PYTHON FOR AI LAB REPORT

R0865976

Problem –

Heart Attack Analysis & Prediction Dataset

About the dataset

14 features

304 Rows.

1 target 0 or 1

0- Low chance of Heart attack
1- High chance of heart attack

Histogram below representing all Labels-

age sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall output

5 categorical columns-

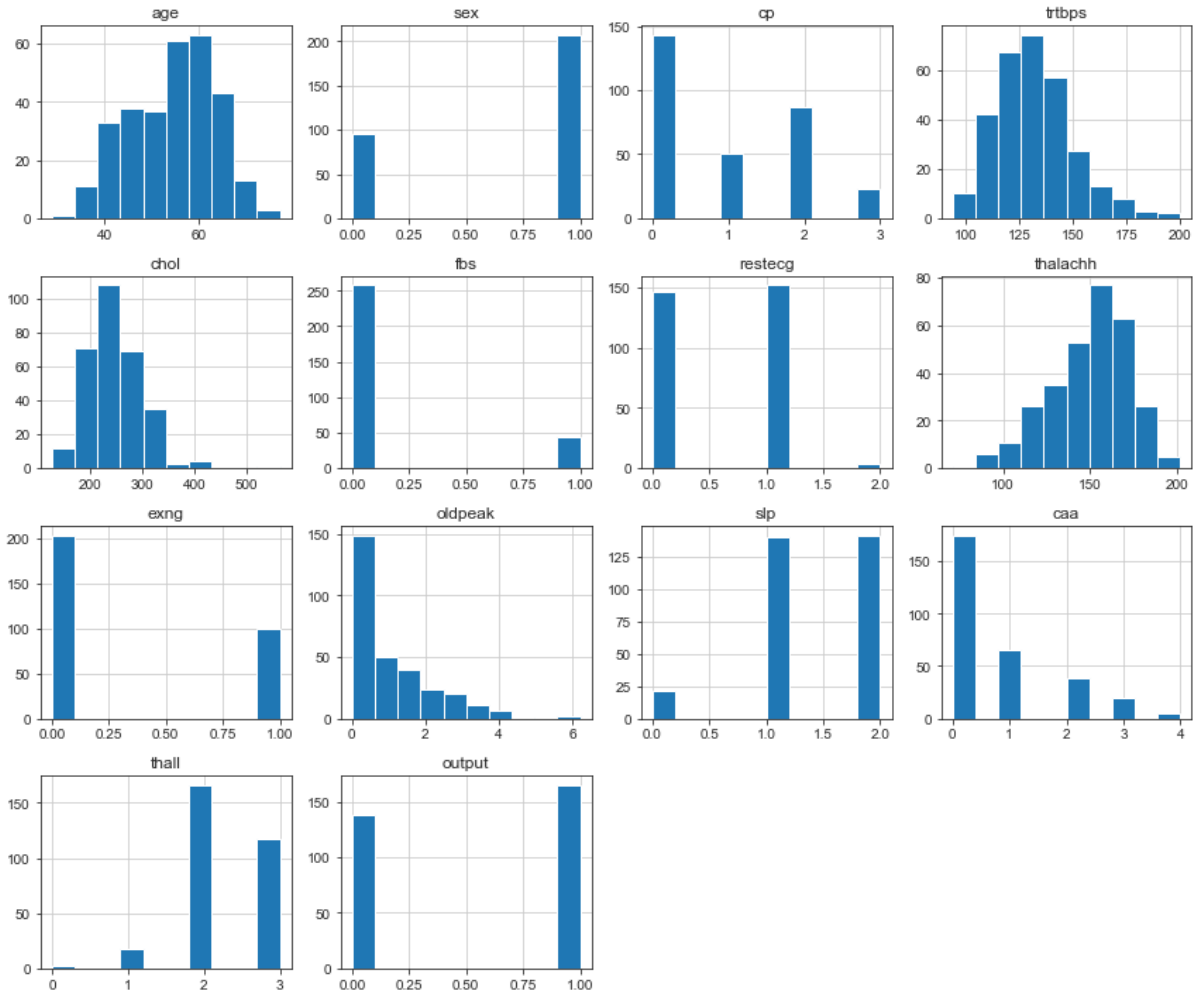
Sex

Thalassemia

Reste_cg type

Chest pain type

Slope type



Data Preprocessing and Feature engineering –

The dataset was too complete.

Removed some values from dataset.

Filled missing values with the mean rounded to a 0 decimal.

Removed Duplicated Rows in the dataset.

SelectKBest and Chi2 for feature importance.

- Standard Scaler()

-KNN + **PCA(n-components =1)**

-SVR

- Min-Max Scaler() 0 to 1

Keras Sequential

-Grid Search CV

Parameters for grid={'C': [0.01,0.1,1, 10,100,1000,10000,100000], 'gamma': [1,0.1,0.01,0.001,0.0001,0.00001,0.000001]

Data Split-

Test Size 20 percent through all models.

No K fold cross validation due to dataset only having 300 rows.

Adding sub-columns using loc to determine important columns.

One hot Encoding to turn categorical values into numerical after dropping non important values.

Tried but did not work-

Logistic Regression which is not really fit for this type of model.

Decision Tree Classifiers.

Methods –

One hot Encoding, Standard Scaler on few columns, PCA

-KNN

Plotted Error rate for Each K value (Lowest at k=9) --- Underfitting.

Accuracy of K-Nearest Neighbors: 0.9016393442622951

	precision	recall	f1-score	support
0	0.93	0.86	0.89	29
1	0.88	0.94	0.91	32

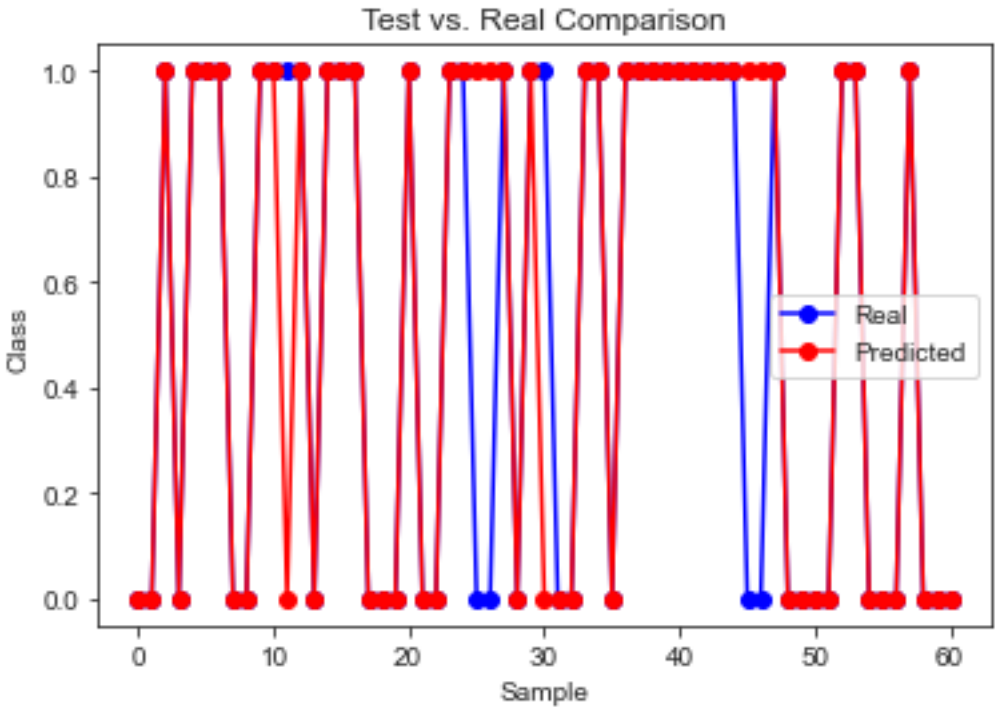
accuracy 0.90 61

macro avg 0.90 0.90 0.90 61

weighted avg 0.90 0.90 0.90 61

KNN + PCA with n components=1

Same metric on test set- 0.85



SVC

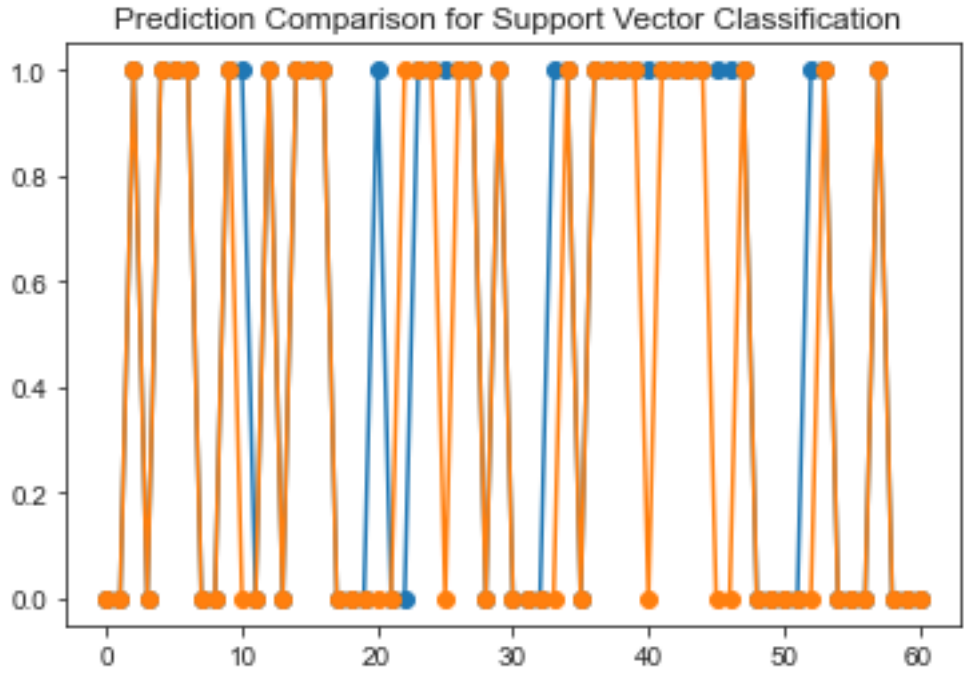
Accuracy of Support Vector Classification: 0.8524590163934426

	precision	recall	f1-score	support
0	0.79	0.93	0.86	29
1	0.93	0.78	0.85	32

accuracy 0.85 61

macro avg 0.86 0.86 0.85 61

weighted avg 0.86 0.85 0.85 61

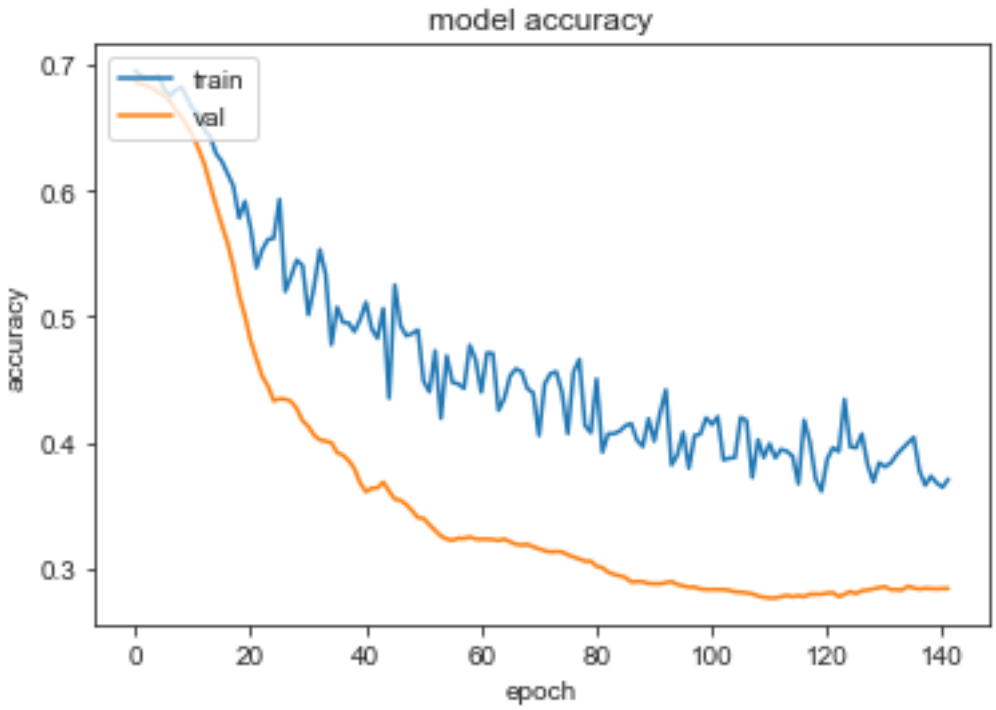


Keras Sequential Model with MIN MAX SCALER and no preprocessing steps-

Value loss: 0.3382

Epoch 125: early stopping

[[0.92590964]] Accuracy Metric



Tuning Hyperparamters –

grid=GridSearchCV(SVC(),param_grid={'C': [0.01,0.1,1, 10,100,1000,10000,100000], 'gamma': [1,0.1,0.01,0.001,0.0001,0.00001,0.000001]})

Accuracy after GridSearch tuning : 0.8624590163934426

	precision	recall	f1-score	support
0	0.84	0.90	0.87	29
1	0.90	0.84	0.87	32
accuracy			0.87	61
macro avg	0.87	0.87	0.87	61
weighted avg	0.87	0.87	0.87	61

The accuracy dropped after tuning parameters because default parameters have been set good enough to give almost 90 percent accuracy.

Tuning exhaustively might lead to overfitting.

Conclusion –

KNN is a slow learner in comparison to SVC(k=9 leads to underfitting .). SVC here can be used over more generalized for classification problems especially with the Standard Scaler since our dataset does not contain negative values.

I would prefer using a Keras with Support Vector Classification for this problem however I have tried Sequential separately above.

SVC would be the more generalized to new data fed to the model.