



PROFESSIONAL TRAINING REPORT

at

Sathyabama Institute of Science and Technology
(Deemed to be University)

Submitted in partial fulfillment of the requirements for the
award of Bachelor of Engineering Degree in Information Technology

By,
Shajith ahamed
Reg.No:39120106



DEPARTMENT OF INFORMATION TECHNOLOGY

SCHOOL OF COMPUTING

**SATHYABAMA INSTITUTE OF SCIENCE AND
TECHNOLOGY**

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI – 600119, TAMILNADU**



NOVEMBER 2021

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY



(DEEMED TO BE UNIVERSITY)

Accredited with Grade “A” by NAAC

(Established under Section 3 of UGC Act, 1956)

JEPPIAAR NAGAR, RAJIV GANDHI SALAI

CHENNAI– 600119

www.sathyabama.ac.in

DEPARTMENT OF INFORMATION TECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Shajith ahamed** (Reg. No: **39120106**) who carried out the project entitled “**Identifying Credit Risks in Banking Domain Using Data Science Algorithm**” under my supervision from June 2021 to November 2021.

Internal Guide

Dr. ajitha,

Professor

Submitted for Viva voce Examination held on _____

Internal Examiner
Examiner

External

DECLARATION

I, **Shajith ahamed** hereby declare that the project report entitled **Identifying CreditRisks in Banking Domain Using Data Science Algorithm** done by me under the guidance of **DR.Ajith** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Information Technology

DATE:

PLACE:

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D, Dean**, School of Computing, **Dr.Aijth** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr.R.Subhashini** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Information Technology** who were helpful in many ways for the completion of the project.

TRAINING CERTIFICATE

CERTIFICATE OF COMPLETION

This certifies that

SHAJITH AHAMED S of Sathyabama Institute of Science and Technology

Has successfully completed Training Program on

Data Science Using R

From 9th Sep 2021 to 30th Oct 2021

A handwritten signature in black ink, appearing to read 'Dharani'.

Dr. Dharani Senthilkumar
CEO - Vani Analytics
<http://www.vanianalytics.com>

Identifying Credit Risks in Banking Domain Using Data Science Algorithm

ABSTRACT

Data Science can be used in various areas in financial sectors like customer segmentation, profitability, high risk loan applicants, credit analysis, fraudulent transactions, cash management and forecasting operations. There are numerous risks related to bank loans both for the banks and the borrowers getting the loans and the number of transactions in banking sector is rapidly growing and huge data volumes are available. Banks need to analyse their customers for loan eligibility so that they can specifically target good customers. In this project, an automated loan eligibility process model has been created, based on customer details such as Gender, Marital Status, Age, Occupation, Income, Property etc. The Decision Tree algorithm has been used to create the loan approval classification model. As the number of transactions in banking sector is rapidly growing and huge data volumes are available. Instead of using standardized schemes, this classification model can be used by bank managers for analysing the customers' behaviour easily and the risks around loan can be reduced.

TABLE OF CONTENTS		
CHAPTER	TITLE	PAGE NO
	ABSTRACT	
1	INTRODUCTION	
2	SYSTEM ANALYSIS	
	2.1 EXISTING SYSTEM	
	2.2 PROPOSED SYSTEM	
3	SYSTEM SPECIFICATION	
	3.1 HARDWARE SPECIFICATION	
	3.2 SOFTWARE SPECIFICATION	
4	SYSTEM IMPLEMENTATION	
	4.1 SYSTEM ARCHITECTURE	
	4.2 MODULES	
5	SYSTEM DESIGN	
	5.1 DATA FLOW DIAGRAM	
	5.1.1 LEVEL 0	
	5.1.2 LEVEL 1	
	5.1.3 LEVEL 2	
	5.1.4 LEVEL 3	
6	CONCLUSION	
	REFERENCES	
	APPENDIX A: SNAP SHOTS	
	APPENDIX B: SOURCE CODE	

CHAPTER 1

INTRODUCTION

Data Science is one of the most motivating and vital area of research with the aim of extracting information from tremendous amount of accumulated data sets. Going on is an era of simplifying almost all complicated works using computers. Manual processing makes the process slow and other problems such as inconsistency and ambiguity on operations.

The analysis of risk in bank loans need understanding what is the meaning of risk. Risk in bank loans involve: credit risk, the risk that the loan won't be return back on time or at all; liquidity risk, the risk that too many deposits will be withdrawn too quickly, leaving the bank short on immediate cash; and interest rate risk, the risk that the interest rates priced on bank loans will be too low to earn the bank adequate money. Data Science techniques aid to distinguish between borrowers who pay back loans at the appointed time from those who don't.

It also helps to expect when the borrower is at default, whether providing loan to a particular customer will result in bad loans. All processes related to banking sector could be analyzed using data mining to detect the customer's behavior. It also helps to analyze whether the customer will make prompt or delay payment if the credit cards are sold to them. Many analysis have been conducted based on data science in the field of financial and banking sector.

The goal of this project, is to focus on the loan evaluation process. The Decision Tree algorithm have been used to create the loan approval classification model using factors such as Loan_ID, Gender, Married, Education, Self_Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Property_Area etc.,. This model helps the bank manager to make the decision of whether the loan have to be approved or not to a particular customer. So that, Bank manager can classify the customers easily and accurately using this classification model.

CHAPTER 2

SYSTEM ANALYSIS

2.1 EXISTING SYSTEM

Existing concept deals with Qualitative observations and simple statistical analysis. The qualitative observations deals with the data that can be observed through human senses. They do not involve measurements or number. The simple statistical analysis includes mean, standard deviation, median, finding the size of data, variance etc., the results produced by this techniques are not precise. Using some programming languages and data mining tools, bank dataset is classified. There is an increase in need to manage and understand data.

2.2 PROPOSED SYSTEM

Loan application dataset is collected and analyzed to classify whether the bank manager have to grant the loan to a particular customer or not. For classification, Data Science techniques of Decision Tree algorithm have been applied on the dataset. Data Science is an interdisciplinary field that incorporates computer science, mathematics, statistics and domain knowledge. Decision Tree classification model is applied to the loan application dataset, the efficient classification accuracy is achieved by delivering the result of whether to approve the loan request of the bank customers or not.

CHAPTER 3

SYSTEM SPECIFICATION

3.1 HARDWARE SPECIFICATION

Processor	:	Processor Intel CORE i3 and above
Internet Connection	:	Existing telephone lines, Data card, Fiber net
RAM	:	4 GB

3.2 SOFTWARE SPECIFICATION

Operating System	:	Windows, Mac, Linux
Language	:	R Programming – R-4.1.1
GUI	:	RStudio

R is a Programming language and environment for Data Manipulation, statistical computing and graphics. Developed at Bell Laboratories by John Chambers and colleagues. Similar to the S language and available across all platforms -Linux, Mac, Windows. It provides variety of Graphical and Statistical techniques – Regression, Clustering, Classification, Association Rule Mining etc.

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and Mac OS. The source code for the R software environment is written primarily in C, FORTRAN, and R.

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- An effective data handling and storage facility,
- A suite of operators for calculations on arrays, in particular matrices,
- A large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either on-screen or on hardcopy and
- a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

RStudio is an Integrated Development Environment (IDE) for R. RStudio makes R easier to use. It includes a code editor, compiler or interpreter, debugger and visualization tools as well as tools for plotting, history, syntax-highlighting editor that supports direct code execution and workspace management. RStudio was founded by JJ Allaire, creator of the programming language ColdFusion.

RStudio is available in open source and commercial editions available in two editions: RStudio Desktop, where the program is run locally as a regular desktop application (Windows, macOS, and Linux) and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote Linux server. RStudio is written in the C++ programming language. RStudio helps to keep R more organized and it adds more functionality to it.

CHAPTER 4

SYSTEM IMPLEMENTATION

4.1 SYSTEM ARCHITECTURE

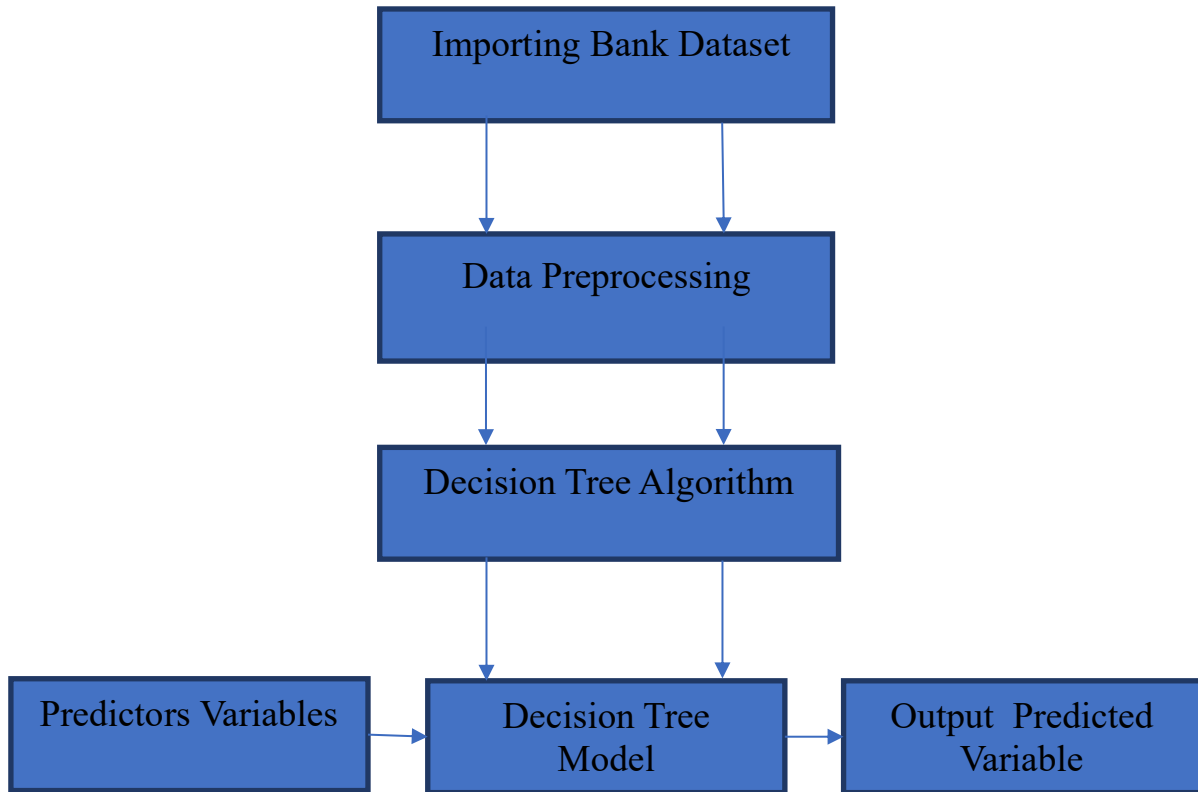


Fig.4.1 Architecture Diagram

The fig 4.1 shows the architectural diagram of this project defines the flow of data for classifying the customers for approving the loan. The first step is data importing, Data have to be loaded in to the R environment for analysis. The second step is, preprocessing the collected raw data into understandable format. Standardization and Normalization is the technique which is used to transform the various format of data into the common format and min-max technique is used for normalization of data values.

It is not necessary to hold all the attributes for doing the analysis, we can hold only the attributes which is affecting the analysis. The missing values

problem have to be solved by simple statistical techniques. Decision Tree algorithm is applied on the preprocessed data to create a predictive Decision Tree model. The algorithm is tested with some predictor variables, the Decision Tree Model will generate the output of whether to approve the loan request of the bank customers or not.

4.2 MODULES

There are four Modules

- Importing Bank Dataset
- Data Preprocessing
- Model Generation Using Decision Tree Algorithm
- Prediction using Decision Tree Model

4.2.1 IMPORTING BANK DATASET

Data is available in any file format like .txt, .csv, .xlsx, .spss etc. Data have to be loaded in to R environment for analysis. Once data have been extracted from the file it should be stored in a data frame. Packages necessary for classification algorithm – Decision Tree have to be installed into the R environment. For Decision Tree, party package have to be installed and loaded in to R environment.

4.2.2 DATA PREPROCESSING

Data preprocessing is the data mining technique that involves transforming raw data into understandable format. The raw data is highly susceptible to noise, missing values, and inconsistency.

Real world data is often incomplete, inconsistent, and is likely to contain errors. The missing Values problem have to be solved by simple statistical techniques.

Data preprocessing is the proven method for resolving such issues. In order to improve the quality of the data consequently, the mining results of raw data is preprocessed so the efficiency process improved. It is not necessary to hold all the attributes for doing the analysis, we can hold only the attributes which is affecting the analysis.

Data standardization is the process by which similar data is collected in various formats is transformed to a common format that enhances the comparison process, allows for collaborative research and large scale analytics. Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Formula used for Normalization:

$$\frac{X - X_{min}}{X_{max} - X_{min}}$$

4.2.3 MODEL GENERATION USING DECISION TREE ALGORITHM

Classification is a technique where we categorize data into a given number of classes. Classification is to determine the extent to which a thing will or will not be a part of a category or type. Classification can be performed on structured or unstructured data. The main goal of a classification problem is to identify the category/class to which a new data will fall under.

Decision trees are powerful and popular tools for classification and prediction. J. Ross Quinlan originally developed ID3 at the University of Sydney. ID3 uses Entropy and Information Gain to construct a decision tree. Decision tree is a classifier in the form of a tree structure. Root node represents entire

population or sample and this further gets divided into two or more homogeneous sets.

Splitting is a process of dividing a node into two or more sub-nodes. When a sub-node splits into further sub-nodes, then it is called decision node. Nodes do not split is called Leaf or Terminal node. When we remove sub-nodes of a decision node, this process is called pruning. It is the opposite process of splitting. A sub section of entire tree is called branch or sub-tree. A node, which is divided into sub-nodes is called parent node of sub-nodes whereas sub-nodes are the child of parent node. The entropy is a measure of the uncertainty or impurity associated with a random variable. Values range from 0 – 1 to represent the entropy of information

$$Entropy(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

Information gain is a measure of the effectiveness of an attribute in classifying the training data. Information gain is used as an attribute selection measure. Pick the attribute that has the highest Information Gain

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Algorithmic steps of Decision Tree algorithm

Step 1: Calculate entropy of the target.

Step 2: The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain

Step 3: Choose attribute with the largest information gain as the decision node, divide the dataset by its branches and repeat the same process on every branch.

Outlook is having max gain, so tree have to start with outlook.

Step4a: A branch with entropy of 0 is a leaf node.

Step4b: A branch with entropy more than 0 needs splitting

The Previous steps have to be repeated till all the data has been classified.

Decision tree algorithm is faster than other traditional statistical techniques. It is very flexible, easy to implement and debug. It handles both continuous and discrete data. It is used for both binary and multiclass classification. Once the preprocessed data is applied to the Decision tree algorithm, Decision tree Model have been created, which is used to classify the customers for loan approval. Decision tree model output can be visualized by classification rules and binary classification tree.

4.2.4 PREDICTION USING DECISION TREE MODEL

By using test data, model can be evaluated. The algorithm is used for prediction by passing predictors variable, the Decision Tree Model will classify the customers for loan approval.

CHAPTER 5

SYSTEM DESIGN

5.1 DATA FLOW DIAGRAM

Data flow diagram (DFD) is a graphical representation of the "flow" of data through an information system, modeling its process aspects. It differs from the flowchart as it shows the data flow instead of the control flow of the program. A

data flow diagram can also be used for the visualization of data processing. The DFD is designed to show how a system is divided into smaller portions and to highlight the flow of data between those parts.

5.1.1 LEVEL 0

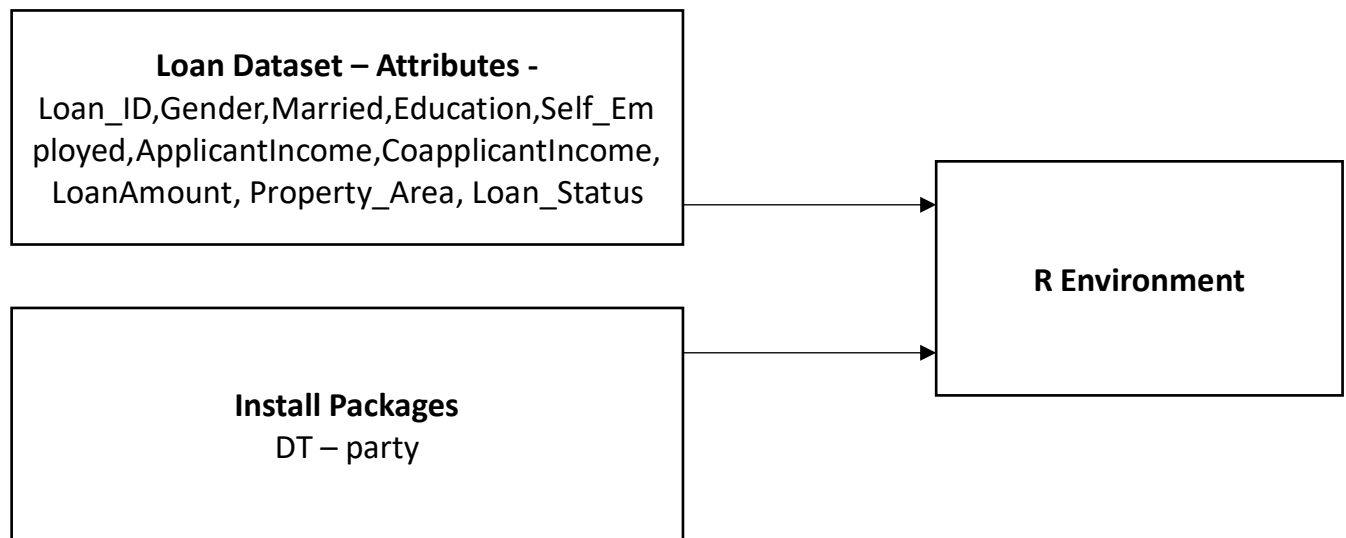


Fig 5.1.1 LEVEL 0

In the fig 5.1.1, Data is imported in to R environment for analysis. The input dataset is Bank dataset, is in txt format have to be loaded into the program to start the analysis. Package necessary for Decision Tree algorithm have to be installed and loaded in to the program. Decision Tree - party.

5.1.2 LEVEL 1

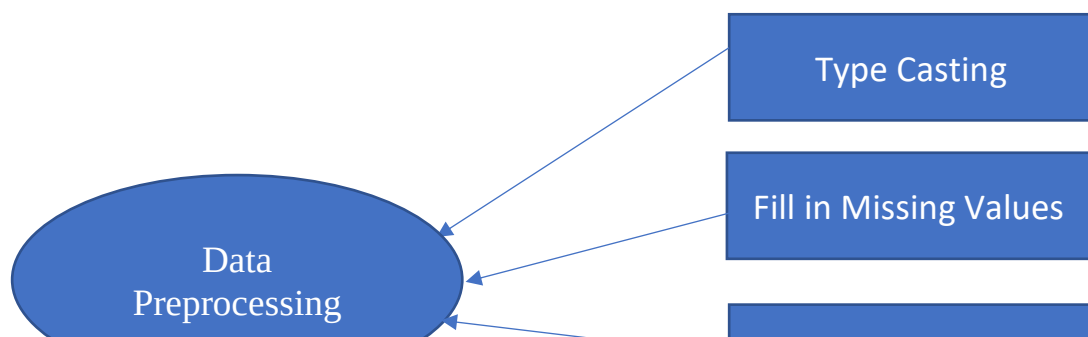


Fig 5.1.2 LEVEL 1

The fig 5.1.2 shows the data preprocessing, which includes attribute selection, standardization and normalization functions. In standardization, raw data is transformed into common, understandable format. In attribute selection, hold only the attributes which is affecting the analysis and it is not necessary to hold all the attributes for doing the analysis. In Normalization, mean of the attribute will be 0 and standard deviation will be 1.

5.1.3 LEVEL 2

In fig 5.1.3, preprocessed data is given as an input to Decision Tree algorithm and Decision Tree model has been created, which is used for classification and prediction.

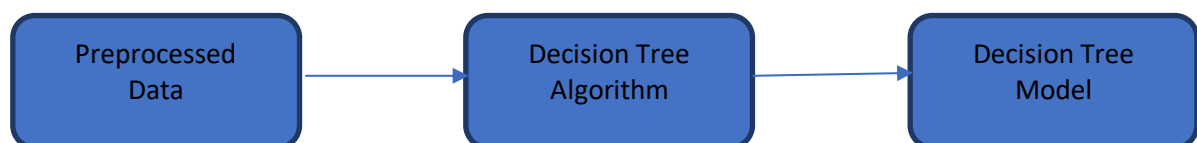


Fig 5.1.3 LEVEL 2

5.1.4 LEVEL 3

In fig 5.1.4, Predictors variables is given as an input and the Decision Tree Model will generate the output of whether to approve the loan request of the bank customers or not. Thus prediction process is implemented successfully. Using this model classification is also achieved by passing more number of customer datasets, it will classify the employee as two groups. One group is the set of customers who will repay the loan and the other group is the set of customers who will not repay the amount.

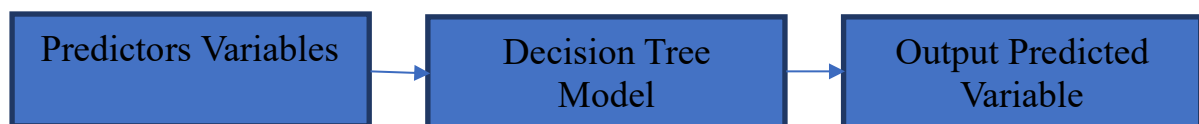


Fig 5.1.4 LEVEL 3

CHAPTER 6

CONCLUSION

The goal of this project is to predict whether the customers are going to pay back their loans or not. Decision Tree algorithm is used to build the classification and predictive model. This model can be used to predict and classify the applications of loans that introduced by the customers to good or bad, by investigate customer behaviour and previous pay back credit. Using this model bank manager can easily decide, whether to approve the loan request of the bank customers or not. The performance of this algorithm is much better when

compared to other statistical techniques. The future enhancement might include, some more data science algorithms and deep learning algorithms.

REFERENCES

1. Khaldoon Alshouiliy, Ali AlGhamdi, "AzureML Based Analysis and Prediction Loan Borrowers Creditworthy", IEEE International Conference on Information and Computer Technologies, 2020.
2. Trilok Nath Pandey, Suman Kumar Mohapatra, "Credit Risk Analysis using Machine Learning Classifiers ", IEEE International Conference on Energy, Communication, Data Analytics and Soft Computing, 2017.
3. Mohammad Ahmad Sheikh; Amit Kumar Goel, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm", IEEE International Conference on Electronics and Sustainable Communication Systems, 2020.

4. G. Arutjothi; C. Senthamarai, “Prediction of loan status in commercial bank using machine learning classifier”, IEEE International Conference on Intelligent Sustainable Systems, 2017.

5. Hua Li; Yumeng Cao, Siwen Li, “XGBoost Model and Its Application to Personal Credit Evaluation”, IEEE Intelligent Systems (Volume: 35, Issue: 3, 2020).

APPENDIX A

SNAP SHOTS

SAMPLE DATASETS

Loan_ID,Gender,Married,Dependents,Education,Self_Employed,ApplicantIncome,CoapplicantIncome,LoanAmount,Loan_Amount_Term,Credit_History,Property_Area,Loan_Status

LP001003,Male,Yes,1,Graduate,No,4583,1508,128,360,1,Rural,N

LP001005,Male,Yes,0,Graduate,Yes,3000,0,66,360,1,Urban,Y

LP001006,Male,Yes,0,Not Graduate,No,2583,2358,120,360,1,Urban,Y

LP001008,Male,No,0,Graduate,No,6000,0,141,360,1,Urban,Y

LP001011,Male,Yes,2,Graduate,Yes,5417,4196,267,360,1,Urban,Y

LP001013,Male,Yes,0,Not Graduate,No,2333,1516,95,360,1,Urban,Y

LP001014,Male,Yes,3+,Graduate,No,3036,2504,158,360,0,Semiurban,N

LP001018,Male,Yes,2,Graduate,No,4006,1526,168,360,1,Urban,Y

LP001020,Male,Yes,1,Graduate,No,12841,10968,349,360,1,Semiurban,N

LP001024,Male,Yes,2,Graduate,No,3200,700,70,360,1,Urban,Y

LP001027,Male,Yes,2,Graduate,,2500,1840,109,360,1,Urban,Y

LP001028,Male,Yes,2,Graduate,No,3073,8106,200,360,1,Urban,Y

LP001029,Male,No,0,Graduate,No,1853,2840,114,360,1,Rural,N

DATA IMPORTING

```
1 library(party)
2
3
4 Loan_data<- read.table("C:\\Program Files\\R\\R-3.3.2\\datasets\\BL.txt", header = TRUE, sep = ",")
5 print(Loan_data)
6 head(Loan_data)
7 dim(Loan_data)
8 str(Loan_data)
9 length(Loan_data)
10 Loan_data[3,]
11 Loan_data[11:20,]
12
13 Loan_data[,3]
14 print(Loan_data$Gender)
15
16 Loan_data$Loan_Status=as.factor(Loan_data$Loan_Status)
```

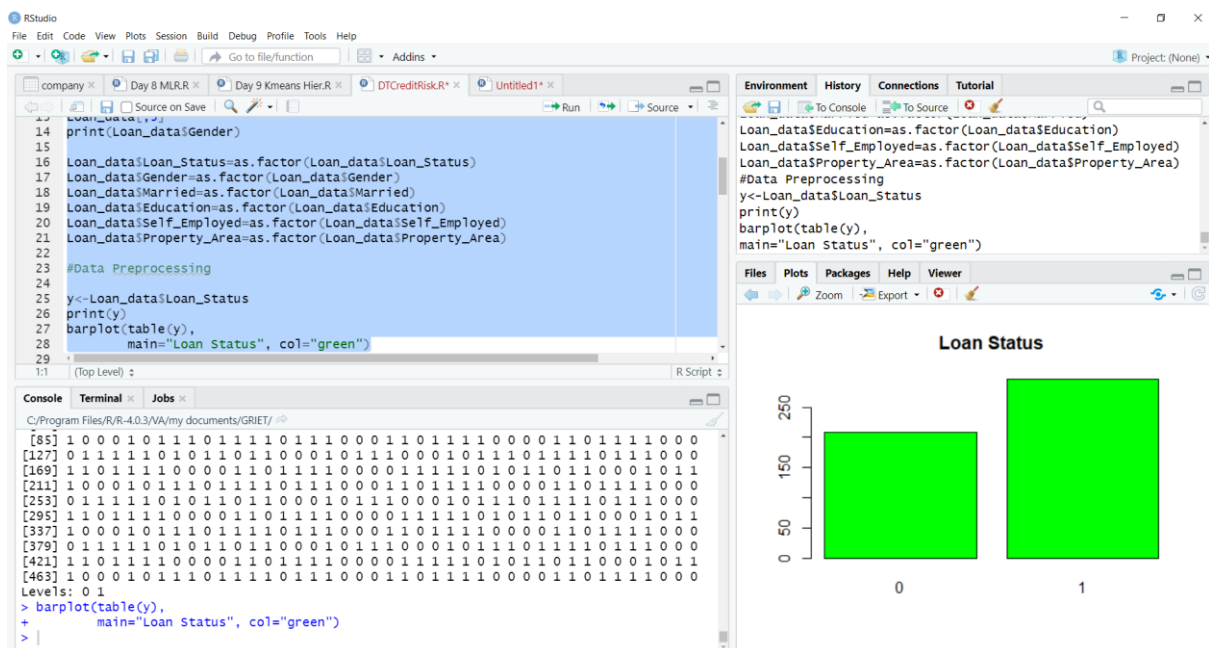
Console output:

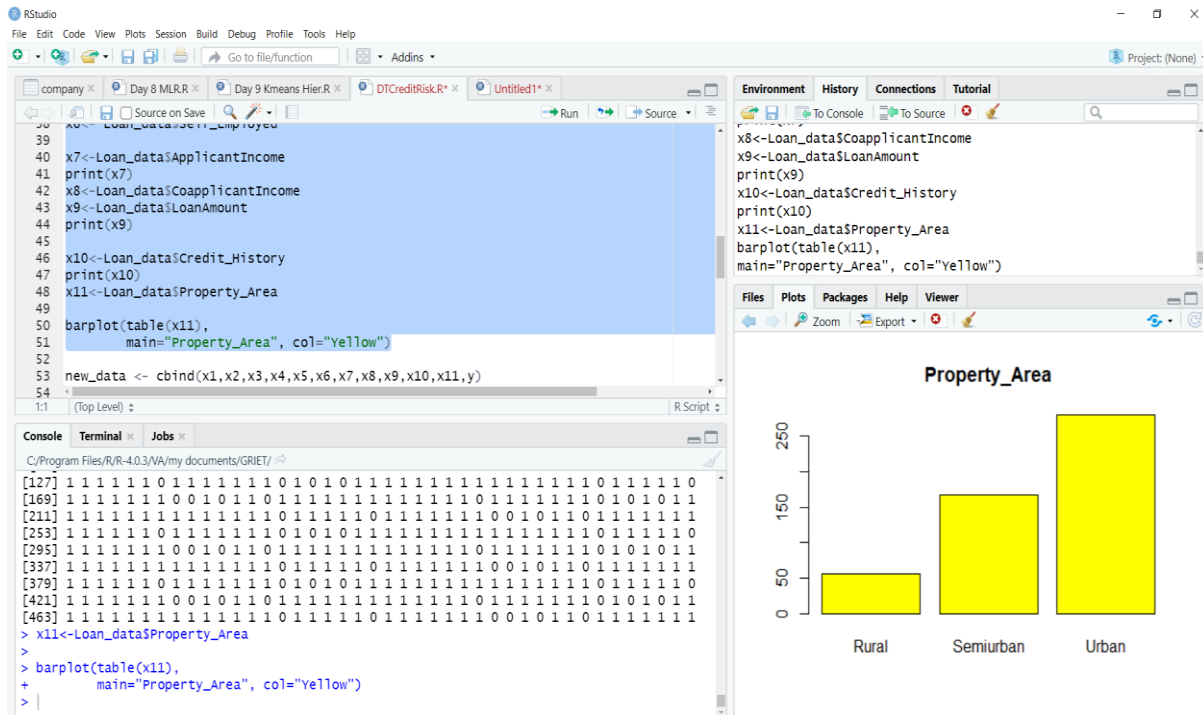
```
> Loan_data<- read.table("C:\\Program Files\\R\\R-3.3.2\\datasets\\BL.txt", header = TRUE, sep = ",")
> print(Loan_data)
  Loan_ID Gender Married Dependents Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount
1 LP001003 Male    Yes         1 Graduate      No           4583             1508           128
2 LP001005 Male    Yes         0 Graduate      Yes           3000              0            66
3 LP001006 Male    Yes         0 Not Graduate No           2583            2358           120
4 LP001008 Male    No          0 Graduate      No           6000              0            141
5 LP001011 Male    Yes         2 Graduate      Yes           5417            4196           267
6 LP001013 Male    Yes         0 Not Graduate No           2333            1516            95
7 LP001014 Male    Yes         5 Graduate      No           3036            2504           158
8 LP001018 Male    Yes         2 Graduate      No           4006            1526           168
9 LP001020 Male    Yes         1 Graduate      No          12841           10968           349
10 LP001024 Male    Yes         2 Graduate      No           3200              700            70
11 LP001028 Male    Yes         2 Graduate      No           3073            8106           200
```

```
library(party)
Loan_data<- read.table("C:\\Program Files\\R\\R-3.3.2\\datasets\\BL.txt", header = TRUE, sep = ",")
print(Loan_data)
head(Loan_data)
dim(Loan_data)
str(Loan_data)
length(Loan_data)
Loan_data[3,]
Loan_data[11:20,]
Loan_data[,3]
print(Loan_data$Gender)
Loan_data$Loan_Status=as.factor(Loan_data$Loan_Status)
```

```
> Loan_data[,3]
[1] "Yes" "Yes" "Yes" "No" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "No" "Yes" "No" "No" "Yes" "Yes" "Yes" "Yes"
[20] "Yes" "Yes" "Yes" "No" "No" "No" "Yes" "No" "Yes" "No" "No" "No" "Yes" "Yes" "Yes" "No" "Yes"
[39] "No" "No" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "No" "Yes" "No" "Yes" "No" "Yes"
[58] "Yes" "No" "Yes" "No" "No" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "No" "Yes" "No" "Yes"
[77] "No" "No" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "No" "No" "No" "Yes" "No" "No" "No" "No" "No"
[96] "Yes" "Yes" "Yes" "Yes" "No" "Yes" "No" "No" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "No"
[115] "Yes" "No" "Yes" "Yes" "Yes" "Yes" "No" "No" "Yes" "No" "No" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes"
[134] "Yes" "Yes" "Yes" "Yes" "No" "Yes" "No" "No" "Yes" "Yes" "Yes" "Yes" "No" "No" "No" "No" "Yes"
[153] "No" "No" "No" "No" "No" "No" "Yes" "Yes" "No" "Yes" "No" "No" "No" "No" "No" "No" "Yes"
[172] "Yes" "Yes" "Yes" "Yes" "No" "Yes" "No" "Yes" "Yes" "No" "Yes" "No" "Yes" "No" "Yes" "Yes" "Yes"
[191] "Yes" "No" "No" "Yes" "Yes" "No" "Yes" "Yes" "Yes" "No" "Yes" "No" "Yes" "No" "Yes" "Yes" "Yes"
[210] "Yes" "Yes" "No" "No" "No" "Yes" "No" "Yes" "No" "No" "Yes" "No" "Yes" "No" "Yes" "No" "Yes"
[229] "No" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "No" "Yes" "No" "Yes" "Yes" "Yes" "Yes"
[248] "No" "Yes" "No" "No" "No" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes"
```

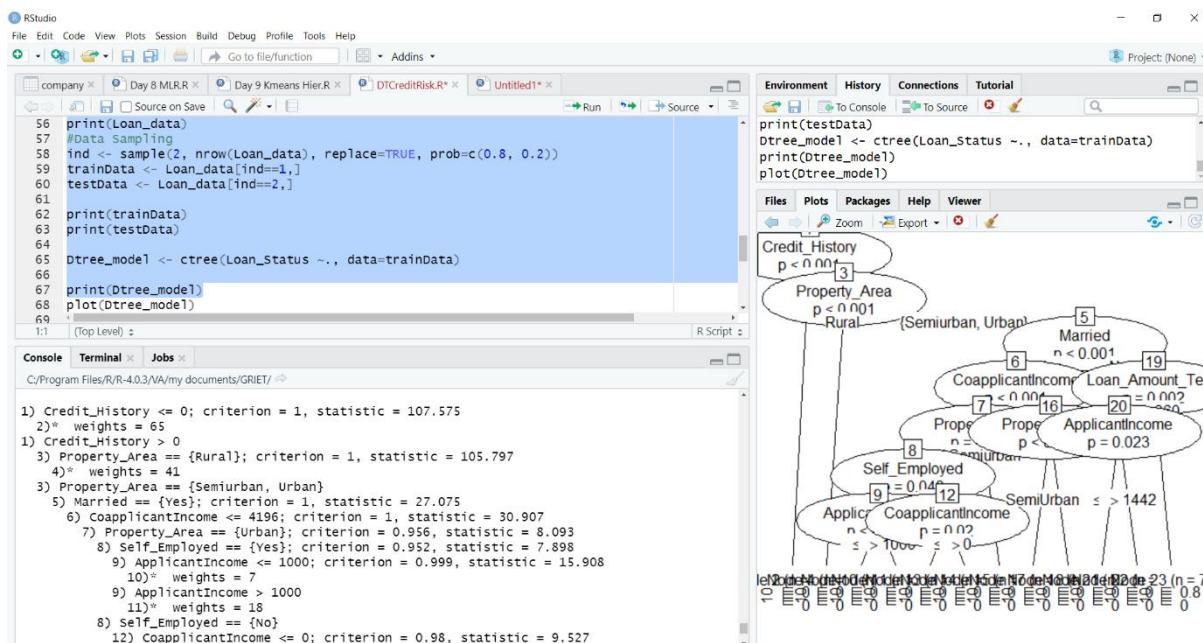
DATA PREPROCESSING



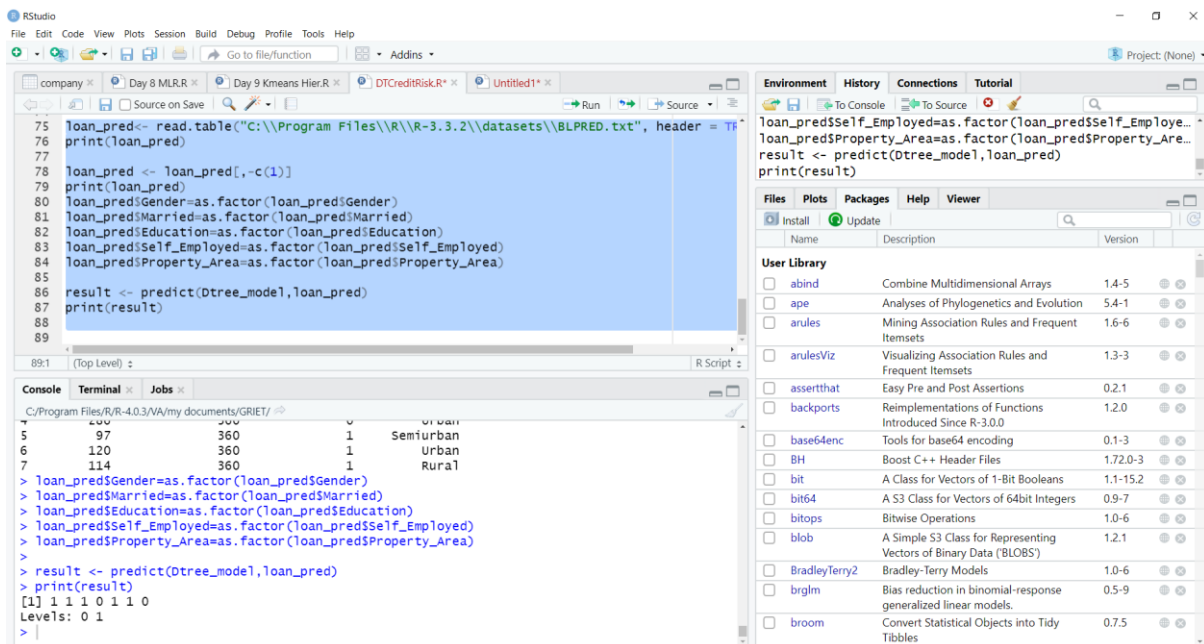


DECISION TREE ALGORITHM

Decision Tree Classification Model



DT Classification Rule Graph



APPENDIX B

SAMPLE CODING

#Loading Package

```
library(party)
```

#Data Importing

```
Loan_data<- read.table("C:\\Program Files\\R\\R-3.3.2\\datasets\\BL.txt", header =
TRUE, sep = ",")
```

```
print(Loan_data)
```

```
head(Loan_data)
```

```
dim(Loan_data)
```

```
str(Loan_data)
```

```
length(Loan_data)
```

```
Loan_data[3,]
```

```
Loan_data[11:20,]
```

```
Loan_data[,3]
```

```
print(Loan_data$Gender)
```

#Data Preprocessing

```
Loan_data$Loan_Status=as.factor(Loan_data$Loan_Status)
```

```
Loan_data$Gender=as.factor(Loan_data$Gender)
Loan_data$Married=as.factor(Loan_data$Married)
Loan_data$Education=as.factor(Loan_data$Education)
Loan_data$Self_Employed=as.factor(Loan_data$Self_Employed)
Loan_data$Property_Area=as.factor(Loan_data$Property_Area)
```

```
y<-Loan_data$Loan_Status
print(y)
barplot(table(y),
        main="Loan Status", col="green")
```

```
x1<- Loan_data$Loan_ID
x2<- Loan_data$Gender
barplot(table(x2))
x3<- Loan_data$Married
barplot(table(x3))
x4<- Loan_data$Dependents
hist(x4)
x5<- Loan_data$Education
x6<- Loan_data$Self_Employed
x7<-Loan_data$ApplicantIncome
```

```
hist(x7, main = "Applicant Income",xlab="Income", ylab="No.of Applicants",col =
    "green")
x8<-Loan_data$CoapplicantIncome
x9<-Loan_data$LoanAmount
print(x9)
x10<-Loan_data$Credit_History
print(x10)
x11<-Loan_data$Property_Area
barplot(table(x11),
        main="Property_Area", col="Yellow")
```

```
new_data <- cbind(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,y)
```

```
#Attribute Selection
```

```
Loan_data <- Loan_data[,-c(1)]
```

```
print(Loan_data)
```

```
#Data Sampling
```

```
ind <- sample(2, nrow(Loan_data), replace=TRUE, prob=c(0.8, 0.2))
```

```
trainData <- Loan_data[ind==1,]
```

```
testData <- Loan_data[ind==2,]
```

```
print(trainData)
```

```
print(testData)
```

```
#Decision Tree Model Generation
```

```
Dtree_model <- ctree(Loan_Status ~., data=trainData)
```

```
print(Dtree_model)
```

```
plot(Dtree_model)
```

```
# Testing Classification Model
```

```
testPred <- predict(Dtree_model, newdata = testData)
```

```
print(testPred)
```

```
#Target for the given predictors
```

```
loan_pred<- read.table("C:\\Program Files\\R\\R-3.3.2\\datasets\\BLPRED.txt", header  
= TRUE, sep = ",")
```

```
print(loan_pred)
```

```
loan_pred <- loan_pred[,-c(1)]
```

```
print(loan_pred)
```

```
loan_pred$Gender=as.factor(loan_pred$Gender)
```

```
loan_pred$Married=as.factor(loan_pred$Married)
```

```
loan_pred$Education=as.factor(loan_pred$Education)
```

```
loan_pred$Self_Employed=as.factor(loan_pred$Self_Employed)
loan_pred$Property_Area=as.factor(loan_pred$Property_Area)
result <- predict(Dtree_model,loan_pred)
print(result)
```