LOAN ELIGIBILITY PREDICTION

PROBLEM STATEMENT

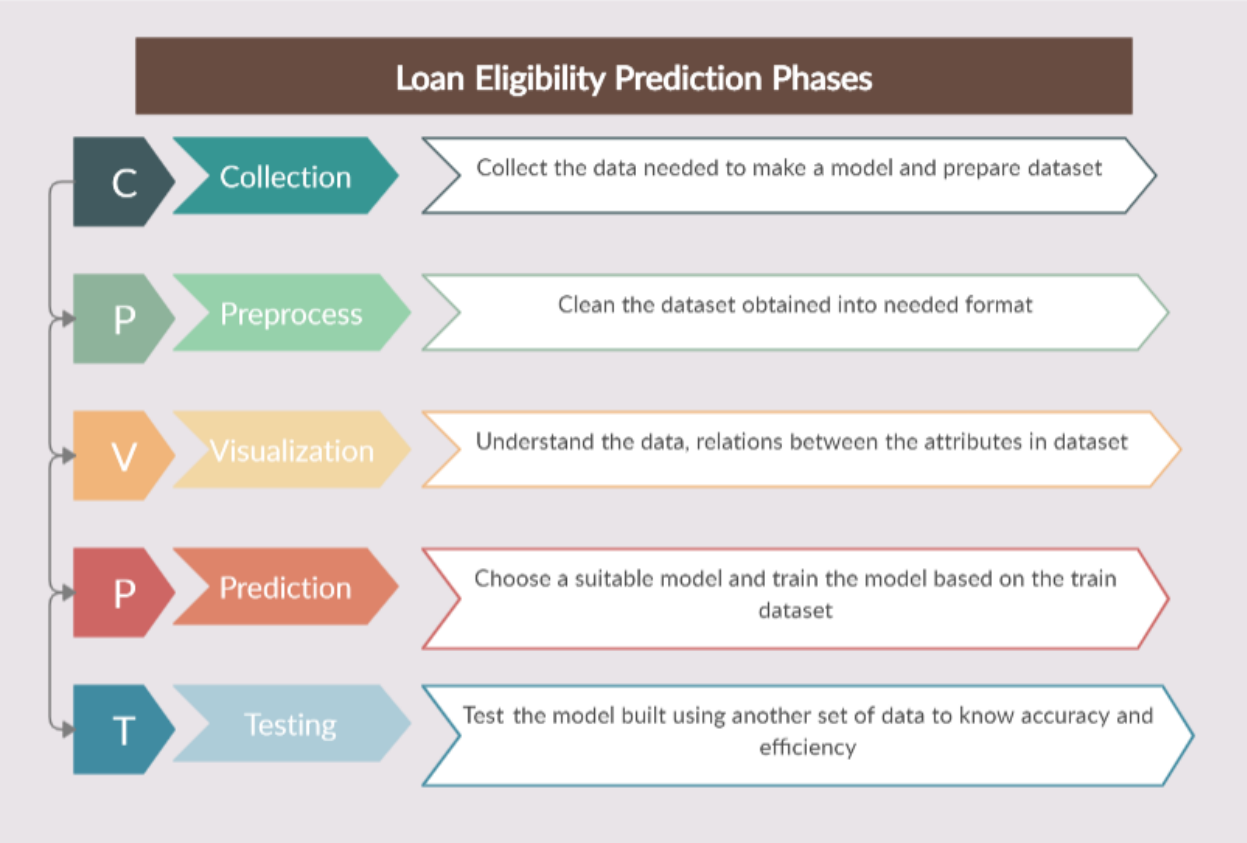
Loans are the core business of banks. The main profit comes directly from the loan's interest. The loan companies grant a loan after an intensive process of verification and validation. However, they still don't have assurance if the applicant is able to repay the loan with no difficulties.

SOLUTION PROPOSED

Build a predictive model to predict using machine learning if an applicant can repay the lending company or not.

SOLUTION USING MACHINE LEARNING

FLOW OF PROJECT



STEP 1: Analysing data

After collecting dataset,analyse the dataset. Some of the details to be noticed in the dataset:

- Shape of dataset(number of columns and rows present in dataset)
- Understand columns present in the dataset.
- Analyse the structure of the dataset.
- Check if missing values are present in the dataset.

Structure of train_data.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

Summary of data

train_data.csv

Summary	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	.	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
Count	614	614	614	614	614	614	614	614	614	4	614	614	614	614	
Unique	614	3	3	5	2	3								3	
Top	LP002888	Male	Yes	0	Graduate	No								Semiurban	
Top Freq.	1	489	398	345	480	500								233	
sum							3317724	995444.91998864		4					
Mean							5403.4592833876195	1621.2457980270997		7					
Min							150	0	9	0	9	12	0		
Max							81000	41667	700	7	700	480	1		
Range							80850	41667	691	7	691	468	1		
Variance							37320390.167181246	8562929.518387228		8					
Std. Dev							6109.041673387181	2926.2483692241894		4					
false															192
true															422

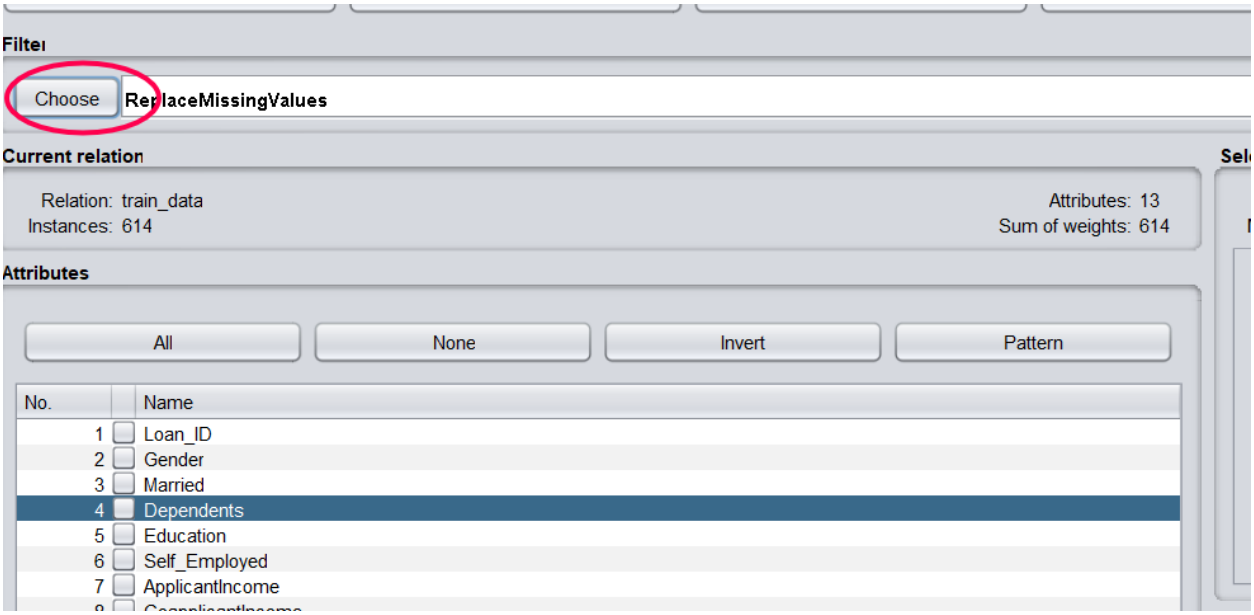
STEP 2: Preprocess data

It is important to remove/fill missing values, reducing noisy data, handling outliers, and removing duplicates to get accurate prediction results. If data is not preprocessed, the prediction results will be deviated and corrupted.

- Can be done using Java or on Weka.

Using Weka:

- Open the file that requires to be preprocessed.
- Choose the filter to be applied such as "ReplaceMissingValues" to replace all the missing values with mode and mean of that columns.
- To remove duplicates use "RemoveDuplicates" filter.
- If you feel any attribute is not needed during prediction or deviates the result, use "RemoveAttribute" filter and delete those columns.
- Save the edited file to use it for prediction.

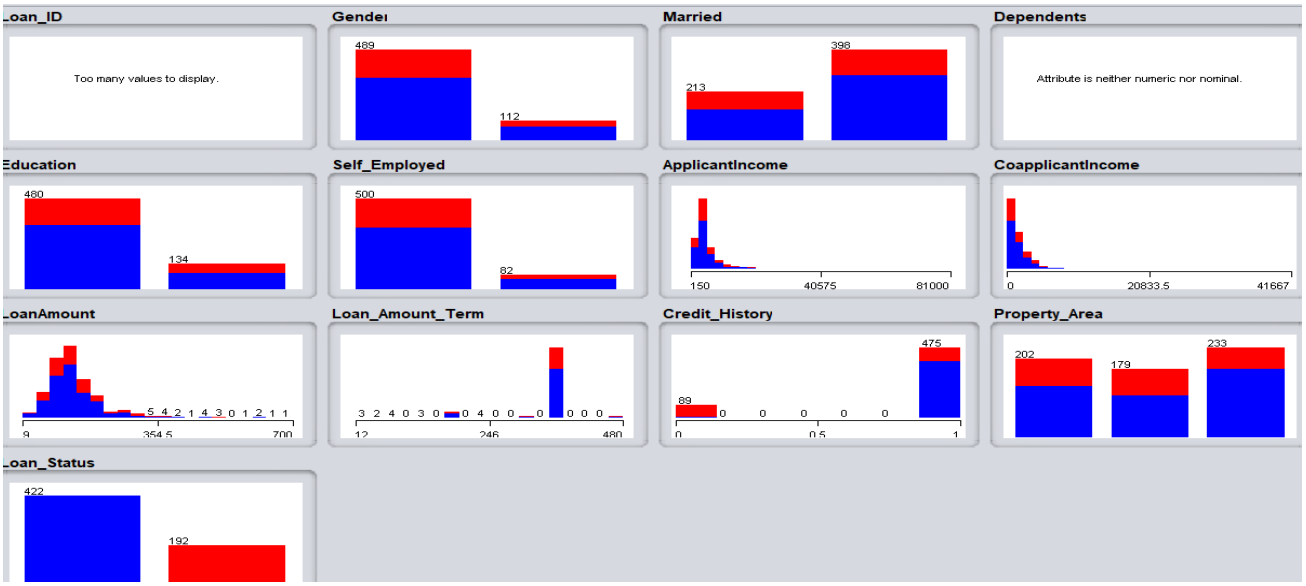


STEP 3: Visualization

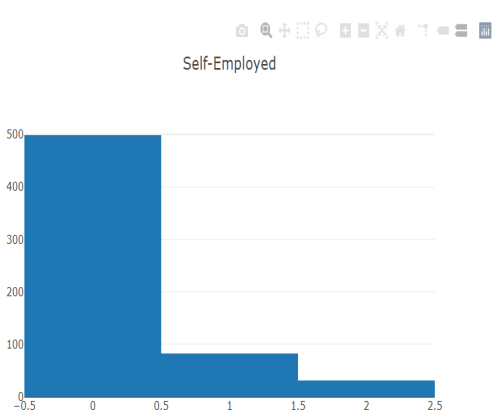
Understand the relationships between different attributes or columns.

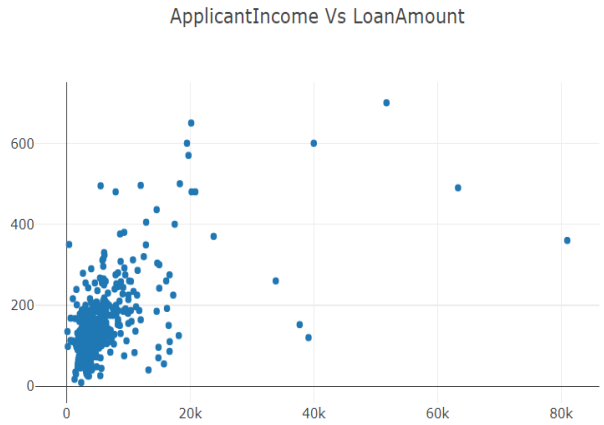
- Can be done using Java or on Weka.
- Plot any graph (scatter plot, bar plot, histogram etc.) and understand how data is classified.

Using Weka:



Using Java:

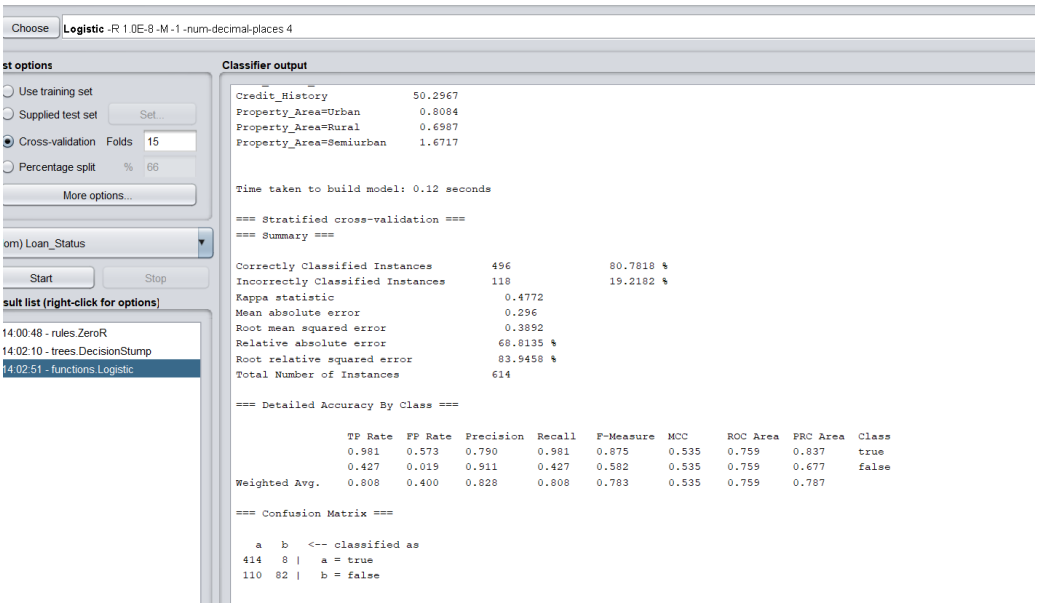




STEP 4: Prediction and Testing

- Choose a model that you think will give accurate results.
- Construct and train the model using training data.
- Pass the testing data to the trained model to check prediction accuracy.

Prediction using Weka: (i took logistic regression)



Using Java: (i took logistic regression)

```
614
Logistic Regression on loan prediction data:
summary :
Correctly Classified Instances         496               80.7818 %
Incorrectly Classified Instances       118               19.2182 %
Kappa statistic                          0.4772
Mean absolute error                      0.2911
Root mean squared error                  0.3815
Relative absolute error                 67.6867 %
Root relative squared error             82.3008 %
Total Number of Instances              614


Confusion matrix:
[414.0, 8.0]
[110.0, 82.0]

Area under the curve
0.7898030213270142
0.7898030213270142
[Correct, Incorrect, Kappa, Total cost, Average cost, KB relative, KB information, Correlation, Complexity 0, Complexity scheme, Complexity improvement, MAE, RMSE, R

Recall :
0.43
0.43

Precision:
0.91
0.91

F1 score:0.58
0.58

Accuracy:0.81
0.81
```

RESULT AND CONCLUSION

Successfully created a logistic model to predict loan eligibility given an instance with accuracy of 81% and 91% precision with the help of machine learning through Java and Weka.