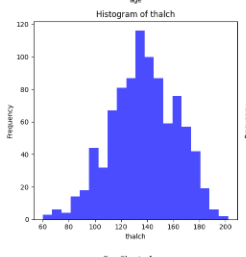
**Data Set :** Heart Disease UCI data set is used in task3 where it predict the disease condition of one patient. Reason for using this dataset in this project , to learn patterns because in this dataset there is missing values , negative values , categorical and numerical feature and so on.

Steps 1 :

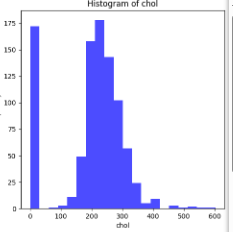
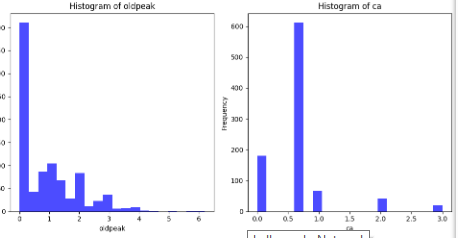
1. Load Data using pandas Library
2. **Clean data**: where duplicate values are handled
3. **Handle Missing values**: missing values are handled using mean strategy for numeric values and for categorical mode is used



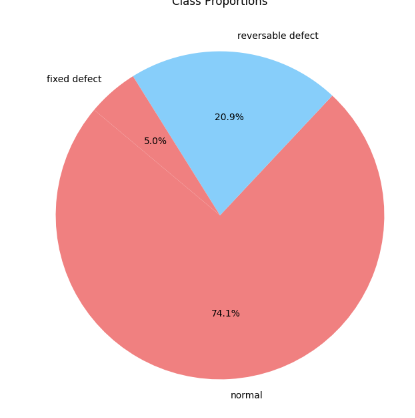
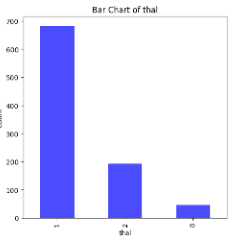
1. **EDA:** this technique helps to see data balancing and outlier existence in dataset and frequency of values in feature this all shows through visualization and by visualization also it shows the distribution structure whether it is normal and skew right or left like in this data set..

These three numeric feature histogram distribution is normal distribution.



These distribution are skewed right means almost values are lies in 0



This is target feature **THAL** Distribution this shows the imbalance of dataset means dataset is not balanced and the result of class distribution and proportion is .

Class Distribution:

thal

normal 682

reversable defect 192

fixed defect 46

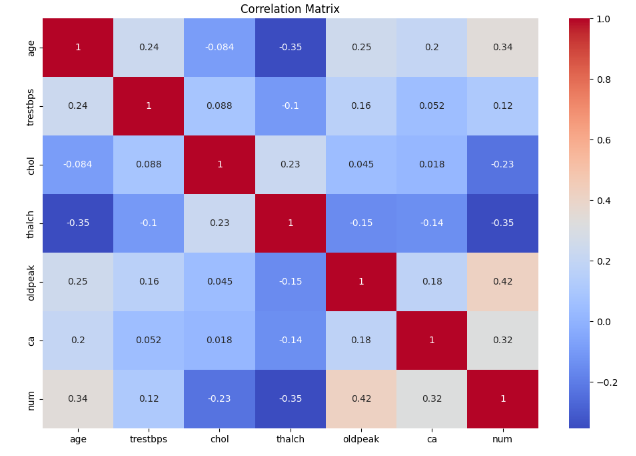
Class Proportions:

thal

normal 0.741304

reversable defect 0.208696

fixed defect 0.050000



This is correlation matrix which shows the relation between features

In this correlation there is not relation above 0.8 and below -0.8 so for this reason features are not drop because of not having strong relation between feature .

1. **Machine Learning Model**: 3 ML model I used here to see perform well Logistic regression , Decision Tree and Linear Discriminant Analysis Algorithm

**Accuracy of Logistic regression**: 0.8152173913043478

**Accuracy of Decision Tree Classifier:** 0.7119565217391305

**Accuracy of LDA:** 0.7717391304347826

Here I chose **Logistic regression** because of its performance.

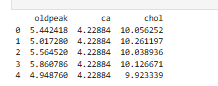
**Feature Engineering Technique**

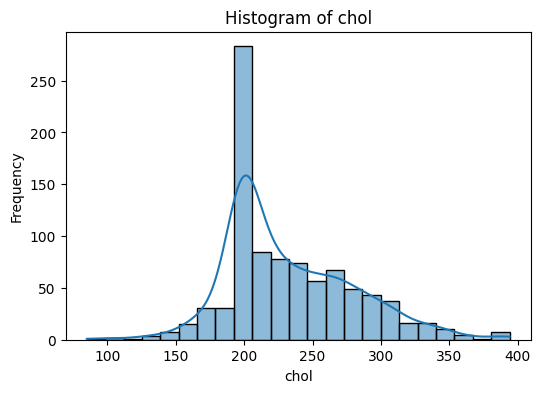
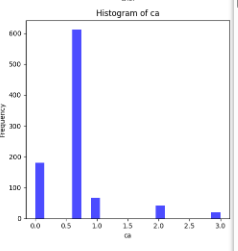
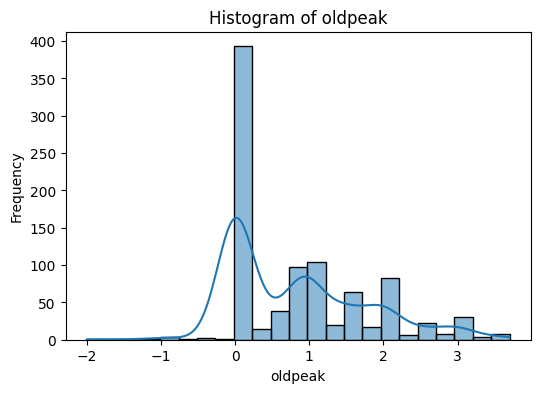
This technique is used for improving performance of model in Feature Engineering Method I use Two techniques **Feature Transformation** , **Feature Generation and Feature Scaling.**

By using this technique the Performance of model is not improved however it is decrees from 0.81 to 0.76

**A)Feature Transformation result**

This technique is used to transform features because these features are skewed feature make it normal I use this feature transformation technique



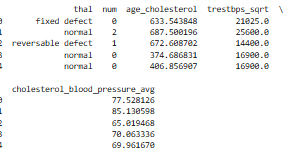
 

These are the graph which are not normal form.

1. **Feature Generation results**

Age\_cholestrol,tresttbps\_sqrt and Cholestrol\_blood\_pressure\_avg

These are new generated features . feature are generated on the basis of their relation like age and cholestrol depends and so on

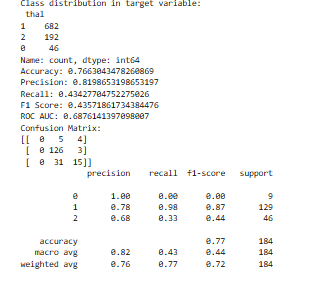


1. **Feature Scaling :**

Feature scaling where we make features value between 0 to 1 there is other methods like normalization, min-max scaling so I use standardization method.

**Result:**

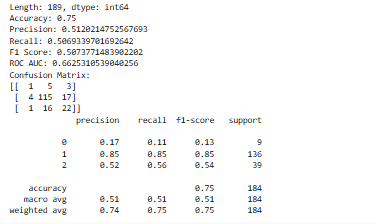
By using these three techniques the overall performance is this..



1. **Polynomial Feature technique**

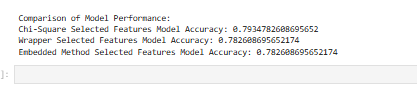
Polynomial features can significantly enhance a model's ability to capture non-linear relationships and interactions between features.

And the result after doing this..



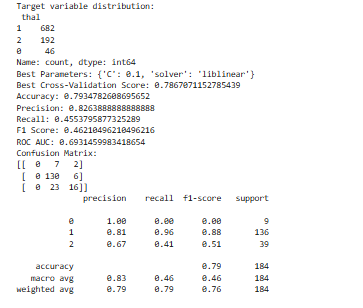
1. **Feature Selection Technique**

In which I use 3 categories filter method wrapper method and Embedded method and the results are …



**Hyper Parameter tuning**

* This result is using feature engineering technique feature transformation and feature generation.



* **This result is without feature Engineering techniques.**

