### CREDIT CARD DEFAULTERS USINGML

##### APROJECTREPORT

*Submittedby*

**Eedara Revanth [Reg No: RA2011042010133] SrikanthS. [Reg No: RA2011042010088]**

*UndertheGuidanceof*

**Dr.Kanmani**

(AssistantProfessor,DepartmentofComputerScienceandEngineering)

*In partial fulfillment of the Requirements for the Degreeof*

#### BACHELOROFTECHNOLOGY



**DEPARTMENTOFCOMPUTERSCIENCEANDENGINEERING**

#### FACULTY OF ENGINEERING AND TECHNOLOGYSRMINSTITUTEOFSCIENCEANDTECHNOLOGYKATTANKULATHUR– 603203

##### November2022



SRM INSTITUTE OF SCIENCE AND TECHNOLOGYKATTANKULATHUR-603203

BONAFIDECERTIFICATE

Certifiedthatthisprojectreporttitled“**CREDIT CARD DEFAULTERS USINGML”**isthebonafideworkof**EedaraRevanth[RegNo:RA2011042010133]Srikanth [Reg No: RA2011042010088],** who carried out the project work undermy supervision. Certified further, that to the best of my knowledge the workreportedherein does not form part of any other thesis or dissertation on the basisof whicha degree or award was conferred on an earlier occasion for this or anyothercandidate.

Dr.Kanmani Dr.MLakshmi

##### SUPERVISOR HEADOFTHEDEPARTMENT

AssistantProfessor DepartmentofDataScience and

DepartmentofDataScienceand Business SystemBusiness System

SignatureofInternalExaminer SignatureofExternalExaminer



OwnWorkDeclaration

Department OfComputerScience andEngineering

**SRMINSTITUTEOFSCIENCEAND TECHNOLOGY**

**OwnWork Declaration Form**

This sheet must be filled in (each box ticked to show that the condition has been met). It must be signed anddated along with your student registration number and included with all assignments you submit – work willnotbe marked unlessthis isdone.

To be completed by the student for all assessmentsDegree/Course : B.TECH

**Student Name :**Eedara RevanthSrikanth

**Registration Number :**RA2011042010133, RA2011042010088

]**Titleof Work :**CREDIT CARD DEFAULTERUSINGML

I / We hereby certify that this assessment compiles with the University’s Rules and Regulations relating toAcademic misconduct and plagiarism\*\*, as listed in the University Website, Regulations, and the EducationCommittee guidelines.

I/Weconfirmthatalltheworkcontained inthisassessmentismy /ourownexceptwhereindicated,and thatI

/Wehave metthe followingconditions:

* Clearlyreferences/listed allsourcesasappropriate
* Referencedandputininvertedcommasallquotedtext(from books,web,etc)
* Given thesourcesofallpictures, dataetc. thatarenotmyown
* Notmadeanyuseof thereport(s) oressay(s)ofanyotherstudent(s)eitherpastorpresent
* AcknowledgedinappropriateplacesanyhelpthatIhavereceivedfromothers(e.g.fellowstudents, technicians, statisticians, external sources)
* Compiledwith anyotherplagiarismcriteriaspecifiedin theCoursehandbook/Universitywebsite

IunderstandthatanyfalseclaimforthisworkwillbepenalisedinaccordancewiththeUniversitypoliciesandregulations.

|  |
| --- |
| **DECLARATION:** |
| I am aware of and understand the University’s policy on Academic misconduct and plagiarism and I certify thatthis assessment is my / our own work, except where indicated by referring, and that I have followed the goodacademic practices noted above. |
| Ifyouareworkingin agroup,pleasewriteyourregistrationnumbersandsign withthedateforeverystudentin  yourgroup. |

Eedara Revanth Srikanth

##### ACKNOWLEDGEMENT

Weexpressourhumblegratitudeto ,ViceChancellor(I/C),SRMInstituteof Science and Technology, for the facilities extended for the project work and his continuedsupport.

Weextendoursincerethanksto ,Professor&Chairperson,SchoolofComputing,SRM Institute ofScienceand Technology,forher invaluablesupport.

Wewishtothank ,Professor&Head,DepartmentofComputerScienceandEngineering, SRM Institute of Science and Technology, for her valuable suggestions andencouragementthroughout the period ofthe project work.

Weareextremely gratefultoourAcademicAdvisor ,AssistantProfessor,Department of Computer Science and Engineering, SRM Institute of Science and Technology,for her great support at all the stages of project work. We would like to convey our thanks toourPanelHead ,AssociateProfessor,DepartmentofComputerScienceandEngineering, SRM Institute of Science and Technology, for his inputs during the projectreviews.

WeregisterourimmeasurablethankstoourFacultyAdvisor, ,AssistantProfessor, Department of Computer Science and Engineering, SRM Institute of Science andTechnology,forleading and helping us to complete ourcourse.

Ourinexpressiblerespectandthankstomyguide, **,**AssistantProfessor,Department of Computer Science and Engineering, SRM Institute of Science and Technology,for providing me an opportunity to pursue my project under her mentorship. She provided methe freedom and support to explore the research topics of my interest. Her passion for solvingrealproblems and makingadifferencein theworld has always beeninspiring.

We sincerely thank staff and students of the Computer Science and Engineering Department,SRMInstituteofScience andTechnology,fortheirhelpduringmyresearch.Finally,wewouldlike to thank my parents, our family members and our friends for their unconditional love,constantsupport and encouragement.

EedaraRevanth (RA2011042010133) Srikanth (RA2011042010088)

##### ABSTRACT

As an increasing number of purchasers depend upon the credit score card to pay their regular purchases in on line and bodily retail store, the quantity of issued credit score playing cards and the overpowering quantity of credit score card debt via way of means of the cardholders have hastily accelerated. Therefore, maximum monetary establishments need to address the problems of credit score card default similarly to the credit score card fraud together with credit score card dump. Both the credit score card verification implemented to the cardholders and the default threat control after card issued are essential to the destiny achievement of maximum monetary establishments According to the Federal Reserve financial statistics, the default charge on credit score loans throughout all industrial banks is at an all-time excessive for the beyond sixty-six months, and it\'s miles possibly to hold to climb in the course of 2022. The delinquency charge suggests the share of beyond-due loans in the borrower’s whole mortgage portfolio. The mountain climbing delinquencies will bring about a good sized amount of cash lose from the lending establishments, together with industrial banks. Therefore, banks ought to have a threat prediction version and be capable of classify the maximum relative traits which can be indicative of humans who\'ve a better possibility of default on credit score. In 2013, patron spending encompassed about 69% of USA gross home product. Of the $3.098 trillion of tremendous patron credit score within the United States within the closing region of 2013, they have been revolving credit score card for over 25% of it ($857.6 billion). A small boom within the accuracy of figuring out excessive-threat loans may want to save you losses of over $eight billion. Because of the dangers inherent in one of these huge part of the economy, constructing fashions for patron spending behaviors to restrict threat exposures on this region is turning into extra critical. For this to be a feasible option, the predictions want to be moderately correct. A strong version isn\'t most effective a beneficial device for the lending establishments to determine on credit score applications, however it could additionally assist the customers to be privy to the behaviors that could harm their credit score scores. The number one motivation at the back of threat prediction is to make use of monetary statistics, for instance, enterprise transactional statistics, change statistics and patron transactions, and so on to foresee the patron’s enterprise overall performance or character credit score card statistics and to lower bathrooms and vulnerability. Several threat prediction fashions are primarily based totally on statistical methods, inclusive of nearest neighbor, discriminant evaluation, and logistic regression. The aim of credit score default prediction is to assist monetary establishments determine whether or not or now no longer to lend to a patron. The ensuing check is mostly a threshold price that lets in the choice-makers to make the lending choice. The well-known version relies upon at the monetary ratios, earnings account, and statistics at the stability sheet..

|  |  |
| --- | --- |
| **LISTOFFIGURES:** |  |
| Figure1:OriginalDatasetSnapshot | 4 |
| Figure2:BarGraphoftheTargetClassesAfterDropping | 7 |
| Figure3:Bar GraphoftheTarget ClassesBeforeDropping | 7 |
| Figure4:DatasetafterScalingandImputing | 7 |
| Figure5:CorrelationMatrixVisualization | 8 |
| Figure6:ResultfromFeatureSelectionusingBackwardEliminationMethod | 9 |
| Figure7:DatasetAfterDroppingColumnsafterFeatureSelection | 9 |
| Figure8:Top10importantfeaturessupportedbyRFECV | 1 |

#### LISTOFTABLES

Table1:ConfusionMatrixObtainedaftertrainingthedata(featureselectionbybackwardelimination)

................................................................................................................................... 11

Table2:ConfusionMatrixObtainedaftertrainingthedata(featureselectionbyRFECVmethod)

.......................................................................................................................................................11

Table 3: Comparison between the feature selection models after training and testing through LogisticRegressionmodel 13

Table4:WorkDivision 14

Table5:Majormodules andclassesused fromSklearn 15

#### LISTOFSYMBOLSANDABBREVIATIONS

1. IDE:IntegratedDevelopmentEnvironment
2. REFCV:RecursiveFeatureEliminationusingCross-Validation
3. CV:CrossValidation
4. RFE:RecursiveFeatureElimination

#### CHAPTER INTRODUCTION

As increasingly more purchasers depend upon the credit score card to pay their ordinary purchases in on line and bodily retail store, the quantity of issued credit score playing cards and the overpowering quantity of credit score card debt via way of means of the cardholders have swiftly increased. Therefore, maximum economic establishments need to address the troubles of credit score card default further to the credit score card fraud which include credit score card dump. Both the credit score card verification carried out to the cardholders and the default hazard control after card issued are critical to the destiny achievement of maximum economic establishments

According to the Federal Reserve financial statistics, the default charge on credit score loans throughout all industrial banks is at an all-time excessive for the beyond sixty-six months, and its miles in all likelihood to preserve to climb at some stage in 2022. The delinquency charge suggests the share of beyond-due loans within the borrower’s complete mortgage portfolio. The mountaineering delinquencies will bring about a huge amount of cash lose from the lending establishments, which include industrial banks. Therefore, banks should have a hazard prediction version and be capable of classify the maximum relative traits which are indicative of humans who've a better chance of default on credit score.

In 2013, purchaser spending encompassed about 69% of USA gross home product. Of the $3.098 trillion of wonderful purchaser credit score within the United States within the remaining area of 2013, they have been revolving credit score card for over 25% of it ($857.6 billion).

A small growth within the accuracy of figuring out excessive-hazard loans may want to save you losses of over $eight billion. Because of the dangers inherent in this kind of massive part of the economy, constructing fashions for purchaser spending behaviors to restrict hazard exposures on this area is turning into extra critical. For this to be a possible option, the predictions want to be moderately accurate. A strong version isn't simplest a beneficial device for the lending establishments to determine on credit score applications, however it may additionally assist the customers to be privy to the behaviors which can harm their credit score scores. The number one motivation in the back of hazard prediction is to make use of economic statistics, for example, enterprise transactional statistics, alternate statistics and patron transactions, and so on to foresee the patron’s enterprise overall performance or person credit score card statistics and to lower bogs and vulnerability. Several hazard prediction fashions are primarily based totally on statistical methods, which includes nearest neighbor, discriminant analysis, and logistic regression. The intention of credit default prediction is to assist economic establishments determine whether or not or now no longer to lend to a patron. The ensuing check is often a threshold fee that lets in the decision-makers to make the lending decision. The popular version relies upon at the economic ratios, profits account, and statistics at the stability sheet.

## ProblemDefinition

Credit card default are increasing heavily because of fraud financial loss increasing drastically. Every year due to fraud, Billions of amounts loss. To predict the default in credit card, there is lack of research. Taking this point in mind we are developing a system which can predict and find the defaulter person, so that financial sector may secure from this fraud.

## Motivation

Machine learning techniques have been around us and has been compared and used for analysis for many kinds ofdata science applications. The major motivation behind this research-based project was to explore the featureselection methods, data preparation and processing behind the training models in the machine learning. With firsthand models and libraries, the challenge we face today is data where beside their abundance, and our cookedmodels, the accuracy we see during training, testing and actual validation has a higher variance. Hence this projectis carried out with the motivation to explore behind the models, and further implement Logistic Regression 2model to train the obtained data. Furthermore, as the whole machine learning is motivated to develop anappropriate computer-based system and decision support that can aid to early detection of credit card defaulter.,

## Objectives

Themainobjectiveof developingthisprojectare:

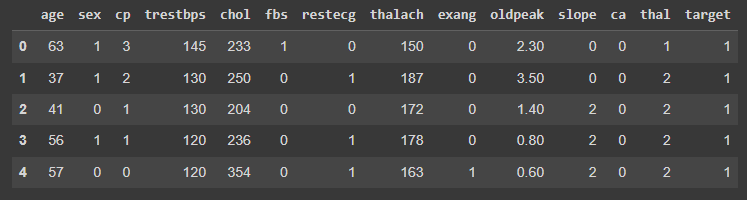
1. To develop machine learning model to predict future possibility of heart disease by implementing LogisticRegression.
2. Todeterminesignificantriskfactorsbasedonmedicaldatasetwhichmayleadtoheartdisease.
3. Toanalyze featureselection methodsand understandtheirworkingprinciple.

### CHAPTER 2RELATEDWORK

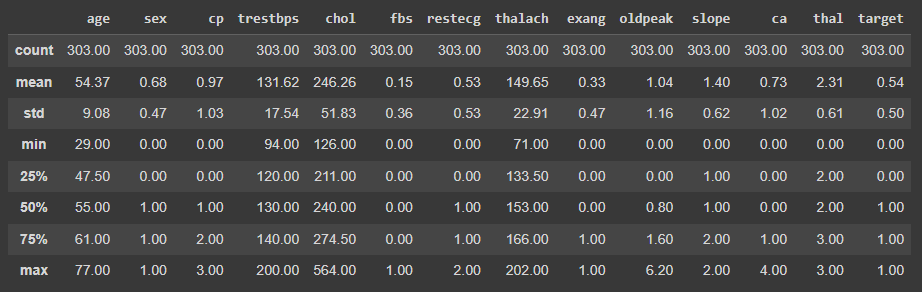
With growing development in the field of medical science alongside machine learning variousexperiments and researches has been carried out in these recent years releasing the relevantsignificant papers. The paper [1] propose heart disease prediction using KStar, J48, SMO, andBayes Net and Multilayer perceptron using WEKA software. Based on performance fromdifferent factor SMO (89% of accuracy) and Bayes Net (87% of accuracy) achieve optimumperformance than KStar, Multilayer perceptron and J48 techniques using k-fold cross validation.The accuracy performance achieved by those algorithms are still not satisfactory. So that if theperformanceofaccuracyisimproved moretogivebatterdecisionto diagnosis disease.

1. In a research conducted using Cleveland dataset for heart diseases which contains 303instances and used 10-fold Cross Validation, considering 13 attributes, implementing 4 differentalgorithms, they concluded Gaussian Naïve Bayes and Random Forest gave the maximumaccuracy of91.2percent.
2. Using the similar dataset of Framingham, Massachusetts, the experiments were carried outusing 4 models and were trained and tested with maximum accuracy K Neighbors Classifier:87%, Support Vector Classifier: 83%, Decision Tree Classifier: 79% and Random ForestClassifier:84%.

**CHAPTER 3DATASETS**

The dataset is publicly available on the Kaggle Website at [4] which is from an ongoingcardiovascular study on residents of the town of Framingham, Massachusetts. It provides patientinformation which includes over 4000 records and 14 attributes. The attributes include: age, sex,chest pain type, resting blood pressure, serum cholesterol, fasting, sugar blood, restingelectrocardiographic results, maximum heart rate, exercise induced angina, ST depression inducedby exercise, slope of the peak exercise, number of major vessels, and target ranging from 0 to 2,where 0 is absence of heart disease. The data set is in csv (Comma Separated Value) format whichis further preparedtodataframeassupportedbypandaslibrary inpython.



**Datasetdescription**

The education data is irrelevant to the heart disease of an individual, so it is dropped. Further withthisdatasetpre-processing andexperiments arethen carriedout.

## CHAPTER 4

**METHODANDALGORTIHM USED**

The main purpose of this project is credit card default are increasing heavily because of fraud financial loss increasing drastically. Every year due to fraud, Billions of amounts loss. To predict the default in credit card, there is lack of research. Taking this point in mind we are developing a system which can predict and find the defaulter person, so that financial sector may secure from this fraud.

## LogisticRegression

Logistic Regression is a supervised classification algorithm. It is a predictive analysis algorithm based on the concept ofprobability. It measures the relationship between the dependent variable (TenyearCHD) and the one or more independentvariables (risk factors) by estimating probabilities using underlying logistic function (sigmoid function). Sigmoid function isusedasacostfunctiontolimit thehypothesis oflogisticregressionbetween0and1(squashing) i.e.0≤ hθ(x)≤ 1.

Inlogisticregressioncost functionisdefinedas:



Logistic Regression relies highly on the proper presentation of data. So, to make the model more powerful, importantfeaturesfrom theavailabledataset areselectedusingBackwardeliminationandrecursiveeliminationtechniques.

RANDOM FOREST:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, ***"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."*** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.**The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.**

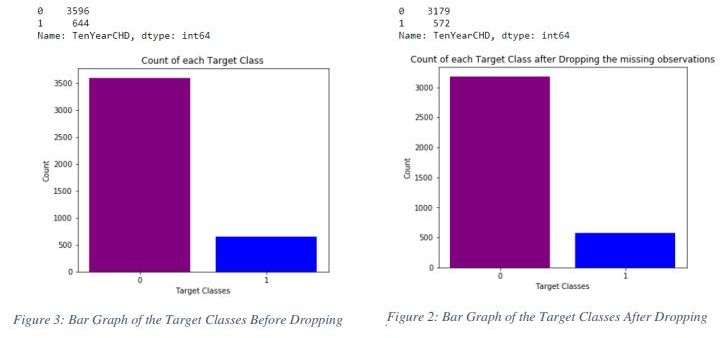
**DECISION TREE:**

The Decision Tree Algorithm is one of the popular supervised type machine learning algorithms that is used for classifications. This algorithm generates the outcome as the optimized result based upon the tree structure with the conditions or rules. The decision tree algorithm associated with three major components as Decision Nodes, Design Links, and Decision Leaves. It operates with the Splitting, pruning, and tree selection process. It supports both numerical and categorical data to construct the decision tree. Decision tree algorithms are efficient for large data set with less time complexity. This Algorithm is mostly used in customer segmentation and marketing strategy implementation in the business.

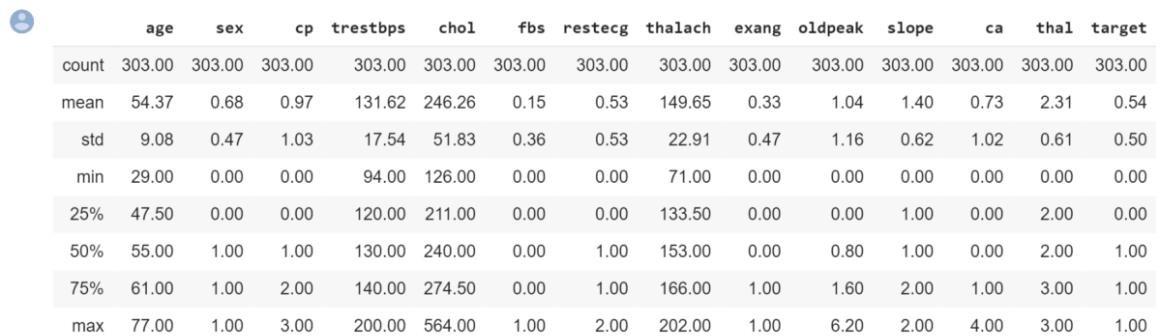
# CHAPTER 5EXPERIMENTS

## Data Preparation

Since the dataset consists of 4240 observations with 388 missing data and 644 observations to be risked for heart disease,two different experiments were performed for data preparation. First, we checked by dropping the missing data, leaving withonly3751 dataand only 572 observations risked forheartdisease.

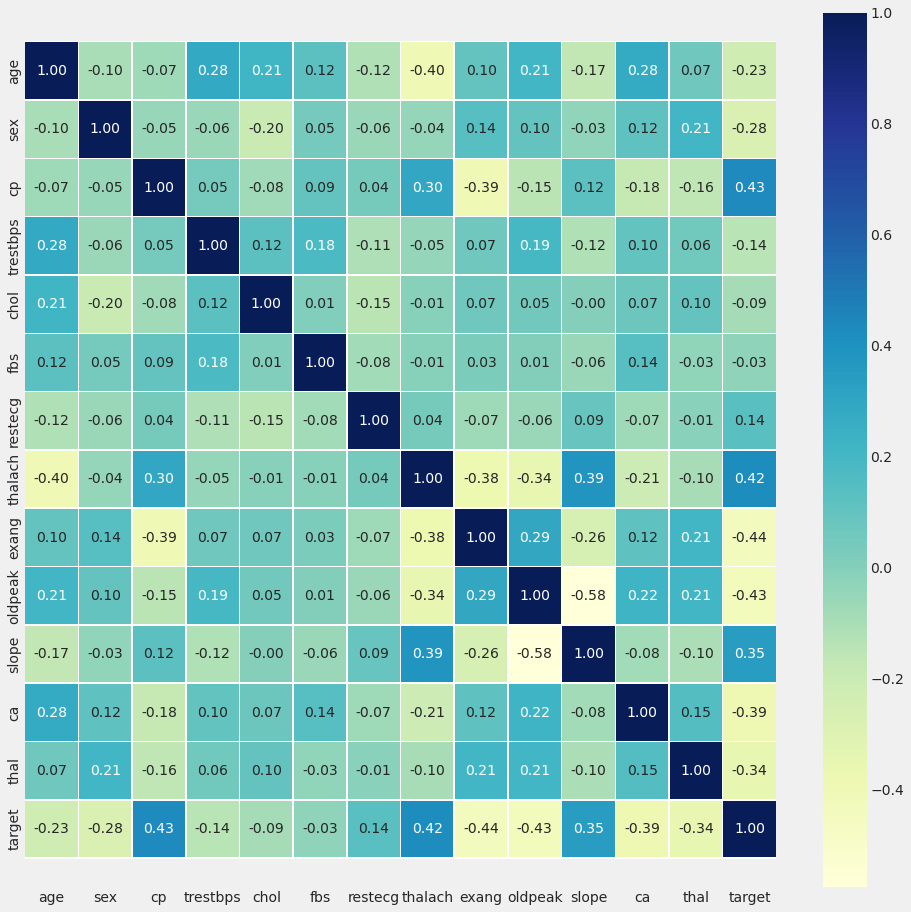


This leads to reduced number of the observations providing irrelevant training to our model. So, we progressed withimputation of data with the mean value of the observations and scaling them using SimpleImputer and StandardScalermodulesof Sklearn.



## ExploratoryAnalysis:

CorrelationMatrixvisualizationBeforeFeatureSelectionshows



It shows that there is no single feature that has a very high correlation with our target value. Also, some of the features havea negative correlation with the target value and some have positive. The data was also visualized through plots and bargraphs.

## FeatureSelection

**FeatureSelectionusingBackwardElimination(P-value) algorithm:**

Further the data was passed through the backward elimination function to select the most relevant features which gavefollowingresult:

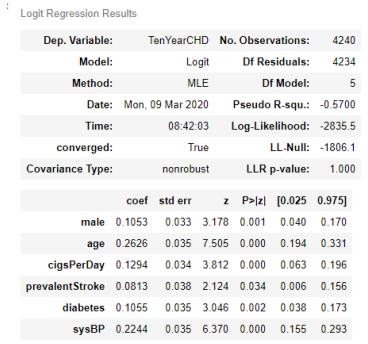


Figure 6: Result from Feature Selection using Backward Elimination MethodAccordingtheresultabovethe columnsweredropped.

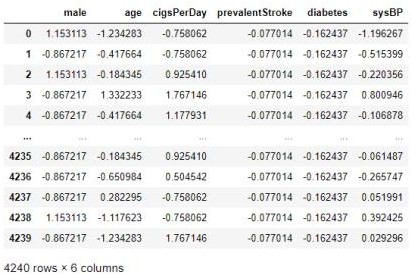


Figure7:DatasetAfterDroppingColumnsafterFeatureSelection

## Trainingand testing

Finally, this resulting data split into 70% train and 30% test data, which was further passed to theLogisticRegressionmodelto fit, predictandscore themodel.

### CHAPTER6

**EVALUATION METRICS**

Fortheevaluationofouroutput fromourtrainingthedata,theaccuracywasanalyzed“Confusionmatrix”.

## Confusion Matrix

A confusion matrix, also known as an error matrix, is a table that is often used to describe the performance of a classificationmodel(or“classifier”)onasetof testdataforwhichthetruevaluesareknown.Itallowsthevisualizationof theperformance of an algorithm. It allows easy identification of confusion between classes e.g. one class is commonlymislabeled as the other. The key to the confusion matrix is the number of correct and incorrect predictions are summarizedwithcountvaluesand broken down byeachclassnotjustthe numberoferrorsmade.

|  |  |
| --- | --- |
| TP=3569 | FP=27 |
| FN=599 | TN=45 |

Table1: ConfusionMatrixObtainedaftertrainingthedata(featureselectionbybackwardelimination)

## Accuracy

Theaccuracyis calculatedas:

Accuracy= 𝑇𝑝+𝑇𝑁

𝑇𝑝+𝑇𝑁+𝐹𝑃+𝐹𝑁

Where,

* TruePositive(TP)=Observationispositive,andispredictedtobepositive.
* FalseNegative(FN)=Observationispositive,butispredictednegative.
* TrueNegative(TN)=Observationisnegative, and ispredictedtobenegative.
* FalsePositive(FP)=Observationisnegative,but ispredictedpositive

The obtained accuracy during training the data after feature selection using backward elimination was 86 % and duringtestingwas 83%.

## Recall

Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number ofpositiveexamples.HighRecall indicatestheclassiscorrectlyrecognized(asmallnumberofFN).Recalliscalculatedas:

Recall=

𝑇𝑃

𝑇𝑃+𝐹𝑁

The obtained recall during training the data after feature selection using backward elimination was and during testing was0.99.

## Precision

To get the value of precision we divide the total number of correctly classified positive examples by the total number ofpredicted positive examples. High Precision indicates an example labelled as positive is indeed positive (a small number ofFP). Precisioniscalculatedas:

Precision=

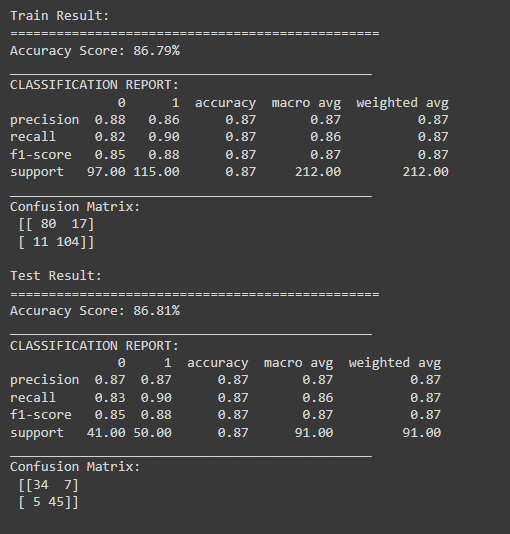
𝑇𝑃

𝑇𝑃+𝐹𝑃

The obtained precision during training the data after feature selection using backward elimination was 0.86 and duringtestingwas 0.84.

# CHAPTER 6DISCUSSIONONRESULTS

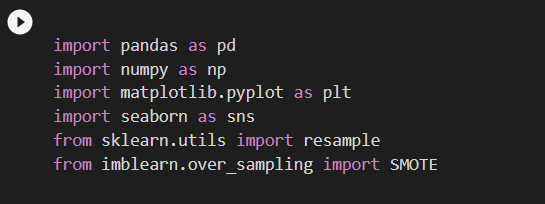
When performing various methods of feature selection, testing it was found that backward elimination gave us thebest results among others. The various methods tried were Backward Elimination with and without KFold,Recursive Feature Elimination with Cross Validation. The accuracy that was seen in them ranged around 86%with 87% being maximum. Though both methods gave similar accuracy but it was seen that in BackwardElimination we found that the number of misclassifications of True Negative was more and it was observed thatthe accuracy had more variance compared to RFEV. The precision of Backward Elimination is 84 and the recallsis 0.83.



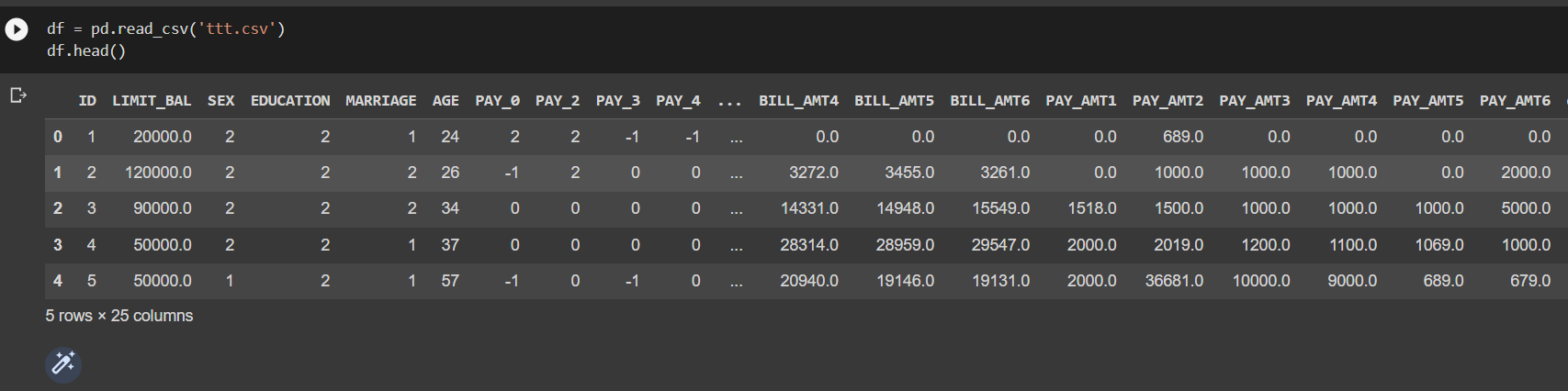
### CHAPTER 9:CODE

The coding portion were carried out to prepare the data, visualize it, pre-process it, building the model and then evaluating it.The code has been written in Python programming language using Google colab. The experiments and all themodelsbuilding are done based on pythonlibraries.

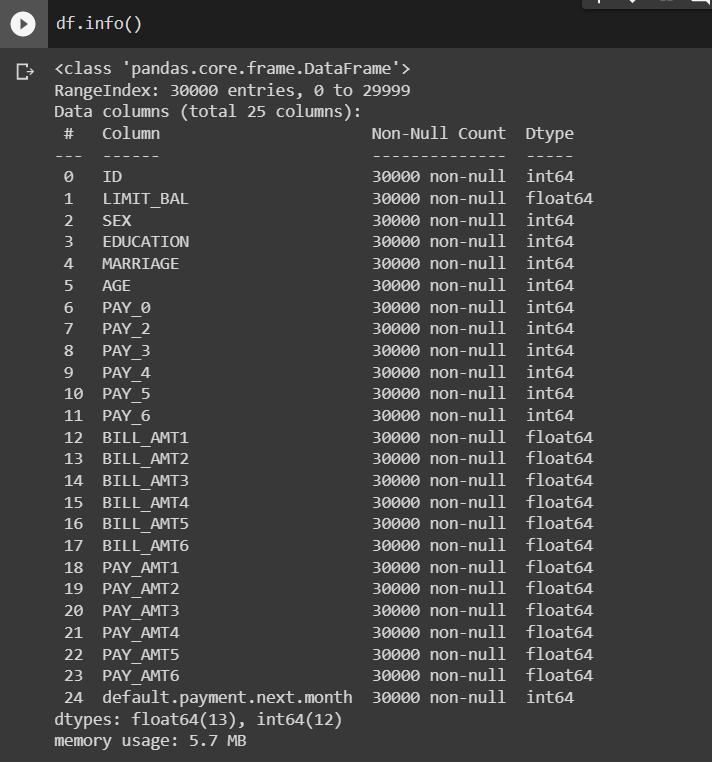
**Importingsomenecessarylibraries:**

****

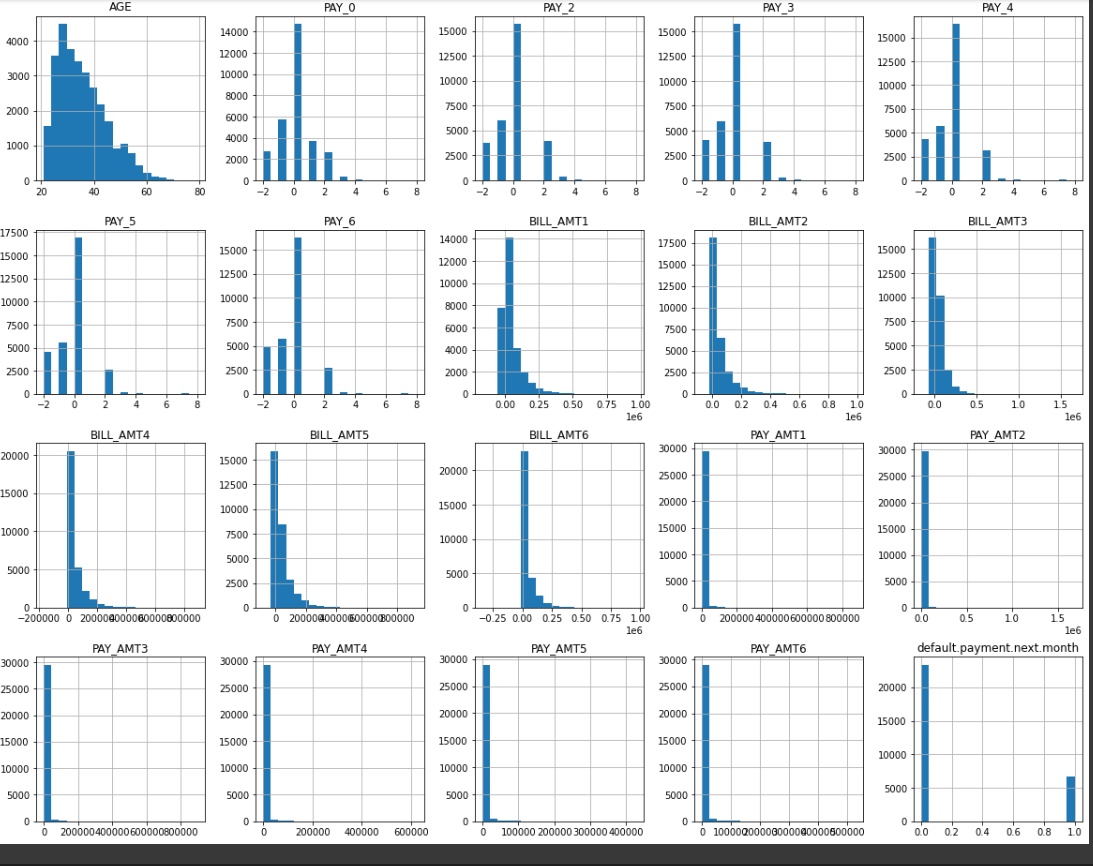
**Importing DATA:**

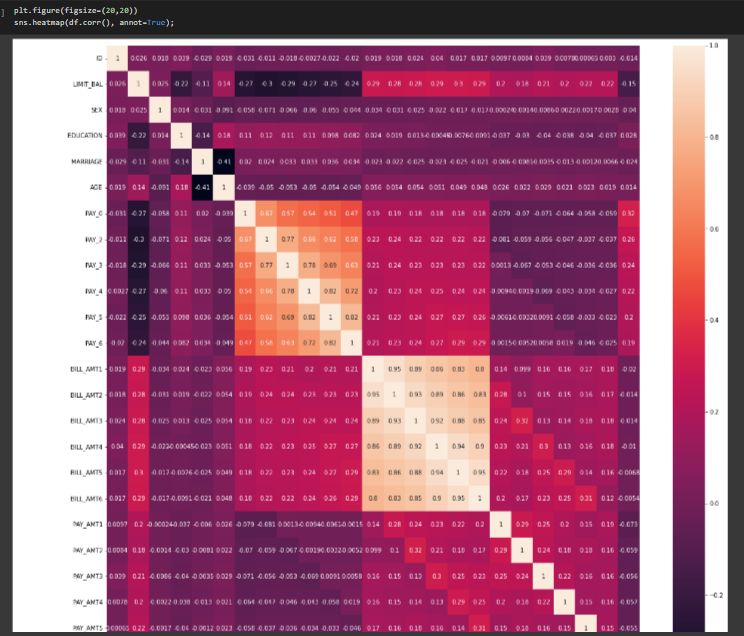
****

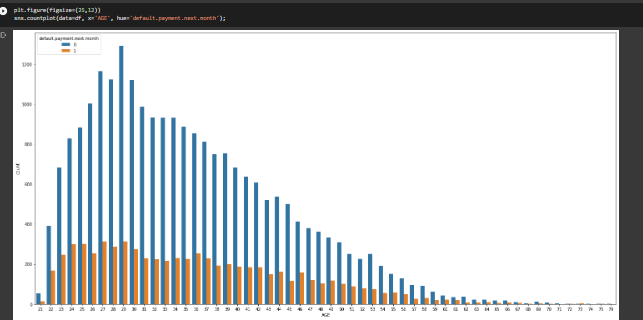
**DATASET INFO:-**

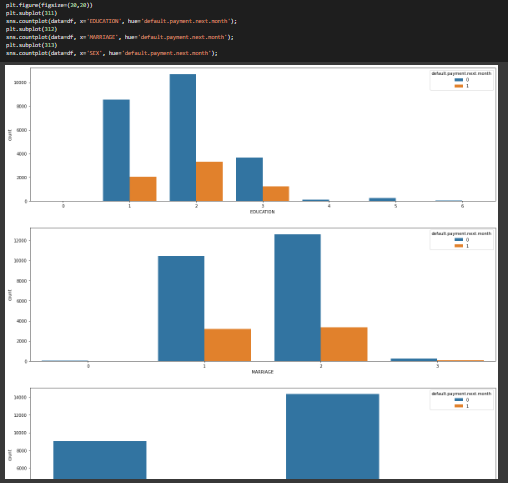
****

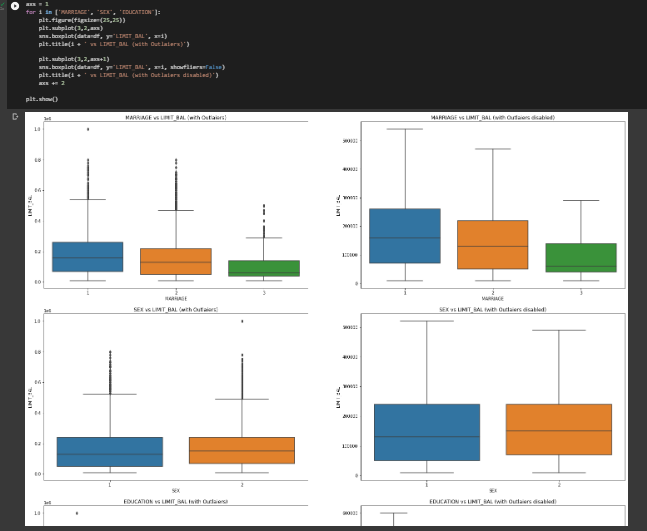
**ExploratoryDataAnalysis:**



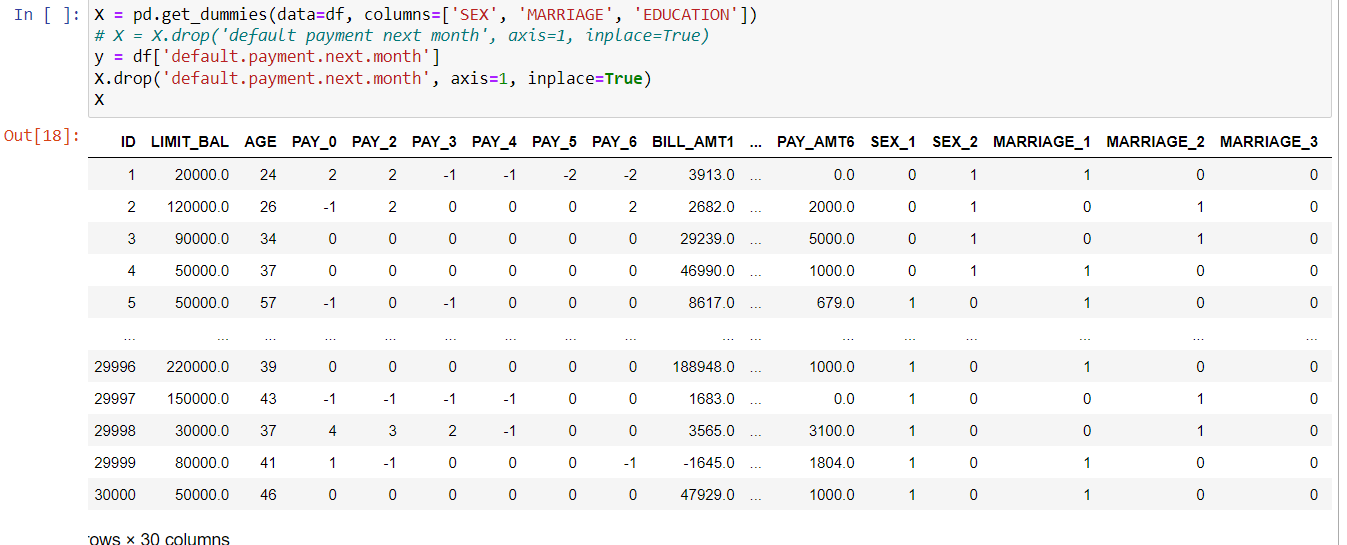
****

****

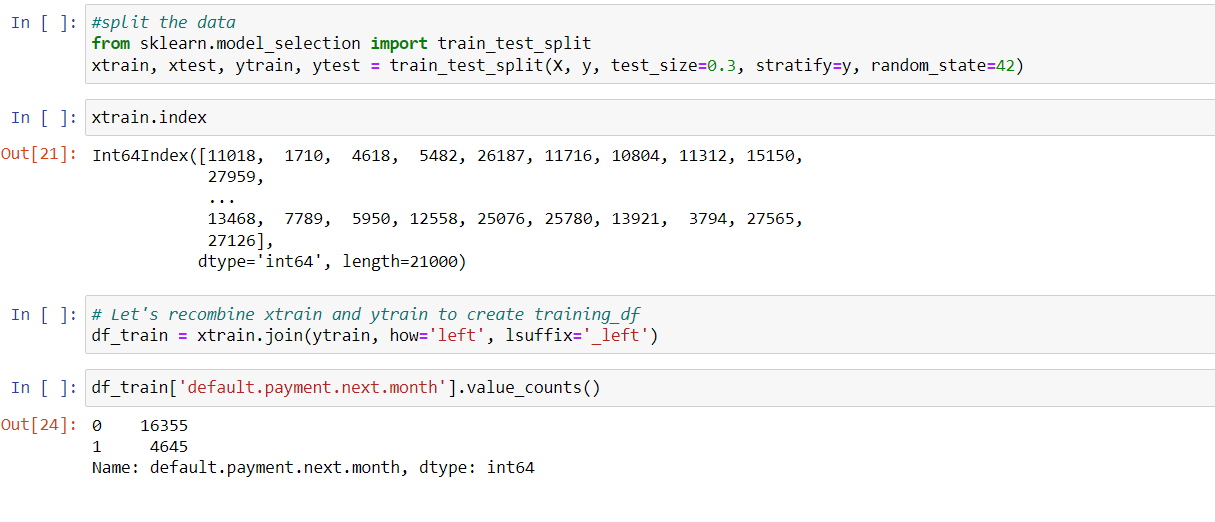
****

****

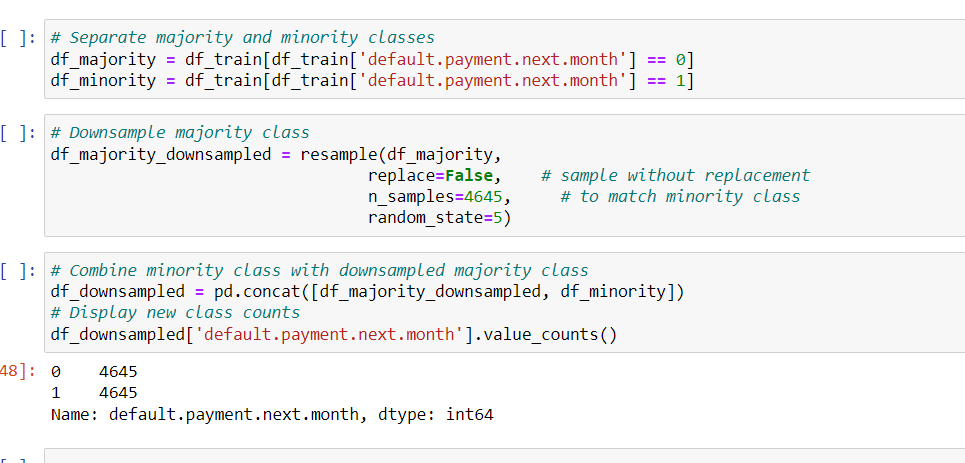
**ANALYSING THE DATA:**

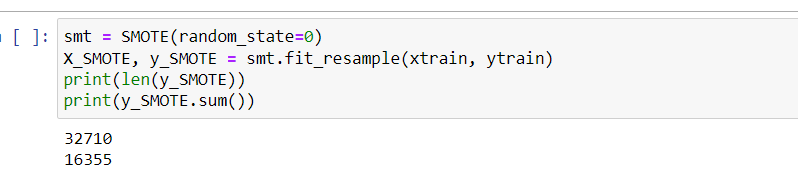
****

**TRAIN TEST AND SPLIT OF THE DATA:-**

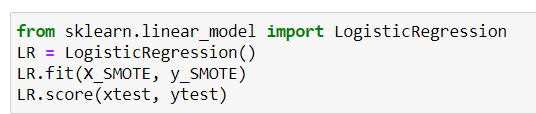
****

**As the data is imbalanced we use smot to arrange the data properly to avoid miss calculations**

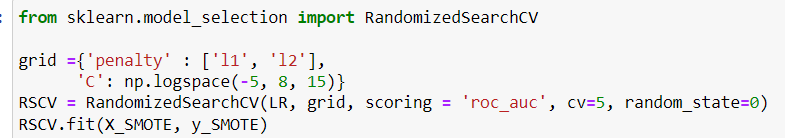
****

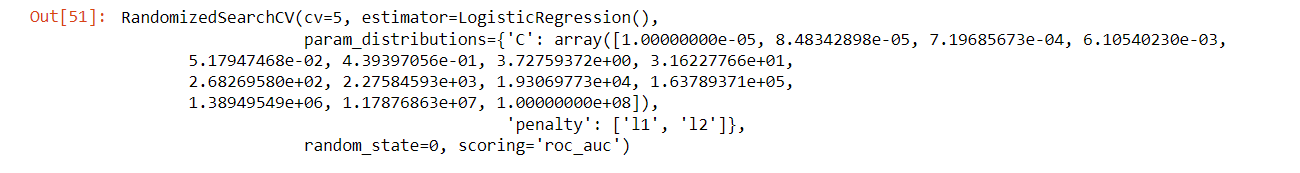
****

**LOGISTIC REGRESSION:**

****

****

****

****

**DECISION TREE CLASSIFIER:**

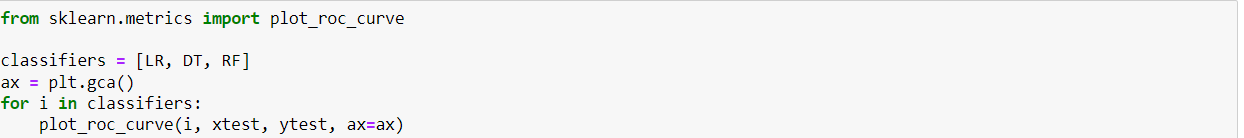
****

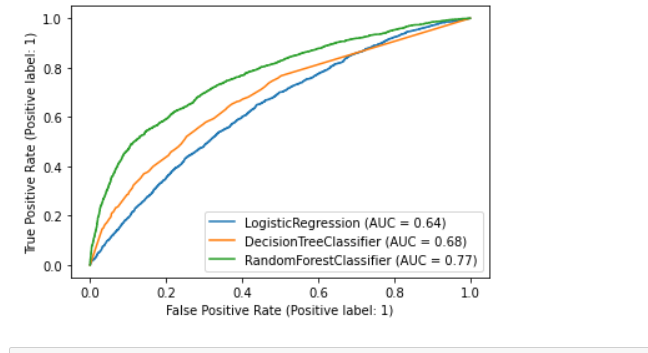
**RANDOM FOREST:**

****

****

**PLOT\_ROC\_CURVE FOR THE THREE REGRESSION MODELS:-**

****

****

### CHAPTER 10CONCLUSION

### CHAPTER 10CONCLUSION

The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes inhigh-risk patients and in turn reduce the complications, which can be a great milestone in the fieldof medicine. This project resolved the feature selection i.e., backward elimination behind themodels and successfully predict the heart disease, with 85% accuracy. The model used wasLogistic Regression. Further for its enhancement, we can train on models and predict the types ofcardiovascular diseases providing recommendations to the users, and also use more enhancedmodels.

### REFERENCES

* + 1. A. H. M. S. U. Marjia Sultana, "Analysis of Data Mining Techniques for Heart DiseasePrediction,"2018.
    2. M. I. K. ,. A. I. ,. S. Musfiq Ali, "Heart Disease Prediction Using Machine LearningAlgorithms".
    3. K. Bhanot, "towarddatascience.com," 13 Feb 2019. [Online]. Available:https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machinelearning-36f00f3edb2c.[Accessed2 March 2020].
    4. [Online]. Available: https:/[/www.kaggle.com/ronitf/he](http://www.kaggle.com/ronitf/heart-disease-uci#heart.csv)a[rt-disease-uci#heart.csv..](http://www.kaggle.com/ronitf/heart-disease-uci#heart.csv)[Accessed

05December2019].

* + 1. M. A. K. S. H. K. M. a. V. P. M Marimuthu, "A Review on Heart Disease Prediction usingMachine Learning and Data Analytics Approach".