LOAN ELIGIBILITY PREDICTION

1. INTRODUCTION:

a. Overview :

Loans are the core business of banks. The main profit comes directly from the loan's interest. The loan companies grant a loan after an intensive process of verification and validation. However, they still don't have assurance if the applicant is able to repay the loan with no difficulties. The main aim of this use-case is to build a predictive model to predict if an applicant is able to repay the lending company or not

b. Purpose:

Loans are the core business of banks. The main profit comes directly from the loan's interest. The loan companies grant a loan after an intensive process of verification and validation. However, they still don't have assurance if the applicant is able to repay the loan with no difficulties. In this project we try to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets. This is done by mining the Data of the previous records of the people to whom the loan was granted before and on the basis of these records the machine was trained using the machine learning model which give the most accurate result. The main objective of this project is to predict whether assigning the loan to particular person will be safe or not.

2.LITERATURE SURVEY:

a. Existing Problems:

Data mining is the process of analyzing data from different perspectives and extracting useful knowledge from it. Different data mining techniques include classification, clustering, association rule mining, prediction and sequential patterns, neural networks, regression etc. Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large.In classification, a training set is used to build the model as the classifier which can classify the data items into its appropriate classes. A test set is used to validate the model.

b. Proposed solution:

Logistic Regression:

Logistic Regression is one of the most popular machine learning algorithm, which is used for predicting the categorical dependent variable using a given set of dependent variable.

Logistic Regression is used when the dependent variable(target) is categorical.

For example,

- To predict whether an email is spam (1) or (0)
- Whether the tumor is malignant (1) or not (0)

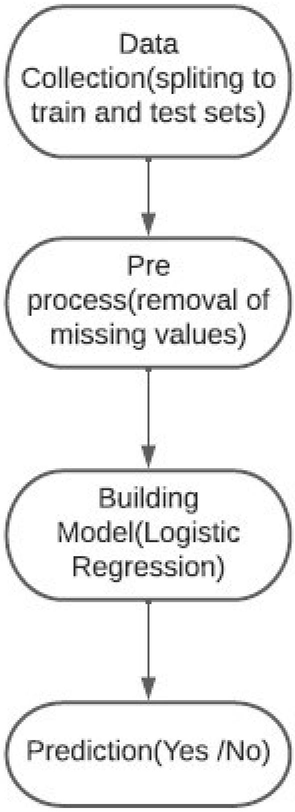
Therefore **Logistic regression is used for solving the classification problems**. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

3.Theoretical Analysis:

a. Block Diagram:

The steps involved in Building the data model is depicted below:

LOAN ELIGIBILITY PREDICTION



b. Software :

The software used for this project are :

- ➤ Java version 10
- ➤ Eclipse neon IDE.
- ➤ WEKA 3.8.5,

4.EXPERIMENTAL INVESTIGATION:

a. Data Collection:

The dataset collected for predicting loan default customers is predicted into Training set and testing set.. The data model which was created using Logistic regression is applied on the training set and based on the test result accuracy, Test set prediction is done. Attributes

LOAN ELIGIBILITY PREDICTION

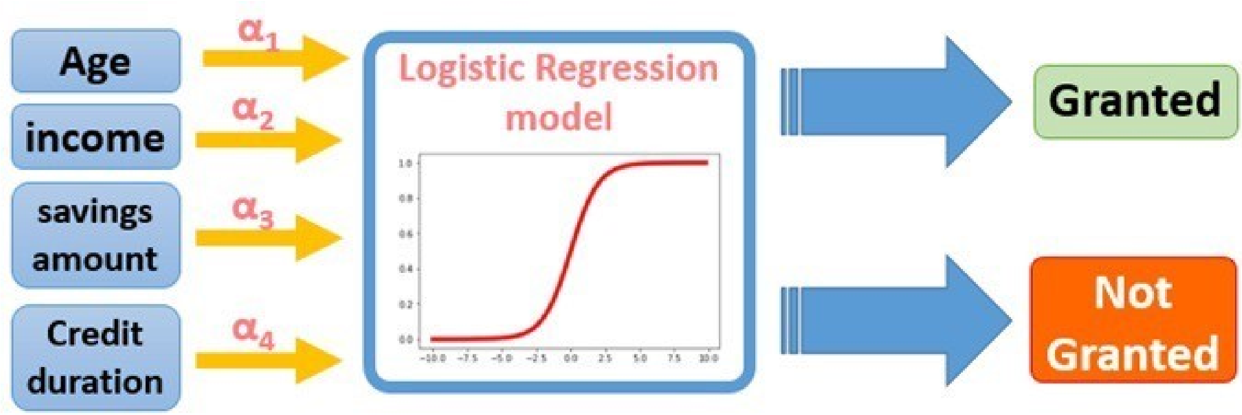
b. Pre processing:

The data which was collected might contain missing values that may lead to inconsistency. To gain better results data need to be preprocessed so as to improve the efficiency o the algorithm.

c. Buliding Model using Logistic Regression Model:

For predicting the loan defaulter's and non defaulter's problem Logistic Regression algorithm is used. The purpose of this algorithm is to find a plane that separates two types. Y variable belongs to 1 or 0.

5.FLOWCHART:



6.RESULT:

The accuracy for the built model is 81%, Precision is 59%.

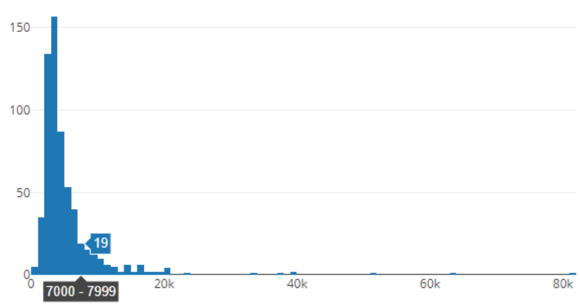
A)ECLIPSE RESULT:

LOAN ELIGIBILITY PREDICTION

```
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
614 rows X 13 cols
                                                                              train_u6lujuX_CVtuZ9i.csv
 Loan_ID  |  Gender  | Married |  Dependents  |  Education   | Self_Employed | ApplicantIncome | CoapplicantIncome | LoanAmount |
--------------------------------------------------------------------------------------------------------------------------------
 LP001002 |    Male  |      No |            0 |     Graduate |            No |            5849 |                 0 |            |
 LP001003 |    Male  |     Yes |            1 |     Graduate |            No |            4583 |              1508 |        128 |
 LP001005 |    Male  |     Yes |            0 |     Graduate |           Yes |            3000 |                 0 |         66 |
 LP001006 |    Male  |     Yes |            0 | Not Graduate |            No |            2583 |              2358 |        120 |
 LP001008 |    Male  |      No |            0 |     Graduate |            No |            6000 |                 0 |        141 |
 LP001011 |    Male  |     Yes |            2 |     Graduate |           Yes |            5417 |              4196 |        267 |
 LP001013 |    Male  |     Yes |            0 | Not Graduate |            No |            2333 |              1516 |         95 |
    Structure of train_u6lujuX_CVtuZ9i.csv
 Index  |     Column Name      |  Column Type  |
-----------------------------------------------
     0  |              Loan_ID  |       STRING  |
     1  |               Gender  |       STRING  |
     2  |              Married  |       STRING  |
     3  |           Dependents  |       STRING  |
     4  |            Education  |       STRING  |
     5  |        Self_Employed  |       STRING  |
     6  |      ApplicantIncome  |      INTEGER  |
     7  |    CoapplicantIncome  |       DOUBLE  |
     8  |           LoanAmount  |      INTEGER  |
     9  |     Loan_Amount_Term  |      INTEGER  |
    10  |       Credit_History  |      INTEGER  |
    11  |        Property_Area  |       STRING  |
    12  |          Loan_Status  |      BOOLEAN  |
                                                                                          train u6lujuX CVtuZ9i.csv
  Summary   |  Loan_ID  |  Gender  | Married |  Dependents  |  Education  | Self_Employed |  ApplicantIncome   |  CoapplicantIncome  |
---------------------------------------------------------------------------------------------------------------------------------------
     Count  |      614  |     614  |     614 |          614 |         614 |           614 |                614 |                 614 |
    Unique  |      614  |       3  |       3 |            5 |           2 |             3 |                    |                     |
       Top  | LP002888  |    Male  |     Yes |            0 |    Graduate |            No |                    |                     |
 Top Freq.  |        1  |     489  |     398 |          345 |         480 |           500 |                    |                     |
       sum  |           |          |         |              |             |               |            3317724 |     995444.91998864 |
      Mean  |           |          |         |              |             |               | 5403.4592833876195 |  1621.2457980270997 |
       Min  |           |          |         |              |             |               |                150 |                   0 |
       Max  |           |          |         |              |             |               |              81000 |               41667 |
     Range  |           |          |         |              |             |               |              80850 |               41667 |
  Variance  |           |          |         |              |             |               | 37320390.167181246 |   8562929.518387228 |
  Std. Dev  |           |          |         |              |             |               |  6109.041673387181 |  2926.2483692241894 |
     false  |           |          |         |              |             |               |                    |                     |
      true  |           |          |         |              |             |               |                    |                     |
```

Writable | Smart Insert | 24 : 51

32°C ENG 07:59 PM 07-05-2021



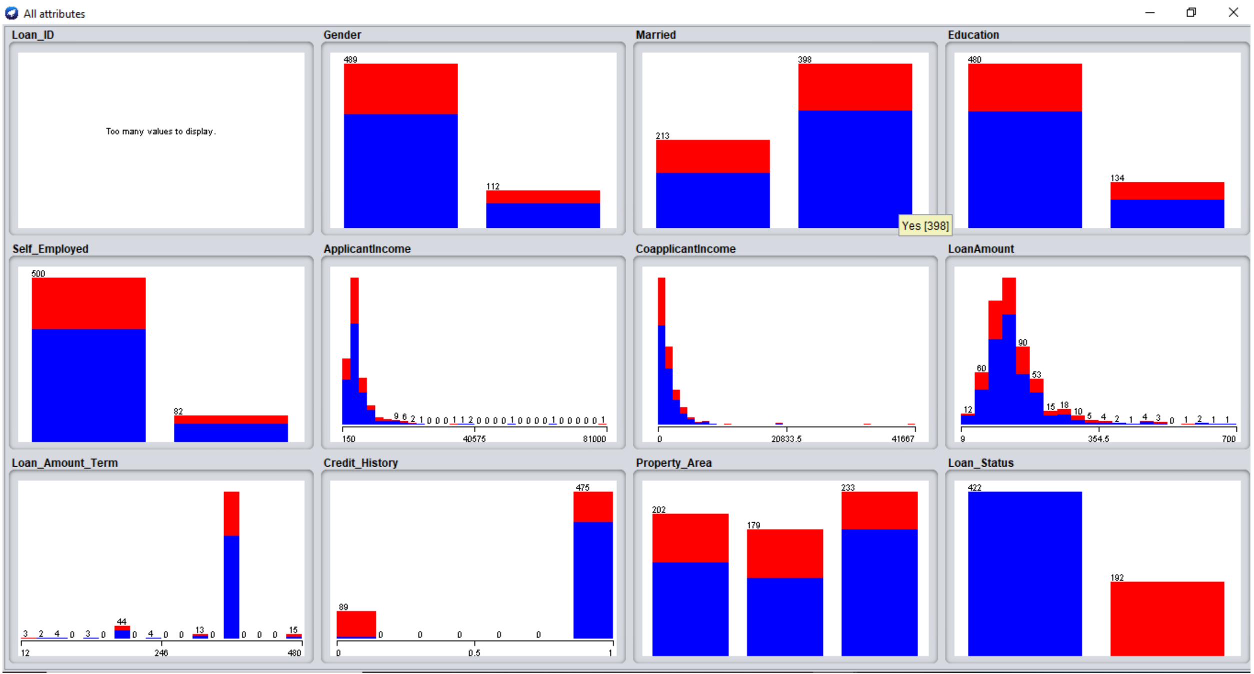
LOAN ELIGIBILITY PREDICTION

```
** Logistic Regression Evaluation with Datasets **

Correctly Classified Instances         498               81.1075 %
Incorrectly Classified Instances       116               18.8925 %
Kappa statistic                          0.4878
Mean absolute error                      0.2909
Root mean squared error                  0.3816
Relative absolute error                 67.635  %
Root relative squared error             82.3114 %
Total Number of Instances              614

 the expression for the input data as per alogorithm is Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
                             Class
Variable                         Y
==================================
Gender=Female               0.0242
Married=Yes                 0.6062
Education=Not Graduate     -0.3887
Self_Employed=Yes           -0.019
ApplicantIncome                  0
CoapplicantIncome               -0
LoanAmount                 -0.0018
Loan_Amount_Term           -0.0011
Credit_History              3.8687
Property_Area=Urban        -0.1756
Property_Area=Rural        -0.3695
Property_Area=Semiurban     0.4887
Intercept                  -2.0897
```

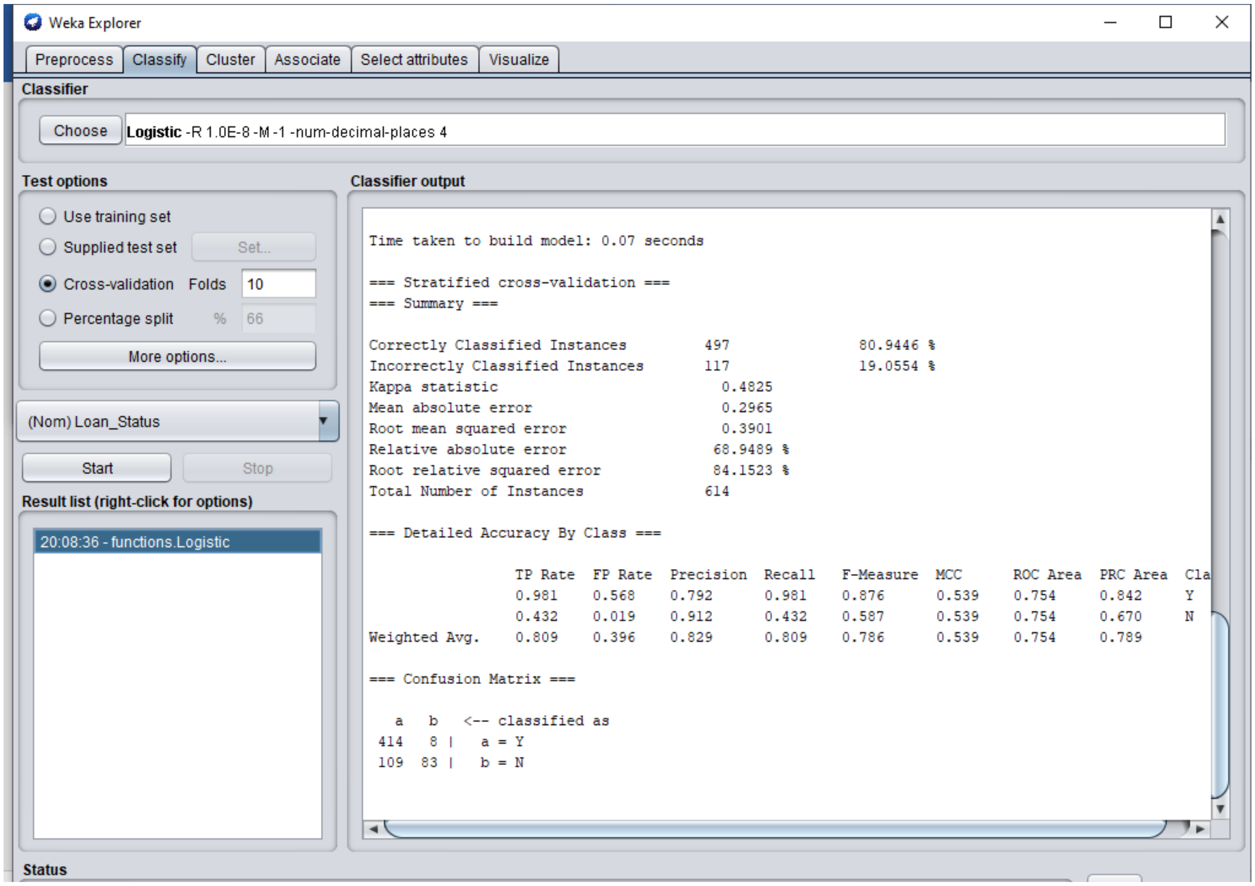
```
                             Class
Variable                         Y
==================================
Gender=Female               1.0245
Married=Yes                 1.8335
Education=Not Graduate       0.678
Self_Employed=Yes           0.9811
ApplicantIncome                  1
CoapplicantIncome                1
LoanAmount                  0.9982
Loan_Amount_Term            0.9989
Credit_History               47.88
Property_Area=Urban          0.839
Property_Area=Rural         0.6911
Property_Area=Semiurban     1.6302

Confusion Matrix...
[414.0, 8.0]
[108.0, 84.0]
--------------------------------------
Area under the curve
0.7876801935229067
----------------------------------------
[Correct, Incorrect, Kappa, Total cost, Average cost, KB relative, KB information, Correlation, Complexity 0, Complexity scheme, Complexity im
Recall--
0.44
precison
0.91
precison
0.59
Accuracy
0.81
```

LOAN ELIGIBILITY PREDICTION

B)WEKA GUI Result:



LOAN ELIGIBILITY PREDICTION



7.ADVANTAGES AND DISADVANTAGES:

7.1 Advantages:

- Compared to othe algorithm,Logistic regression will provide probablility prediction along with the classification result.
- Logistic regression can be used for large set of data.
- One of the great advantages of Logistic Regression is that when you have a complicated linear problem and not a whole lot of data it's still able to produce pretty useful predictions.

7.2 Disadvantages:

- Data preparation can be tedious in Logistic Regression as both scaling and normalization are important requirements of Logistic Regression.

- Logistic Regression is not immune to missing data unlike some other machine learning models such as decision trees and random forests which are based on trees.

8.APPLICATOIN:

It can be used for banking sectors for predicting the eligibility of loan for the customers and predicting the customers's loan status whether he will be able to pay the loan or notby using the previous records.

9.CONCLUSION:

The analytical process started from data cleaning and processing, Missing value imputation with micepackage, thenexploratory analysis and finally model building and evaluation. The best accuracy on public test set is 0.81. Most of the Time, Applicants with high income sanctioning low amount is to more likely get approved which make sense, more likely to pay back their loans.

10.BIBILOGRAPHY:

http://www.ijetjournal.org
https://www.javatpoint.com/logistic-regression-in-machine-learning
https://holypython.com/log-reg/logistic-regression-pros-cons/

11.APPENDIX:

Source Code:

A) Data Analysis Code:

```
package org.ml;
import java.io.IOException;
```

LOAN ELIGIBILITY PREDICTION

```
import tech.tablesaw.api.Table;
import tech.tablesaw.plotly.Plot;
import tech.tablesaw.plotly.components.Figure;
import tech.tablesaw.plotly.components.Layout;
import tech.tablesaw.plotly.traces.HistogramTrace;

public class DataAnalysis {

    public static void main(String[] args) {
        try {
            Table bank_data =
Table.read().csv("M:\\Oracle\\org.ml\\src\\main\\java\\org\\ml\\train_u6lujuX_CVtuZ9i.csv");
            System.out.println(bank_data.shape());
            System.out.println(bank_data.first(7));
            System.out.println(bank_data.structure());
            System.out.println(bank_data.summary());



            Layout layout1 = Layout.builder().title("Distribution of age").build();
            HistogramTrace trace1 = HistogramTrace.builder(bank_data.nCol("ApplicantIncome")).build();
            Plot.show(new Figure(layout1,trace1));

        } catch (IOException e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
        }

    }

}
```

B) Logistic Regression Model:

```
package org.ml;

import java.util.Arrays;

import weka.classifiers.Classifier;
import weka.classifiers.Evaluation;
importweka.classifiers.functions.LinearRegression;
importweka.core.Instance;
import weka.core.Instances;
import weka.core.converters.ConverterUtils.DataSource;

public class Regression {
```

LOAN ELIGIBILITY PREDICTION

```
public static void main(String[] args) throws Exception {
  DataSource source = new DataSource("M:\\Oracle\\org.ml\\src\\main\\java\\org\\ml\\train.arff");
Instances dataset = source.getDataSet();
dataset.setClassIndex(dataset.numAttributes()-1);

Classifier classifier = new weka.classifiers.functions.Logistic();

DataSource source1 = new DataSource("M:\\Oracle\\org.ml\\src\\main\\java\\org\\ml\\test.arff");
Instances dataset1 = source.getDataSet();
dataset1.setClassIndex(dataset1.numAttributes()-1);




classifier.buildClassifier(dataset);
//System.out.println(classifier);

Evaluation eval = new Evaluation(dataset);
    eval.evaluateModel(classifier, dataset1);
    /** Print the algorithm summary */
    System.out.println("** Logistic Regression Evaluation with Datasets **");
    System.out.println(eval.toSummaryString());
    System.out.print(" the expression for the input data as per alogorithm is ");
    System.out.println(classifier);


    double confusion[][] = eval.confusionMatrix();
    System.out.println("Confusion Matrix...");
    for (double[] row : confusion)
        System.out.println( Arrays.toString(row));
 System.out.println("-------------------------------------");

 System.out.println("Area under the curve");
 System.out.println(eval.areaUnderROC(0));
 System.out.println("---------------------------------------");


 System.out.println(eval.getAllEvaluationMetricNames());;


 System.out.println("Recall--");
```

LOAN ELIGIBILITY PREDICTION

```
    System.out.println(Math.round(eval.recall(1)*100.0)/100.0);

    System.out.println("precison");
    System.out.println(Math.round(eval.precision(1)*100.0)/100.0);

    System.out.println("precison");
    System.out.println(Math.round(eval.fMeasure(1)*100.0)/100.0);

    System.out.println("Accuracy");
    double acc = eval.correct()/(eval.correct()+eval.incorrect());
    System.out.println(Math.round(acc*100.0)/100.0);








   }

}
```