Capstone Project

Cardio Vascular Risk Prediction

TEAM MEMBERS

Saransh Srivastava,

Harish Patil

CONTENT

ABSTRACT

- The independent variables such as age, education, is_smoking etc… are the determinants of the dependent variable ‘TenYearCHD’. We were provided with already classified labels in our data set.

- Our experiment can help understand what could be the reason for the classification of such labels by feature selection, data analysis and prediction with machine learning algorithms taking into account previous trends to determine the correct classification.

PROBLEM STATEMENT

- The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. It includes over 4,000 records and 15 attributes. Variables
- Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

STEPS INVOLVED

- Exploratory Data Analysis
- Null values Treatment and Outliers
- Numerical and categorical Features
- Label encoding
- Correlation Analysis
- Train test Split
- Scaling
- Smote Technique
- Fitting different models
    - a) Logistic Regression
    - b)Random Forest Classifier
    - c)SVM
- Tuning the hyperparameters for better accuracy

HANDLING NULL VALUES

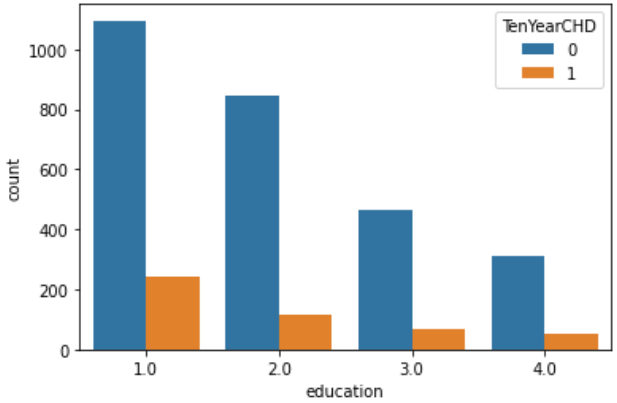
- id 0.000000
- age 0.000000
- education 2.566372
- sex 0.000000
- is_smoking 0.000000
- cigsPerDay 0.648968
- BPMeds 1.297935
- prevalentStroke 0.000000
- prevalentHyp 0.000000
- diabetes 0.000000
- totChol 1.120944
- sysBP 0.000000
- diaBP 0.000000
- BMI 0.412
- heartRate 0.029499
- glucose 8.967552
- TenYearCHD 0.000000

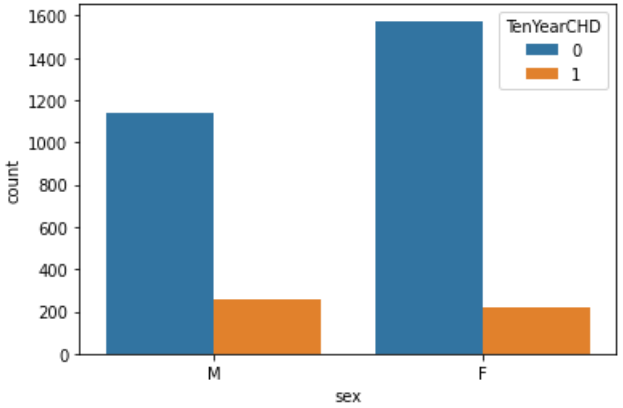
Treatment of Null Values

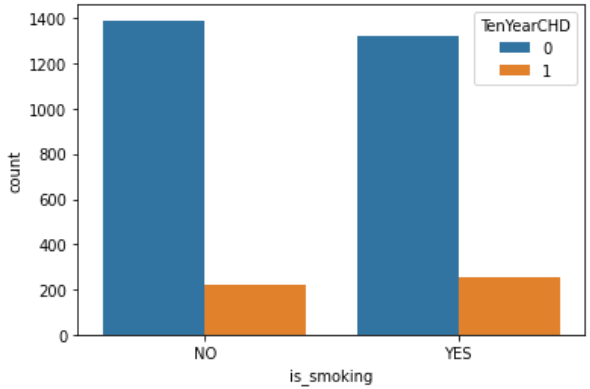
- Filling the rows which has higher than 5% null and lower than 30% null values.
- Dropping the rows which has lower than 5% null values.
- Dropping the id column.

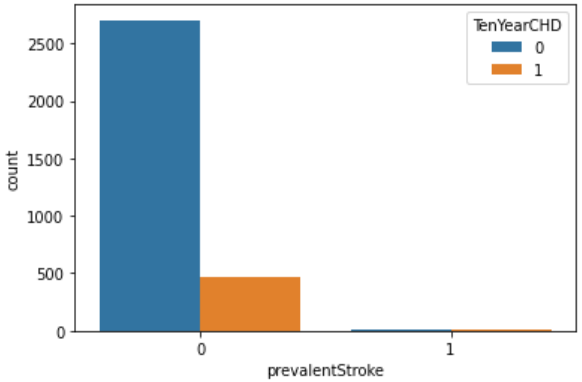
EDA

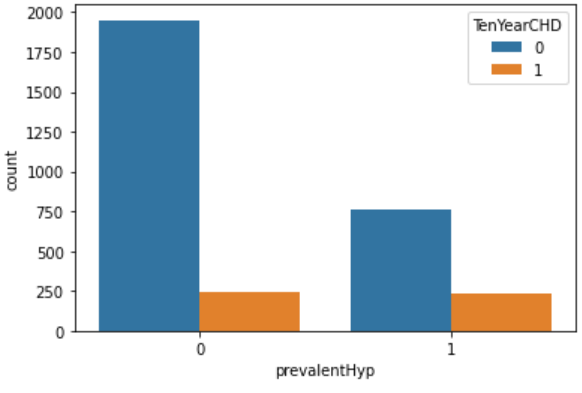
Count plot of all the numerical values in the dataset where hue is the dependent variable.

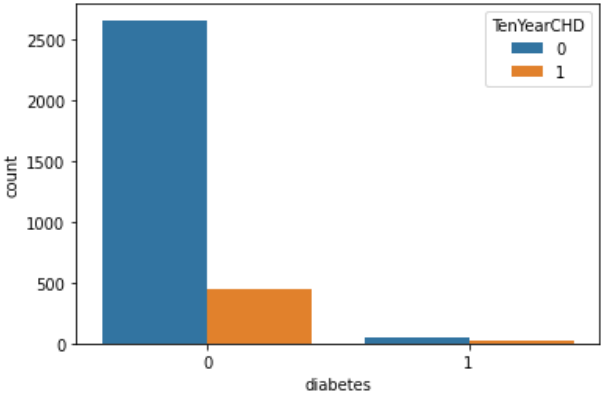






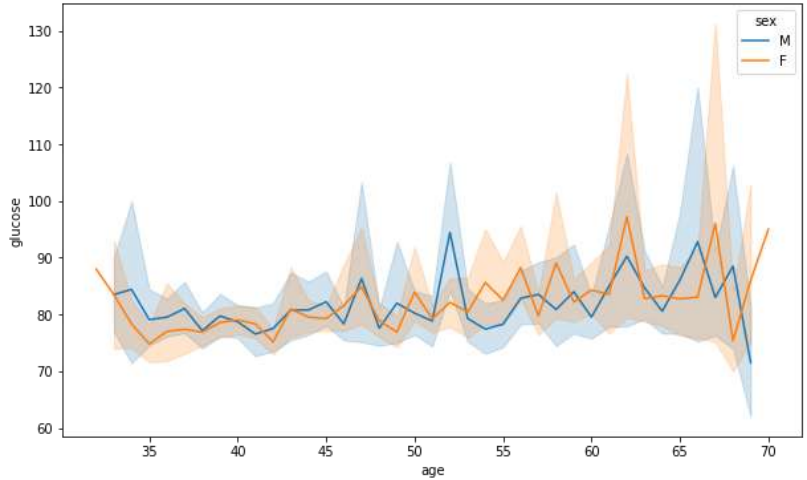


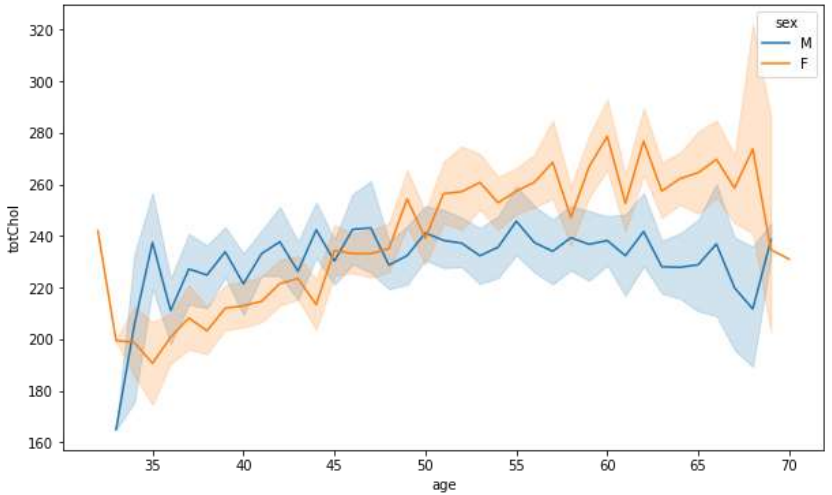


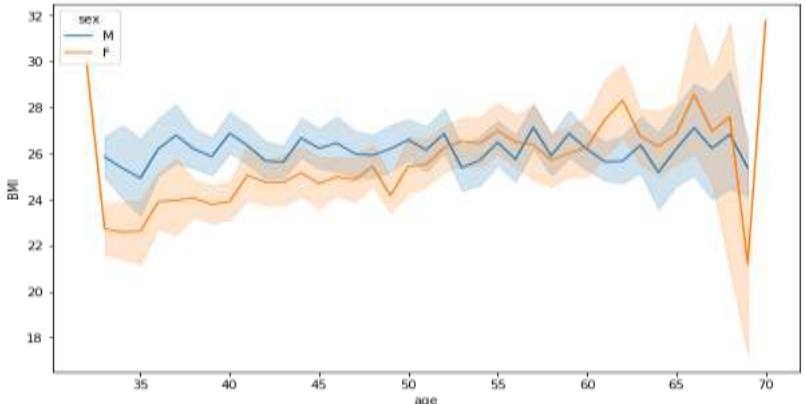


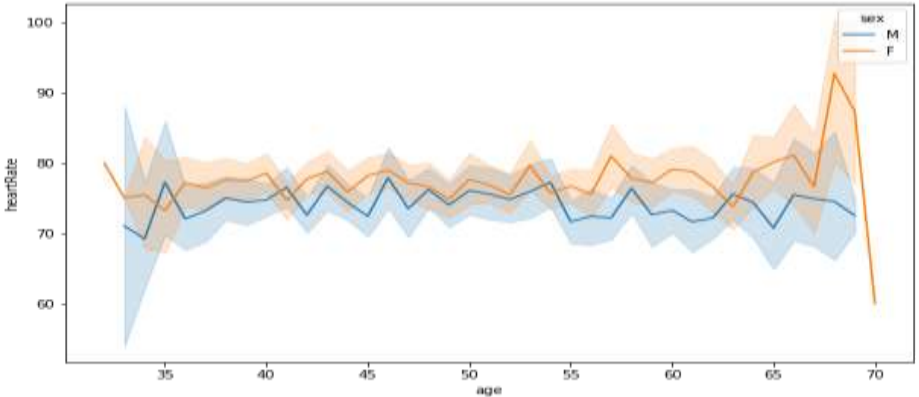
EDA

Finding as the person gets old their glucose level, Cholesterol, BMI and HeartRate increases or not.

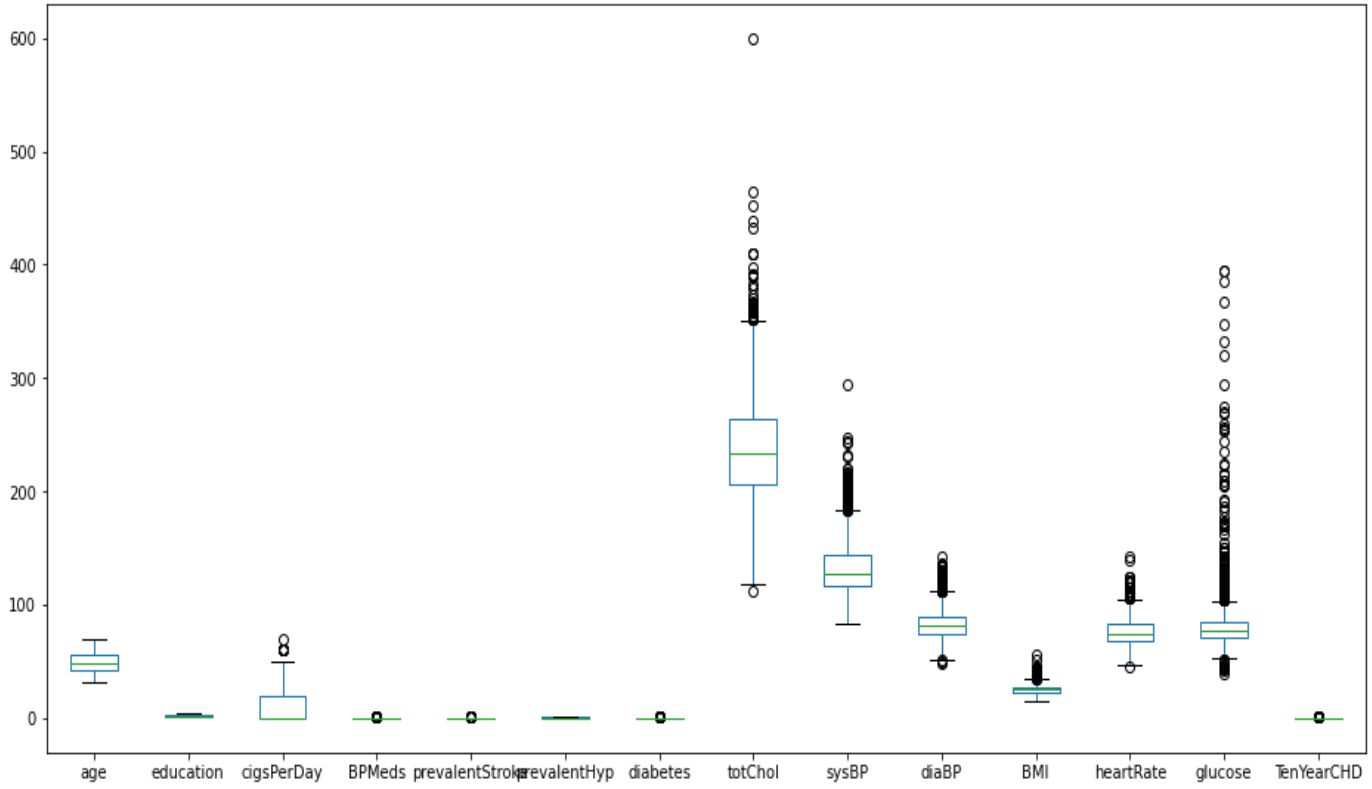








HANDLING OUTLIERS



We used IQR quantile technique to remove the outliers from the columns :-

- `cigsPerDay`
- `totChol`
- `sysBP`
- `diaBP`
- `BMI`
- `heartrate`
- `glucose`

MULTICOLLINEARITY

sysBp means systolic blood pressure and diaBP means diastolic blood pressure are correlated to each other but we cannot drop these columns as
Blood pressure is measured using two numbers: The first number, called systolic blood pressure, measures the pressure in your arteries when your heart beats. The second number, called diastolic blood pressure, measures the pressure in your arteries when your heart rests between beats.

LABEL ENCODING

- We had two categorical features [is_smoking] where values were ‘YES’ or ’NO’ and [sex] where values were ‘M’ or ‘F’. So, in order to encode this we just mapped the variables into 0 or 1 respectively.

```
#Mapping the Variables
df['is_smoking']=df['is_smoking'].map({'YES':1,'NO':0})
df['sex']=df['sex'].map({'M':0,'F':1})
```

FEATURE SELECTION & SCALING

In Feature selection we remove non-informative or redundant predictors from the model.

After selecting the features and splitting them into training & testing datasets(80:20), we scaled the data by using Standard Scaler on our independent features.

Implementing LOGISITIC REGRESSION

After fitting Logistic Regression on our training datasets and predicting our y_pred on X_train and X_test our accuracy was 85.57% on both training and testing but in classification problems we cannot rely on accuracy solely so in our confusion matrix on training:-

[2166,10]
[358, 17]

Here we can see in the diagonal of this matrix that it is indicating a class imbalanced problem [534] True Positive and [12] True Negative and to clarify this we will use classification report on training :-

	precision	recall	f1-score	support
0	0.86	1.00	0.92	2176
1	0.63	0.05	0.08	375

CONTINUED

confusion matrix on testing :-

```
[534, 4]
[88, 12]
```

Here we can see in the diagonal of this matrix that it is indicating a class imbalanced problem [534] True Positive and [12] True Negative and to clarify this we will use classification report on testing :-

```
              precision    recall  f1-score   support

           0       0.86      0.99      0.92       538
           1       0.75      0.12      0.21       100
```

As, we can see recall and f1-score for class 1 is very low so, the results are clearly biased in order to solve this we will use **Smote Technique.** To use Smote Technique we are using SMOTETomek module from imblearn.combine.

Implementation of Logistic Regression after using Smote Technique.

Earlier our X & y shape was `(3189, 15), (3189,)` but after using Smote Technique it turn to (5380, 15), (5380,). So, again we will split & scale the data as before and fit our Logistic Regression on training data.
Our y_pred on X_train and X_test our accuracy was 66.82% and 67.47% on both training and testing.
Confusion Matrix on training :-

```
[1418, 730]
[698, 1458]
```

```
Classification Report on training : -
              precision    recall  f1-score   support

           0       0.67      0.66      0.67      2148
           1       0.67      0.68      0.67      2156
```

CONTINUED

Confusion Matrix :-

[361, 181]
[169, 365]

```
Classification Report: -
              precision    recall  f1-score   support

           0       0.68      0.67      0.67       542
           1       0.67      0.68      0.68       534
```

Now, the precision & recall improved by a lot on label 1 but still on the whole this should not be our final Model so we will try another model.

Implementing Random Forest Classifier

After fitting RFC on our training datasets and predicting our y_pred on X_train and X_test our accuracy was 73.03% and 50% on both training and testing we can see its clearly underfitted but in classification problems we cannot rely on accuracy solely so in our
confusion matrix on training :-

```
[1498 650]
[499 1657]
```

Classification Report on training:-

	precision	recall	f1-score	support
0	0.75	0.70	0.72	2148
1	0.72	0.77	0.74	2156

CONTINUED

confusion matrix:-

- **[0 534]**
- **[4 538]**

This is again not the best confusion matrix.

Classification Report

	precision	recall	f1-score	support
0	1.00	0.01	0.01	542
1	0.50	1.00	0.67	534

On the whole, this model is more poorer than Logistic Regression in our case so, we will definitely reject this model.

Implementing SVM

After fitting SVM on our training datasets and predicting our y_pred on X_train and X_test our accuracy was 71.11% and 72.21% on both training and testing but in classification problems we cannot rely on accuracy solely so in our confusion matrix for training :-

```
[1496 591]
[652 1565]
```

confusion matrix for testing :-

```
[383 140]
[159 394]
```

Confusion Matrix has improved by a lot compared to our both models used. Let's see our classification report

Classification Report on training :-

	precision	recall	f1-score	support
0	0.70	0.72	0.71	2087
1	0.73	0.71	0.72	2217

Continued

Classification report on testing :-

	precision	recall	f1-score	support
0	0.71	0.73	0.72	523
1	0.74	0.71	0.72	553

Comparing both classification report its not that different and its improved by a lot compared to both models used.

So, our next step would be to try Hyperparameter tune it and get the best parameters and see the results.

Hyperparameter Tuning

For Hyperparameter tuning we would like to use Grid Search CV.
So, to import this we will use sklearn library.

Firstly, we will assign our parameters of SVM :-

```
param_grid={'C':[0.1,1,10,100,1000],'gamma':[1,0.1,0.001,0.0001],'kernel':['rbf']}
```

Now, using GridSearchCV we will fit our estimator which is SVM and these parameters.

Our best parameters were :-

```
{'C': 10, 'gamma': 1, 'kernel': 'rbf’}
```

Now, we will use these parameters again perform SVM.

Implementing SVM after using GridSearchCV

After fitting SVM on our training datasets and predicting our y_pred on X_train and X_test our accuracy was 99.97% and 93.12% on both training and testing but in classification problems we cannot rely on accuracy solely so in our confusion matrix for training :-

```
[2147 0]
[1 2156]
```

Our False Positive and False Negative has decreased by a lot for training.

confusion matrix for testing :-
[517 49]
[25 485]

Same for testing our FP & FN has decreased by a lot.

Continued

Classification Report on Training:-

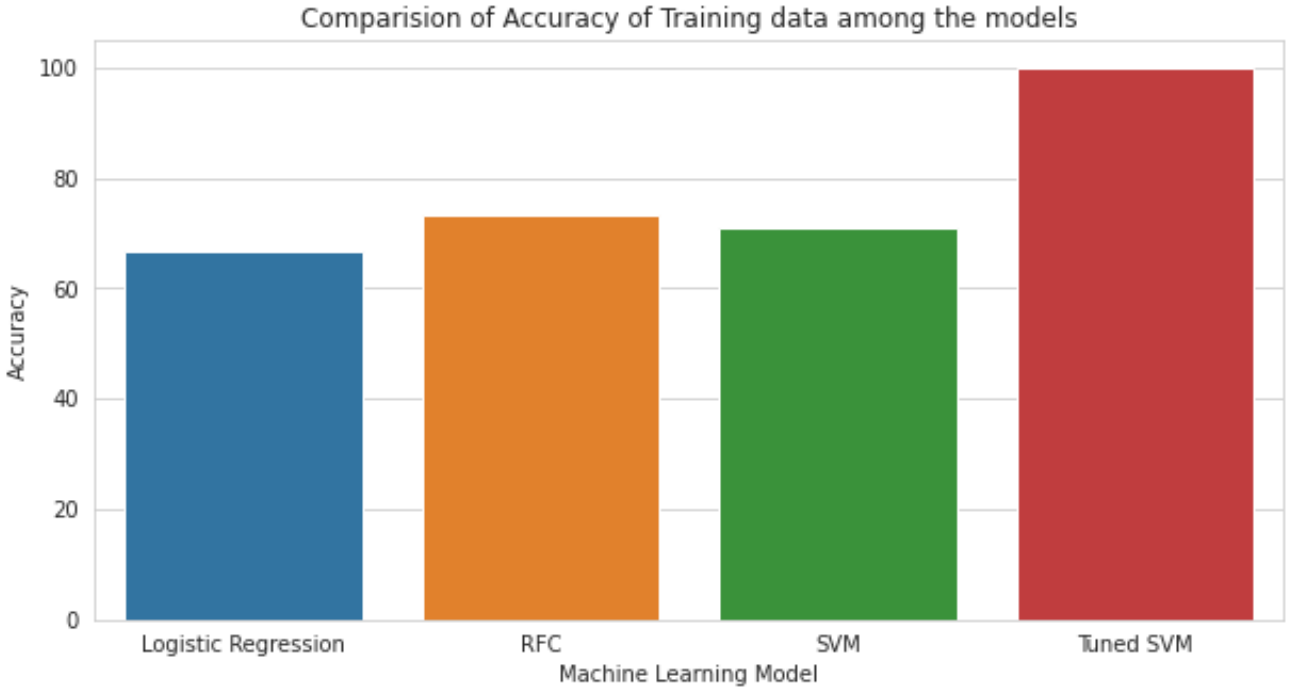
	precision	recall	f1-score	support
0	1.00	1.00	1.00	2147
1	1.00	1.00	1.00	2157

Classification Report on Testing :-

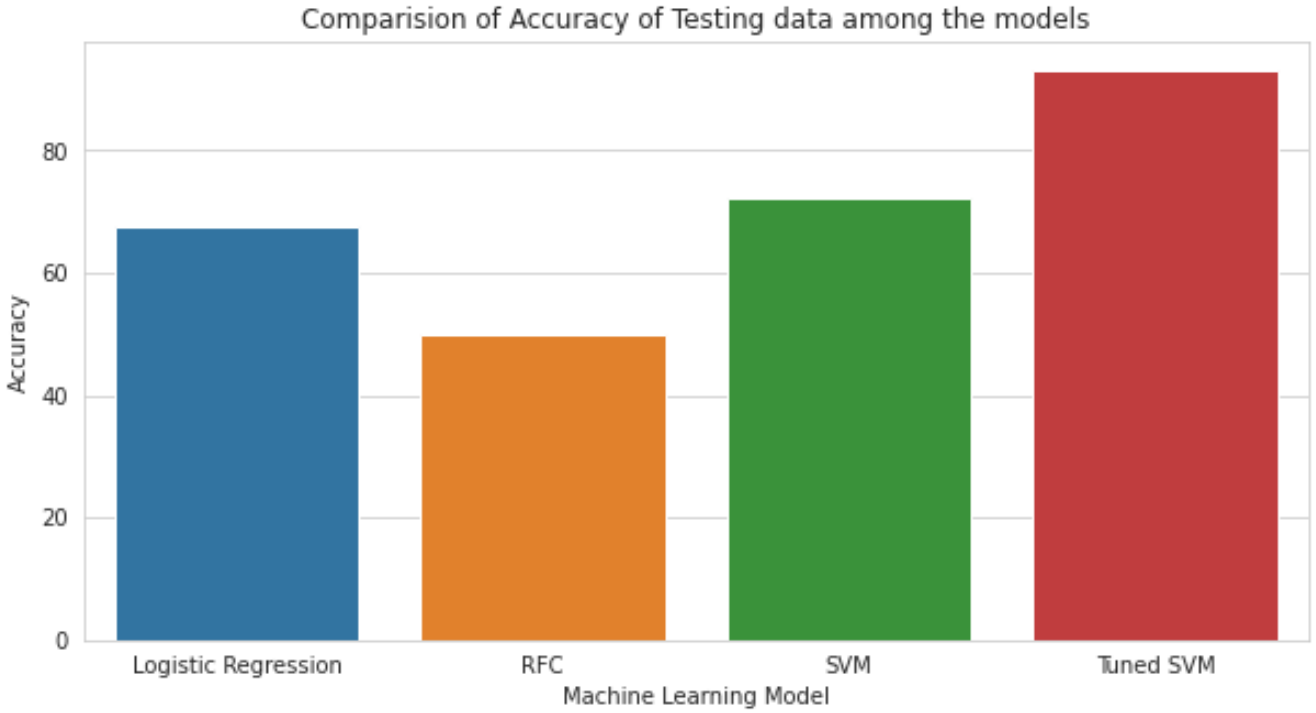
	precision	recall	f1-score	support
0	0.95	0.91	0.93	566
1	0.91	0.95	0.93	510

After using the best parameters our Classification Report on both labels has improved by a lot and we have got the best model which can be used for predicting.

COMPARISION OF EVALUATION METRICS AMONG THE MODELS BEING USED



COMPARISION OF EVALUATION METRICS AMONG THE MODELS BEING USED



CONCLUSION

We have used three models – Logistic Regression, Random Forest Classification, Support Vector Machine. Firstly, by applying Logistic Regression & looking at the Evaluation Metrics we came to a conclusion that it's a Imbalanced dataset as total no. of values in one label is much higher than the other.

After that we have applied Smote Technique to deal with imbalanced dataset problem and after using we again start modelling and on in Logistic Regression and Random Forest Classification we were not satisfied with the metrics so we tried SVM and it gave better results but there was need to tune it.
So, by using GridSearchCV we got the best parameters and after using those parameters our metrics got improved by a lot, so we finalised SVM as our final model.

We can conclude a patient's is at risk more when they are old as when a person gets old their Cholesterol level increases, whether there was prevalent hypertension and their Systolic and Diastolic Blood Pressure is high.

THANKYOU