

|                              |  |
|------------------------------|--|
| <b>Course</b>                |  |
| <b>Stage / Year</b>          |  |
| <b>Module</b>                |  |
| <b>Semester</b>              |  |
| <b>Assignment</b>            |  |
| <b>Date of Title Issue</b>   |  |
| <b>Assignment Deadline</b>   |  |
| <b>Assignment Submission</b> |  |
| <b>Assignment Weighting</b>  |  |

## **Objective of Assignment**

To successfully apply a set of data mining skills imparted through lectures and lab session to a previously unseen dataset using Weka to achieve knowledge discovery and producing a written technical paper format report.

## **1. Description of your dataset(s) and findings**

**1) Title:** German Credit Data

**2) Data description:**

**o The problem domain** - Credit classification, The original dataset contains 1000 entries with 20 categorial/symbolic attributes prepared by Prof. Hofmann. In this dataset, each entry represents a person who takes a credit by a bank. Each person is classified as good or bad credit risks according to the set of attributes.

**o The source of the data** -

UCI Machine Learning Repository

Professor Dr. Hans Hofmann

Institut f"ur Statistik und "Okonometrie

Universit"at Hamburg

FB Wirtschaftswissenschaften

Von-Melle-Park 5

2000 Hamburg 13

## **o The agencies working with the data**

- Open Knowledge Foundation Germany , AlgorithmWatch

## **o The intended use of the data**

- This dataset represents entries of people taking a credit by a bank. Good or Bad credit risk is analysed for each person based on the set of attributes.

## **o The attribute types of the data-**

| <b>Attributes:</b>                                  | <b>Type:</b> |
|---|--------------|
| Status of existing of checking account              | Nominal      |
| Duration in months                                  | Numeric      |
| Credit history                                      | Nominal      |
| Purpose   | Nominal      |
| Credit amount                                       | Numeric      |
| Saving accounts/bonds                               | Nominal      |
| Present employment since                            | Nominal      |
| Installment rate in percentage of disposable income | Numeric      |
| Personal and Sex                                    | Nominal      |
| Other debtors/guarantors                            | Nominal      |
| Present residence since                             | Numeric      |
| property  | Nominal      |
| Age in years  | Numeric      |
| Other installment plans                             | Nominal      |
| Housing   | Nominal      |

|  |         |
|--|---------|
| No of existing credits at this bank                  | Numeric |
| job  | Nominal |
| No of people being liable to provide maintenance for | Numeric |
| Telephone  | Nominal |
| Foreign worker                                       | Nominal |

## Description of Attributes

Attribute 1: (qualitative)

Status of existing checking account

A11 : ... < 0 DM

A12 : 0 <= ... < 200 DM

A13 : ... >= 200 DM / salary assignments for at least 1 year

A14 : no checking account

Attribute 2: (numerical)

Duration in month

Attribute 3: (qualitative)

Credit history

A30 : no credits taken/ all credits paid back duly

A31 : all credits at this bank paid back duly

A32 : existing credits paid back duly till now

A33 : delay in paying off in the past

A34 : critical account/ other credits existing (not at this bank)

Attribute 4: (qualitative)

Purpose

A40 : car (new)

A41 : car (used)

A42 : furniture/equipment

A43 : radio/television

A44 : domestic appliances

A45 : repairs

A46 : education

A47 : (vacation - does not exist?)

A48 : retraining

A49 : business

A410 : others

Attribute 5: (numerical)

Credit amount

Attribute 6: (qualitative)

Savings account/bonds

A61 : ... < 100 DM

A62 : 100 <= ... < 500 DM

A63 : 500 <= ... < 1000 DM

A64 : .. >= 1000 DM

A65 : unknown/ no savings account

Attribute 7: (qualitative)

Present employment since

A71 : unemployed

A72 : ... < 1 year

A73 : 1 <= ... < 4 years

A74 : 4 <= ... < 7 years

A75 : .. >= 7 years

Attribute 8: (numerical)

Installment rate in percentage of disposable income

Attribute 9: (qualitative)

Personal status and sex

A91 : male : divorced/separated

A92 : female : divorced/separated/married

A93 : male : single

A94 : male : married/widowed

A95 : female : single

Attribute 10: (qualitative)

Other debtors / guarantors

A101 : none

A102 : co-applicant

A103 : guarantor

Attribute 11: (numerical)

Present residence since

Attribute 12: (qualitative)

Property

A121 : real estate

A122 : if not A121 : building society savings agreement/ life insurance

A123 : if not A121/A122 : car or other, not in attribute 6

A124 : unknown / no property

Attribute 13: (numerical)

Age in years

Attribute 14: (qualitative)

Other installment plans

A141 : bank

A142 : stores

A143 : none

Attribute 15: (qualitative)

Housing

A151 : rent

A152 : own

A153 : for free

Attribute 16: (numerical)

Number of existing credits at this bank

Attribute 17: (qualitative)

Job

A171 : unemployed/ unskilled - non-resident

A172 : unskilled - resident

A173 : skilled employee / official

A174 : management/ self-employed/

highly qualified employee/ officer

Attribute 18: (numerical)

Number of people being liable to provide maintenance for

Attribute 19: (qualitative)

Telephone

A191 : none

A192 : yes, registered under the customers name

Attribute 20: (qualitative)

foreign worker

A201 : yes

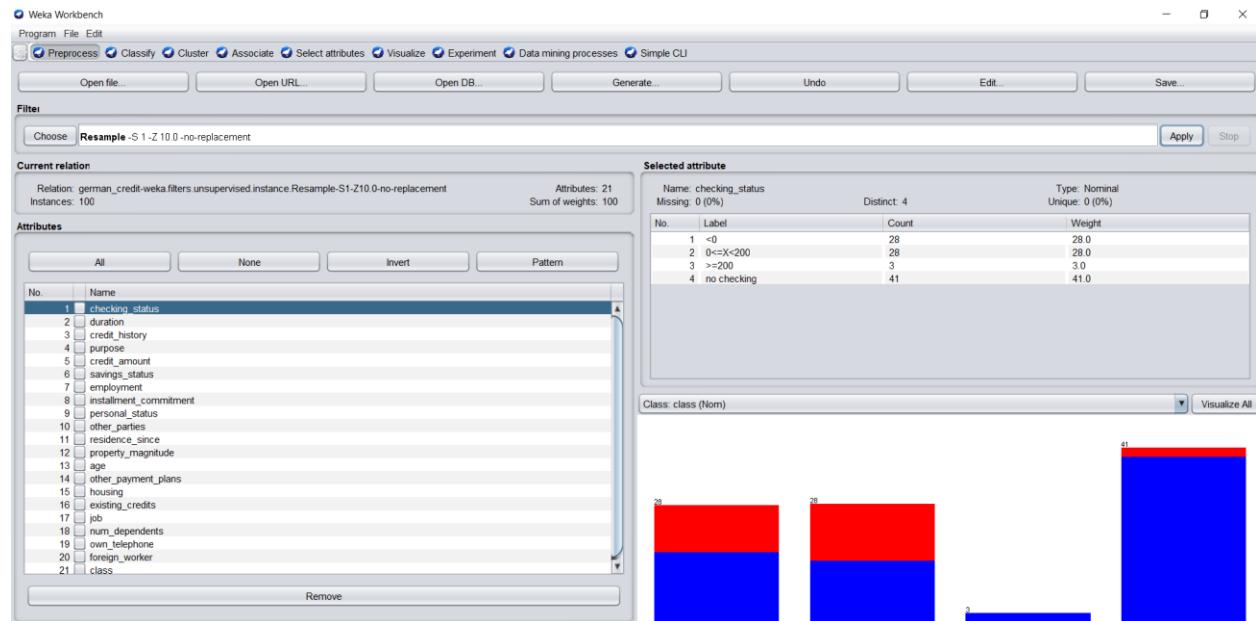
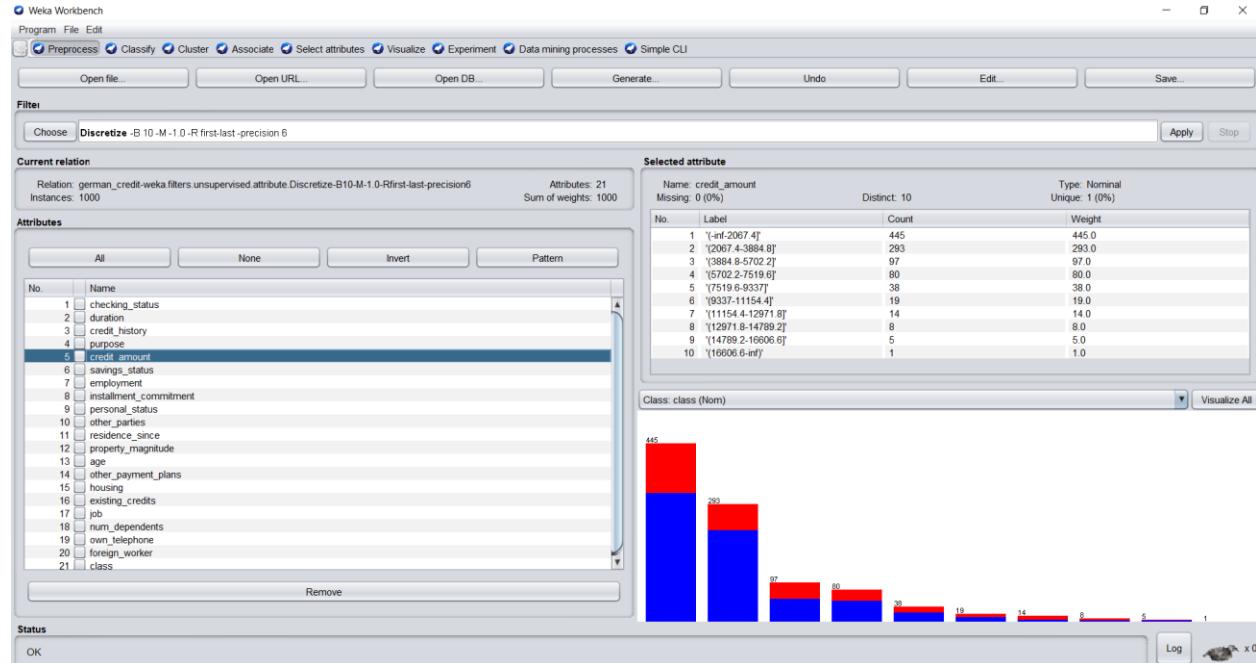
A202 : no

**Summary of Dataset:**

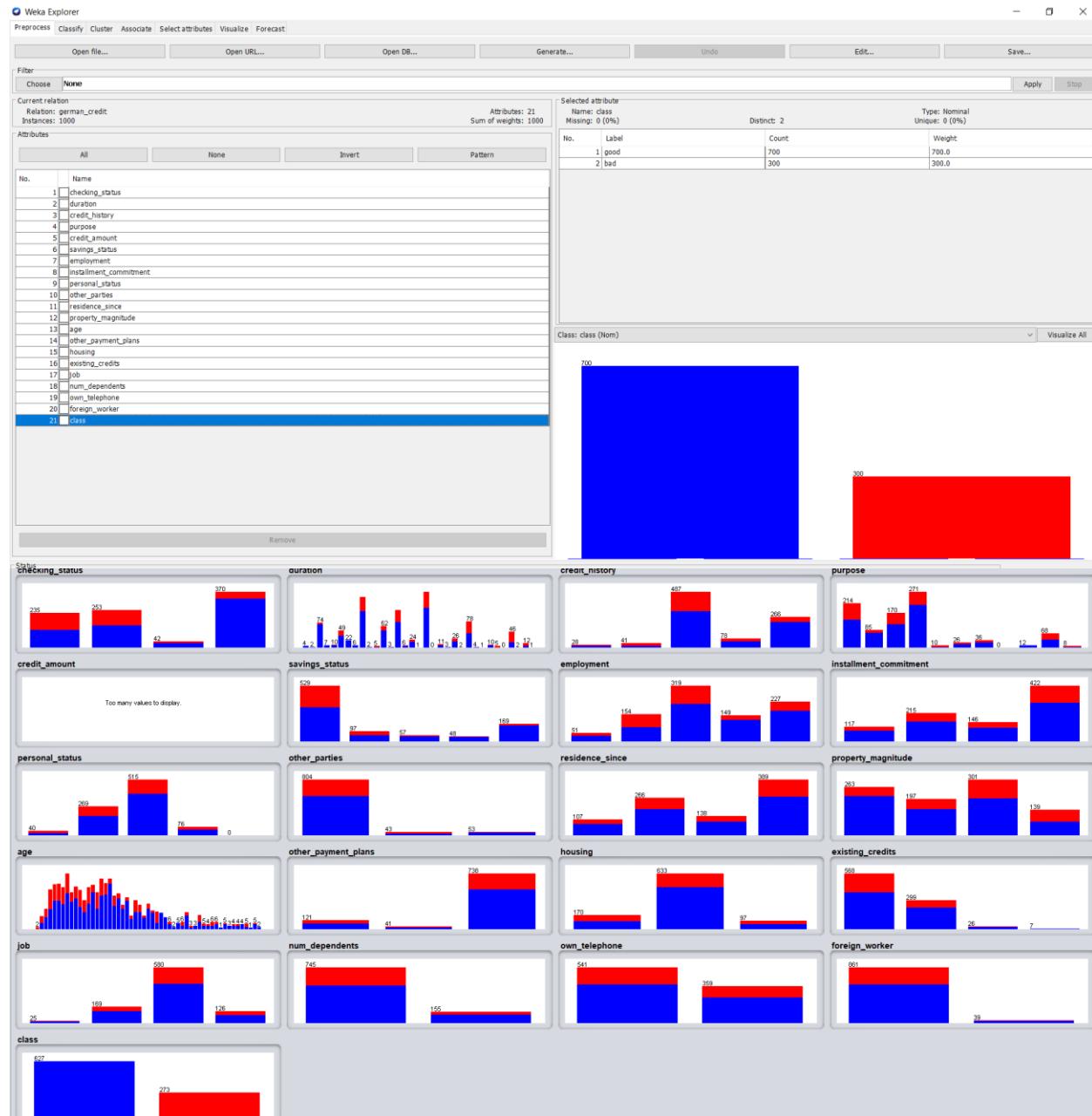
The original dataset is made up of 1000 entries with 20 categorial/symbolic attributes prepared by Prof. Hofmann. This dataset, each entry represents a person who takes a credit by a bank. Each person is classified as good or bad credit risks according to the set of attributes.

# Screen Shots of Dataset:

## All Attributes:



## Visualization of Attributes:



## **Summary on Weka:**

```
%  
% 1. Title: German Credit data  
%  
% 2. Source Information  
%  
% Professor Dr. Hans Hofmann  
% Institut f"ur Statistik und "Okonometrie  
% Universit"at Hamburg  
% FB Wirtschaftswissenschaften  
% Von-Melle-Park 5  
% 2000 Hamburg 13  
%  
% 3. Number of Instances: 1000  
%  
% Two datasets are provided. the original dataset, in the form provided  
% by Prof. Hofmann, contains categorical/symbolic attributes and  
% is in the file "german.data".  
%  
% For algorithms that need numerical attributes, Strathclyde University  
% produced the file "german.data-numeric". This file has been edited  
% and several indicator variables added to make it suitable for  
% algorithms which cannot cope with categorical variables. Several  
% attributes that are ordered categorical (such as attribute 17) have  
% been coded as integer. This was the form used by StatLog.  
%  
%  
% 6. Number of Attributes german: 20 (7 numerical, 13 categorical)  
%   Number of Attributes german.numer: 24 (24 numerical)  
%  
%  
% 7. Attribute description for german
```

**TYPE TO ENTER A CAPTION.**

%

% Attribute 1: (qualitative)

% Status of existing checking account

% A11 : ... < 0 DM

% A12 : 0 <= ... < 200 DM

% A13 : ... >= 200 DM /  
salary assignments for at least 1 year

% A14 : no checking account

%

% Attribute 2: (numerical)

% Duration in month

%

% Attribute 3: (qualitative)

% Credit history

% A30 : no credits taken/  
all credits paid back duly

% A31 : all credits at this bank paid back duly

% A32 : existing credits paid back duly till now

% A33 : delay in paying off in the past

% A34 : critical account/  
other credits existing (not at this bank)

%

% Attribute 4: (qualitative)

% Purpose

% A40 : car (new)

% A41 : car (used)

% A42 : furniture/equipment

% A43 : radio/television

% A44 : domestic appliances

% A45 : repairs

% A46 : education

% A47 : (vacation - does not exist?)

% A48 : retraining

% A49 : business

% A410 : others

%

% Attribute 5: (numerical)

### **Objective:**

To identify fraudulent Credit Card Transactions so that the customer isn't charged for items they didn't purchase.

### **Summary of Findings:**

The dataset is preprocessed using Numeric to Nominal filter as most of the data is numeric and qualitative. The DataMining techniques used are J48 Tree which is a Classification Technique giving 98% accuracy by varying the parameters. Similarly, Voted Perceptron an advanced machine learning technique available in WEKA did not produce satisfactory results as the maximum accuracy stood around 88%. Classes were also used to Cluster evaluation technique which was part of WEKA software. In Conclusion, using only the credit amount attribute , fraudulent transactions can be identified.

## **2. Preprocessing**

The dataset did not consist of any missing or duplicate values

The Set of preprocessing techniques analysed are:

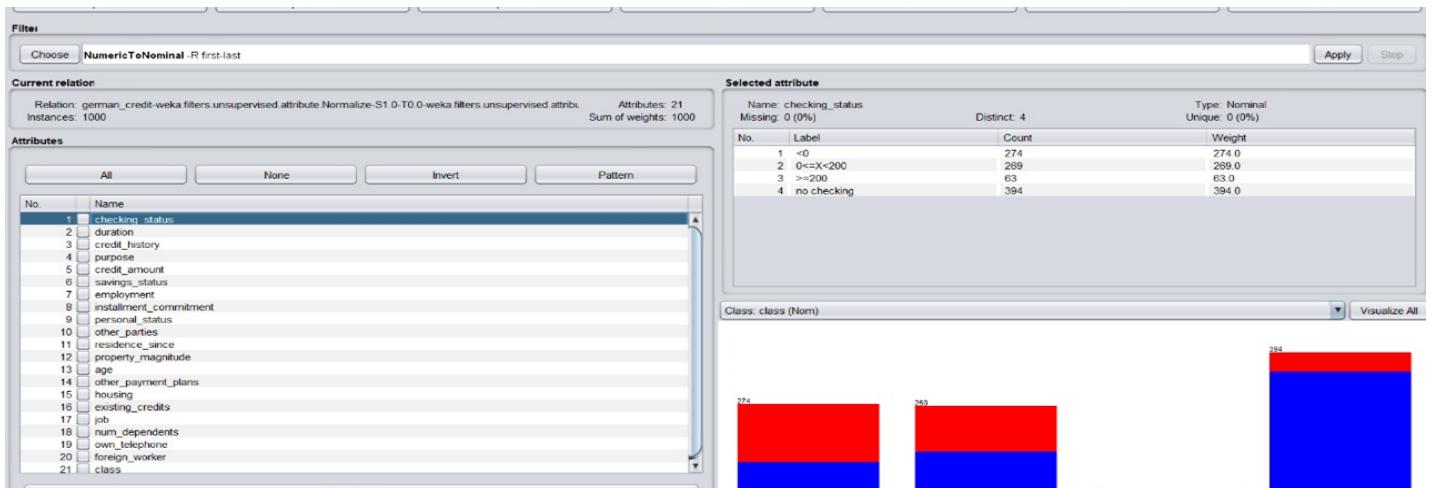
- 1) Numeric to Nominal
- 2) Nominal to Binary
- 3) Normalise
- 4) Discretise

The dataset consists of nominal and numeric values but class output is nominal-good or bad.

In this case we will have to analyze the dataset after applying the above mentioned preprocessing techniques.

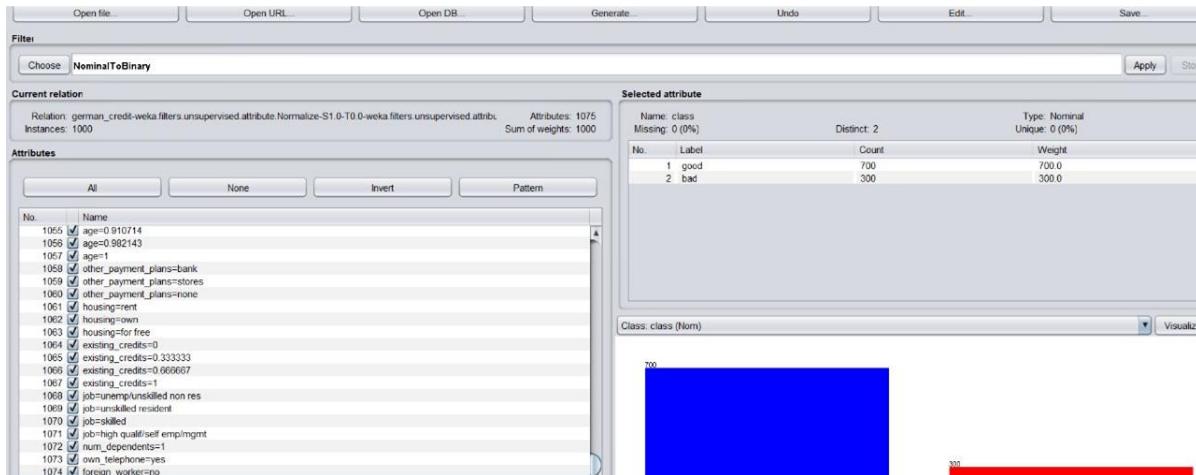
The preprocessing technique that gives best result in the form of Nominal good or bad.

Many preprocessing techniques were explored to clean the data to get desired results.



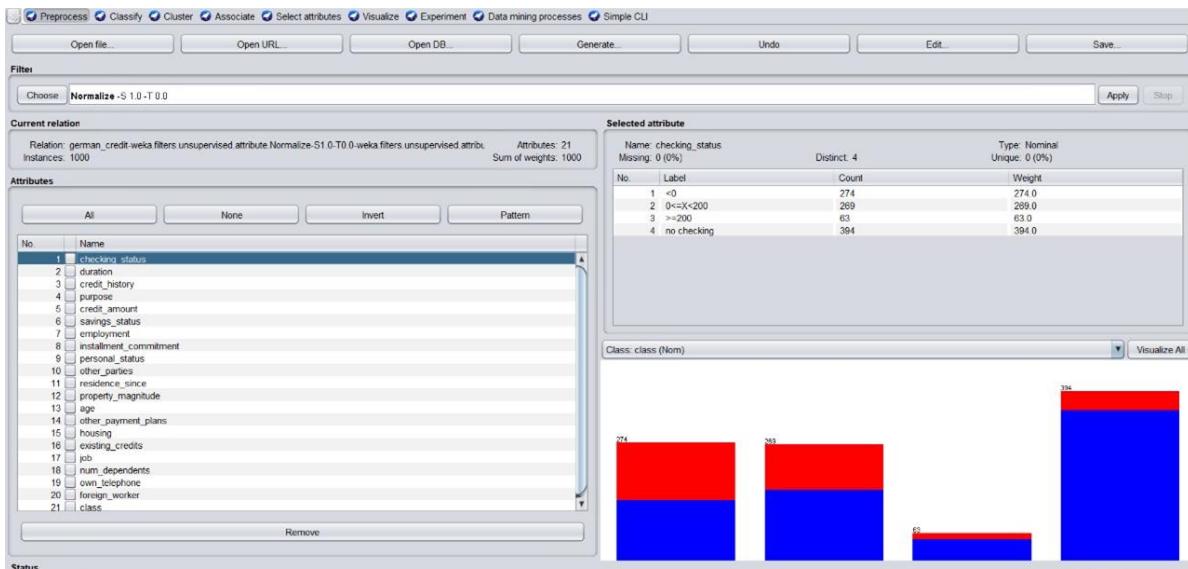
## NUMERIC TO NOMINAL

Numeric to nominal is used as all attribute values are converted to nominal form which is the technique being used.



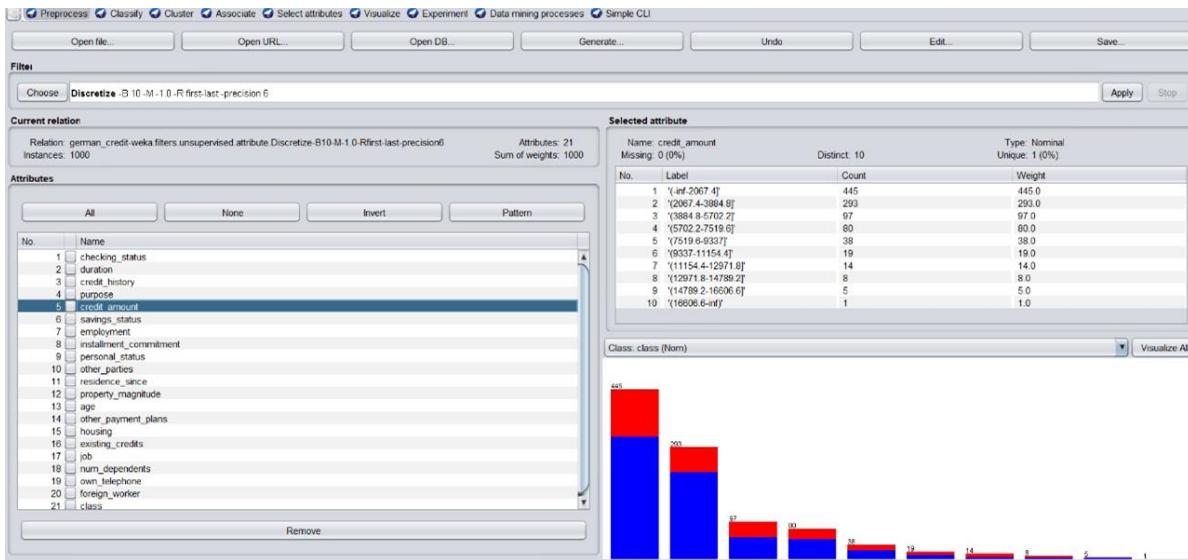
## NOMINAL TO BINARY

Nominal to binary is also not good enough as it results in more attributes.



## NORMALIZE

Normalise is another technique which was used but did not produce desired results.



## DISCRETISE

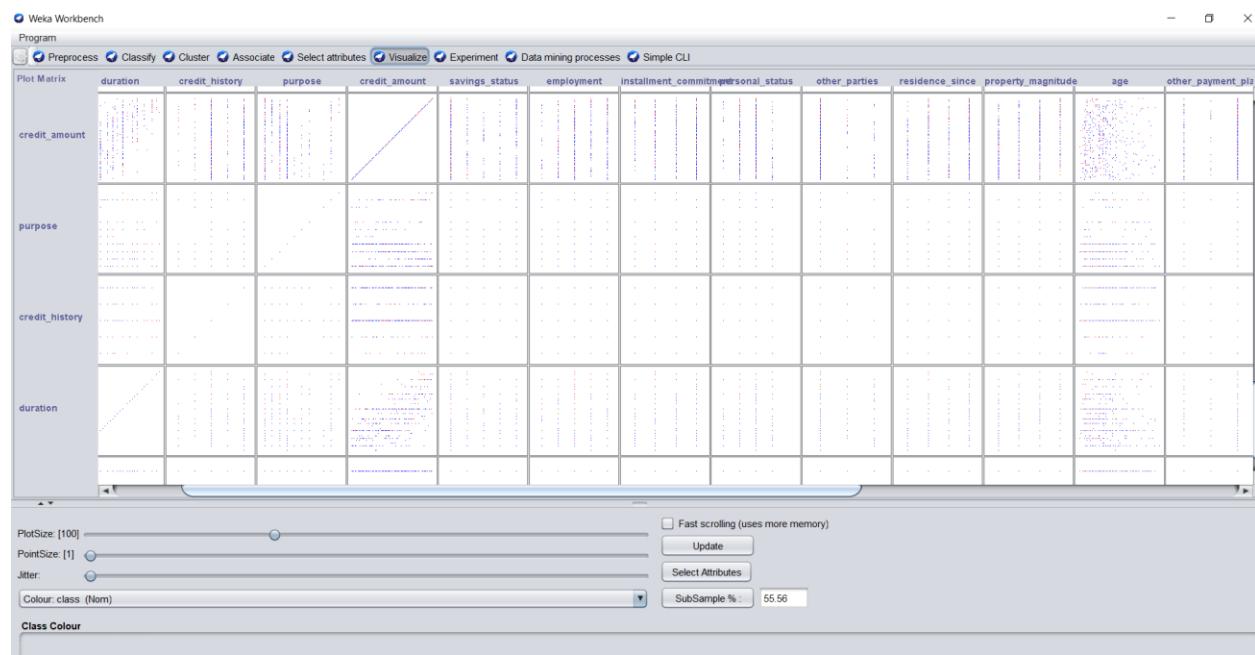
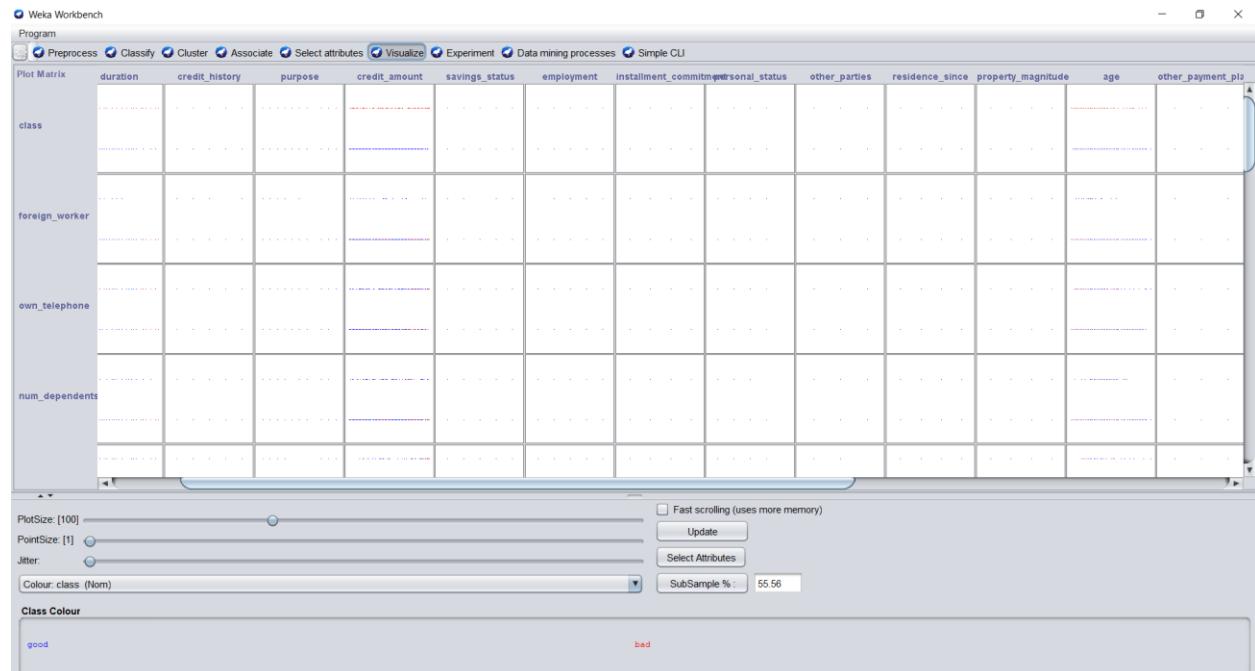
Discretise does not work as labels are changed that leads to misinformation.

- 1) The outcome is given in dataset.arff present in the zip file.

Normalize in the range [-1,1]. For ML algorithms most of the cases normalization is important as attribute values can differ in order of magnitudes.

2) Discretize all numerical values to 3 nominals. This will allow us to use J48.

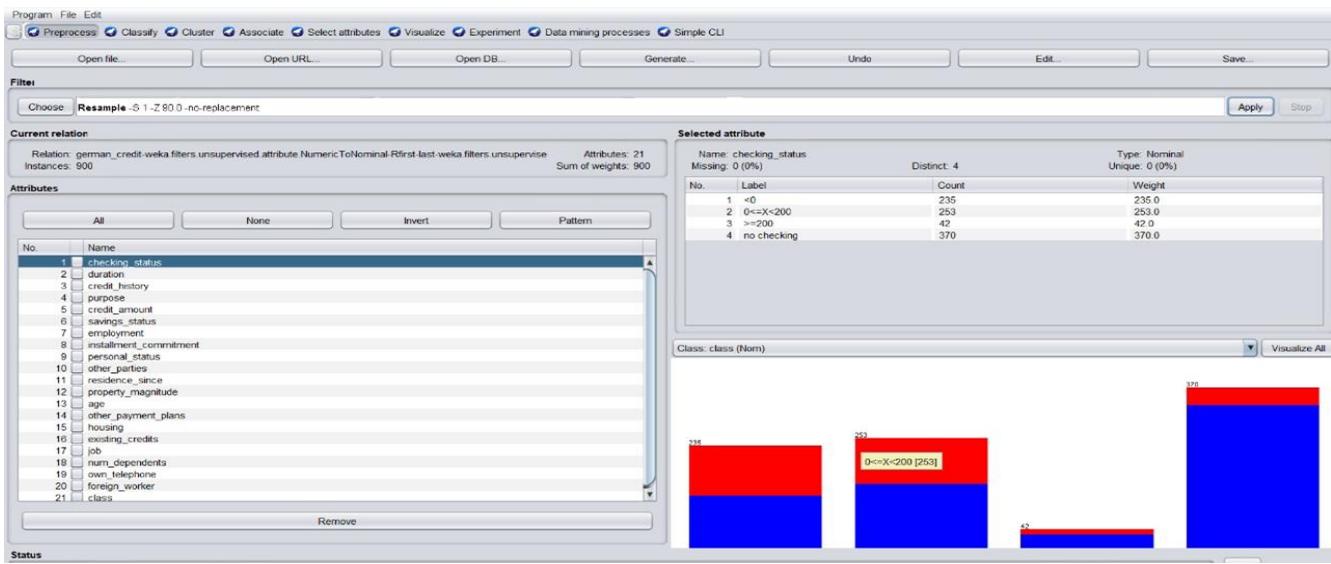
## Visualisation of dataset



### 3. Divided dataset into training and test set

The dataset is divided into training and test set into (9:1) ratio.

#### Training Set



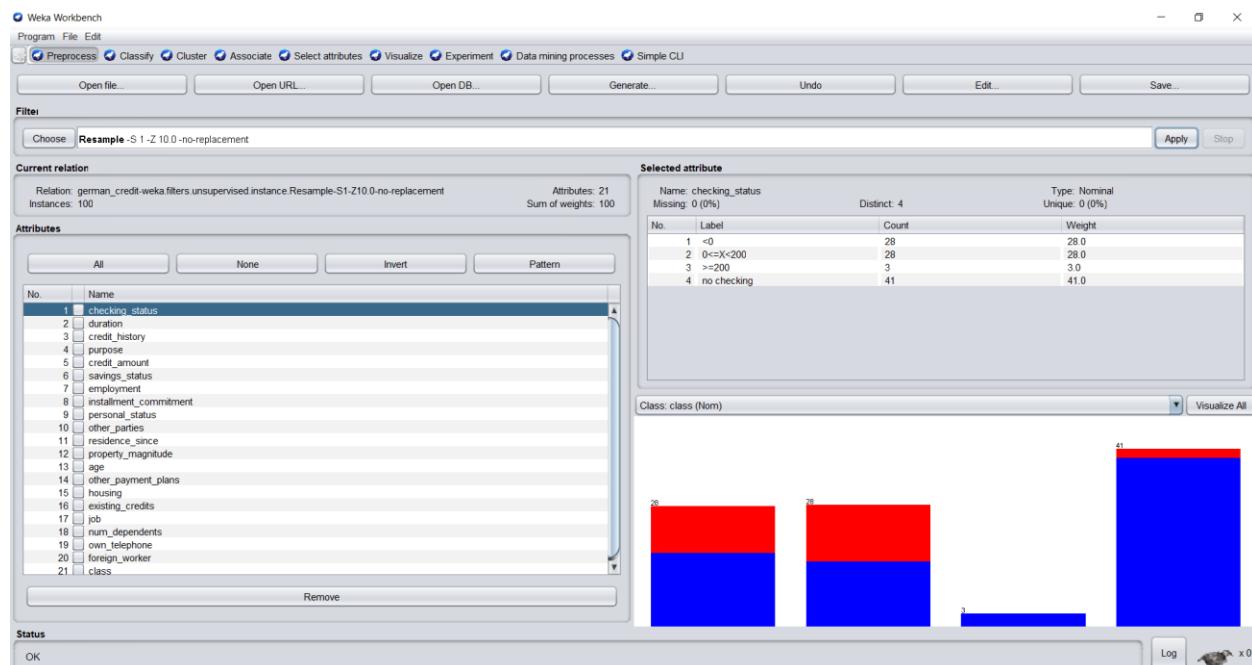
Raw view of dataset(Training)

Viewer

Relation: german\_credit-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last-weka.filters.unsupervised.instance.Resample-S1-Z100.0-weka.filters.unsupervised.instance.Resample-S1-Z90.0-no-replacement

| No. | 1. checking_status | 2. duration | 3. credit_history   | 4. purpose   | 5. credit_amount | 6. savings_status | 7. employment | 8. installment_commitment | 9. personal_status   | 10. other_parties | 11. residence_since | 12. property_magnitude | 13. age | 14. other_payment_plans | 15. housing |
|-----|--------------------|-------------|---------------------|--------------|------------------|-------------------|---------------|---------------------------|----------------------|-------------------|---------------------|------------------------|---------|-------------------------|-------------|
|     | Nominal            | Nominal     | Nominal             | Nominal      | Nominal          | Nominal           | Nominal       | Nominal                   | Nominal              | Nominal           | Nominal             | Nominal                | Nominal | Nominal                 | Nominal     |
| ... | no checking        | 24          | existing paid       | furniture... | 3062             | 500(<X 1000)      | =7            | 4                         | male single          | none              | 3                   | no known property      | 32      | none                    | rent        |
| ... | (0                 | 18          | critical/other e... | furniture... | 1049             | (100              | (1            | 4                         | female dividdepim... | none              | 4                   | life insurance         | 21      | none                    | rent        |
| ... | no checking        | 18          | critical/other e... | radio/tv     | 6070             | (100              | =7            | 3                         | male single          | none              | 4                   | car                    | 33      | none                    | own         |
| ... | (0                 | 12          | existing paid       | furniture... | 652              | (100              | =7            | 4                         | female dividdepim... | none              | 4                   | life insurance         | 24      | none                    | rent        |
| ... | no checking        | 18          | existing paid       | new car      | 2662             | no known savin... | 4(<X 7        | 4                         | male single          | none              | 3                   | life insurance         | 32      | none                    | own         |
| ... | no checking        | 18          | critical/other e... | furniture... | 3780             | (100              | (1            | 3                         | male divsep          | none              | 2                   | car                    | 35      | none                    | own         |
| ... | no checking        | 24          | existing paid       | radio/tv     | 3181             | (100              | (1            | 4                         | female dividdepim... | none              | 4                   | life insurance         | 26      | none                    | own         |
| ... | no checking        | 15          | existing paid       | radio/tv     | 1386             | no known savin... | 1(<X 4        | 4                         | male mar/wid         | none              | 2                   | real estate            | 40      | none                    | rent        |
| ... | (0=>200            | 24          | delayed previo...   | furniture... | 2064             | (100              | unemployed    | 3                         | female dividdepim... | none              | 2                   | life insurance         | 34      | none                    | own         |
| ... | no checking        | 12          | critical/other e... | radio/tv     | 2331             | no known savin... | =7            | 1                         | male single          | co applicant      | 4                   | real estate            | 49      | none                    | own         |
| ... | no checking        | 6           | existing paid       | furniture... | 2978             | 500(>X 1000       | 1(<X 4        | 1                         | male single          | none              | 2                   | car                    | 32      | none                    | own         |
| ... | (0                 | 18          | all paid            | new car      | 1442             | (100              | 4(<X 7        | 4                         | male single          | none              | 4                   | no known property      | 32      | none                    | for free    |
| ... | (0                 | 36          | existing paid       | new car      | 1842             | (100              | (1            | 4                         | female dividdepim... | none              | 4                   | car                    | 34      | none                    | own         |
| ... | (0                 | 12          | existing paid       | new car      | 2579             | (100              | (1            | 4                         | male single          | none              | 1                   | real estate            | 33      | none                    | own         |
| ... | no checking        | 18          | existing paid       | radio/tv     | 433              | (100              | unemployed    | 3                         | female dividdepim... | co applicant      | 4                   | real estate            | 22      | none                    | rent        |
| ... | (0                 | 16          | critical/other e... | new car      | 2625             | (100              | =7            | 2                         | male single          | guarantor         | 4                   | life insurance         | 43      | bank                    | rent        |
| ... | (0=>200            | 48          | existing paid       | furniture... | 9960             | (100              | (1            | 1                         | female divdepim...   | none              | 2                   | car                    | 26      | none                    | own         |
| ... | no checking        | 30          | delayed previo...   | business     | 4272             | 100(>X 500        | 1(<X 4        | 2                         | male single          | none              | 2                   | life insurance         | 26      | none                    | own         |
| ... | no checking        | 48          | existing paid       | business     | 3914             | no known savin... | 1(<X 4        | 4                         | male divsep          | none              | 2                   | real estate            | 38      | bank                    | own         |
| ... | no checking        | 36          | existing paid       | business     | 7409             | no known savin... | =7            | 3                         | male single          | none              | 2                   | life insurance         | 37      | none                    | own         |
| ... | (0=>200            | 9           | existing paid       | furniture... | 2030             | no known savin... | 4(<X 7        | 2                         | male single          | none              | 1                   | car                    | 24      | none                    | own         |
| ... | no checking        | 18          | existing paid       | radio/tv     | 2051             | (100              | (1            | 4                         | male single          | none              | 1                   | real estate            | 33      | none                    | own         |
| ... | no checking        | 21          | existing paid       | business     | 1572             | =1000(<X 4        | =7            | 4                         | female dividdepim... | none              | 4                   | real estate            | 36      | bank                    | own         |
| ... | (0=>200            | 21          | critical/other e... | business     | 3652             | (100              | 4(<X 7        | 2                         | male single          | none              | 3                   | life insurance         | 27      | none                    | own         |
| ... | (0                 | 36          | existing paid       | radio/tv     | 2302             | (100              | 1(<X 4        | 4                         | male divsep          | none              | 4                   | car                    | 31      | none                    | rent        |
| ... | (0                 | 18          | critical/other e... | new car      | 3968             | (100              | =7            | 1                         | female dividdepim... | none              | 4                   | real estate            | 33      | bank                    | rent        |
| ... | (0                 | 8           | critical/other e... | new car      | 731              | (100              | =7            | 4                         | male single          | none              | 4                   | real estate            | 47      | none                    | own         |
| ... | (0=>200            | 24          | existing paid       | radio/tv     | 5084             | no known savin... | =7            | 2                         | female dividdepim... | none              | 4                   | car                    | 42      | none                    | own         |
| ... | (0                 | 12          | critical/other e... | used car     | 1526             | (100              | =7            | 4                         | male single          | none              | 4                   | no known property      | 66      | none                    | for free    |
| ... | (0=>200            | 48          | no credits/all p... | business     | 3844             | 100(>X 500        | 4(<X 7        | 4                         | male single          | none              | 4                   | no known property      | 34      | none                    | for free    |
| ... | no checking        | 18          | critical/other e... | radio/tv     | 1169             | no known savin... | 1(<X 4        | 4                         | male single          | none              | 3                   | life insurance         | 29      | none                    | own         |
| ... | (0                 | 30          | no credits/all p... | furniture... | 4583             | (100              | 1(<X 4        | 2                         | male divsep          | guarantor         | 2                   | real estate            | 32      | none                    | own         |
| ... | (0                 | 12          | all paid            | new car      | 697              | (100              | (1            | 4                         | male single          | none              | 2                   | car                    | 46      | bank                    | own         |
| ... | no checking        | 18          | critical/other e... | radio/tv     | 1800             | (100              | 1(<X 4        | 4                         | male single          | none              | 2                   | car                    | 24      | none                    | own         |
| ... | (0=>200            | 6           | all paid            | new car      | 931              | 100(>X 500        | (1            | 1                         | female dividdepim... | none              | 1                   | life insurance         | 32      | stores                  | own         |
| ... | (0                 | 24          | existing paid       | furniture... | 7721             | no known savin... | (1            | 1                         | female dividdepim... | none              | 2                   | life insurance         | 30      | none                    | own         |
| ... | no checking        | 36          | critical/other e... | business     | 6304             | no known savin... | =7            | 4                         | male single          | none              | 4                   | real estate            | 36      | none                    | own         |
| ... | no checking        | 15          | critical/other e... | education    | 1532             | 100(>X 500        | 1(<X 4        | 4                         | female dividdepim... | none              | 3                   | car                    | 31      | none                    | own         |
| ... | no checking        | 12          | critical/other e... | new car      | 682              | 100(>X 500        | 4(<X 7        | 4                         | female dividdepim... | none              | 3                   | car                    | 51      | none                    | own         |
| ... | no checking        | 12          | critical/other e... | radio/tv     | 1934             | (100              | =7            | 2                         | male single          | none              | 2                   | no known property      | 26      | none                    | own         |
| ... | no checking        | 36          | critical/other e... | furniture... | 7127             | (100              | (1            | 2                         | female dividdepim... | none              | 4                   | life insurance         | 23      | none                    | rent        |
| ... | (0                 | 24          | existing paid       | furniture... | 2996             | no known savin... | 1(<X 4        | 2                         | male mar/wid         | none              | 4                   | car                    | 20      | none                    | own         |
| ... | (0                 | 9           | existing paid       | radio/tv     | 1364             | (100              | 4(<X 7        | 3                         | male single          | none              | 4                   | real estate            | 59      | none                    | own         |

## Test Set



## Raw view dataset (Test Set)

Viewer

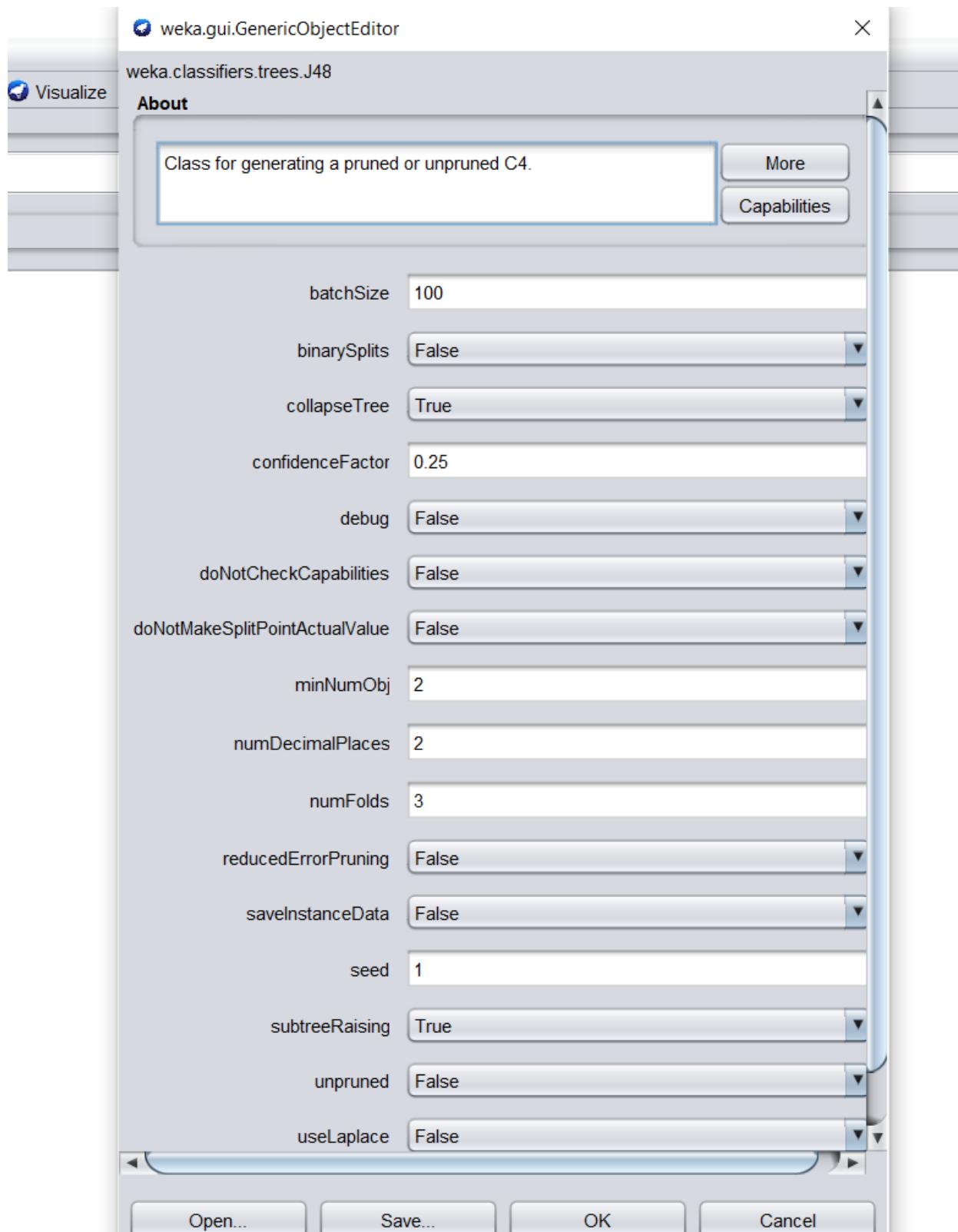
Relation: german\_credit-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last-weka.filters.unsupervised.instance.Resample-S1-Z10.0

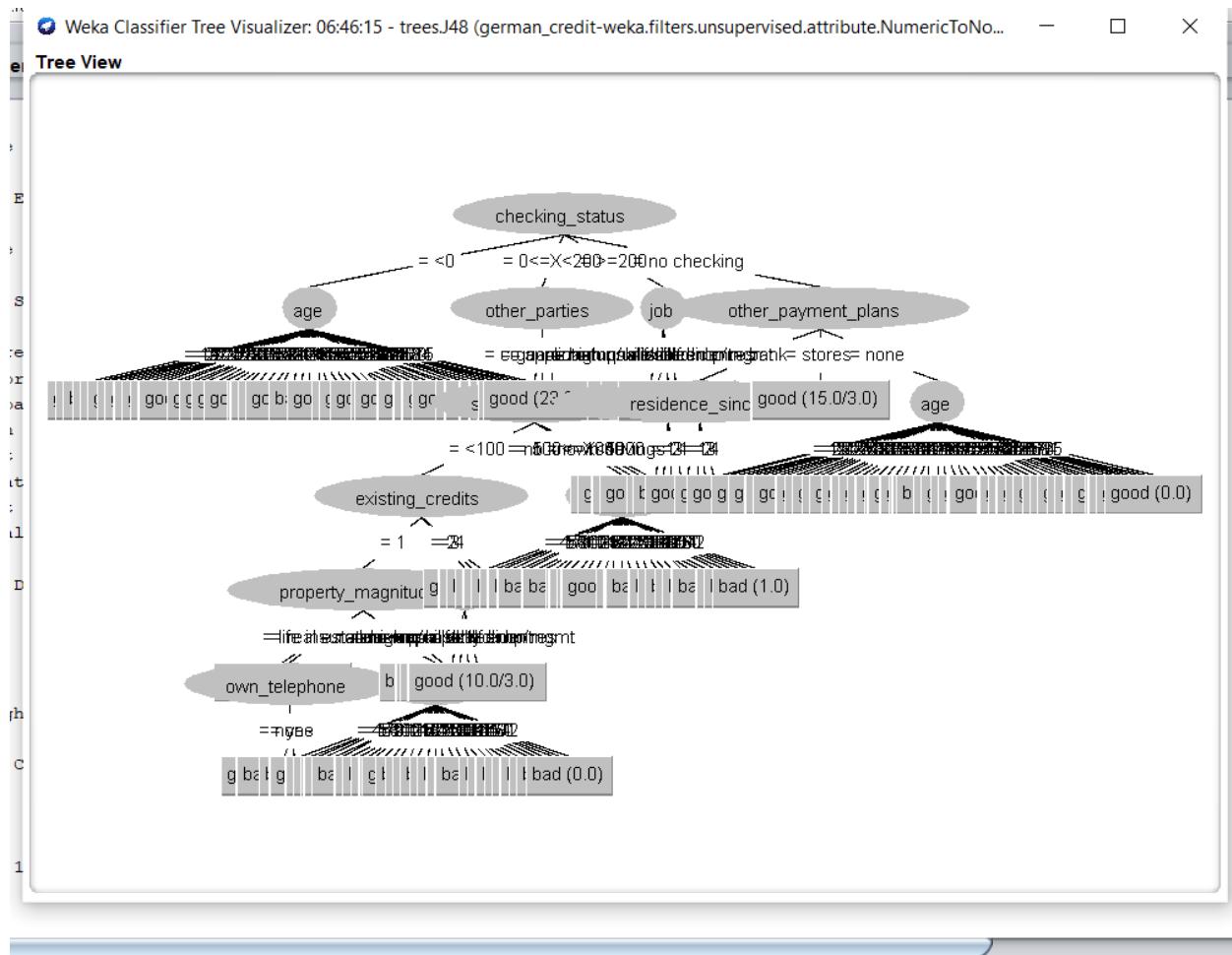
| No | 1: checking_status | 2: duration   | 3: credit_history   | 4: purpose    | 5: credit_amount | 6: savings_status | 7: employment | 8: installment_commitment | 9: personal_status  | 10: other_parties | 11: residence_since | 12: property_magnitude | 13: age | 14: other_payment_plans | 15: housing |
|----|--------------------|---------------|---------------------|---------------|------------------|-------------------|---------------|---------------------------|---------------------|-------------------|---------------------|------------------------|---------|-------------------------|-------------|
|    | Nominal            | Nominal       | Nominal             | Nominal       | Nominal          | Nominal           | Nominal       | Nominal                   | Nominal             | Nominal           | Nominal             | Nominal                | Nominal | Nominal                 | Nominal     |
| 58 | (0                 | 24            | existing paid       | new car       | 1207             | (100              | (1            | 4                         | male single         | none              | 1                   | no known property      | 22      | bank                    | for free    |
| 59 | (0                 | 24            | existing paid       | furniture/... | 4057             | (100              | 4=(X)7        | 3                         | female div/dep/m... | none              | 4                   | life insurance         | 24      | none                    | rent        |
| 60 | (0=>X200           | 7             | existing paid       | radio/tv      | 1009             | (100              | (1            | 1                         | male div/sep        | none              | 3                   | car                    | 43      | none                    | own         |
| 61 | (0=>X200           | 7             | existing paid       | radio/tv      | 1924             | (100              | 1=(X)4        | 1                         | female div/dep/m... | guarantor         | 1                   | real estate            | 45      | none                    | own         |
| 62 | no checking        | 10            | existing paid       | radio/tv      | 3077             | (100              | 1=(X)4        | 2                         | male single         | none              | 4                   | life insurance         | 36      | none                    | own         |
| 63 | no checking        | 12            | existing paid       | radio/tv      | 5103             | (100              | (1            | 3                         | male mar/wid        | none              | 4                   | car                    | 52      | none                    | own         |
| 64 | no checking        | 24            | critical/other e... | radio/tv      | 2775             | (100              | 4=(X)7        | 2                         | male single         | none              | 2                   | life insurance         | 31      | bank                    | own         |
| 65 | no checking        | 18            | critical/other e... | new car       | 3857             | (100              | 1=(X)4        | 4                         | male div/sep        | none              | 4                   | life insurance         | 40      | none                    | own         |
| 66 | (0                 | 30            | existing paid       | used car      | 1961             | (100              | 1=(X)4        | 4                         | female div/dep/m... | none              | 2                   | life insurance         | 46      | none                    | own         |
| 67 | no checking        | 15            | delayed previo...   | used car      | 3594             | (100              | (1            | 1                         | male single         | none              | 1                   | life insurance         | 54      | none                    | own         |
| 68 | no checking        | 24            | existing paid       | new car       | 2255             | no known savin... | (1            | 4                         | male single         | none              | 2                   | car                    | 20      | none                    | rent        |
| 69 | no checking        | 15            | existing paid       | furniture/... | 2221             | 5000=>(1000       | 1=(X)4        | 2                         | female div/dep/m... | none              | 4                   | car                    | 31      | none                    | rent        |
| 70 | (0                 | 36            | existing paid       | radio/tv      | 2302             | (100              | 1=(X)4        | 4                         | male div/sep        | none              | 4                   | car                    | 26      | none                    | rent        |
| 71 | no checking        | 9             | existing paid       | new car       | 3577             | 1000=>(500        | 1=(X)4        | 1                         | male single         | guarantor         | 2                   | real estate            | 40      | none                    | own         |
| 72 | no checking        | 30            | critical/other e... | radio/tv      | 3077             | no known savin... | >7            | 3                         | male single         | none              | 2                   | car                    | 48      | bank                    | own         |
| 73 | (0=>X200           | 48            | no credits/all p... | business      | 12204            | no known savin... | 1=(X)4        | 2                         | male single         | none              | 2                   | car                    | 23      | none                    | own         |
| 74 | =>Z=200            | 18            | existing paid       | new car       | 1961             | (100              | >7            | 3                         | female div/dep/m... | none              | 2                   | no known property      | 53      | none                    | for free    |
| 75 | (0                 | 48            | no credits/all p... | furniture/... | 7119             | (100              | 1=(X)4        | 3                         | male single         | none              | 4                   | life insurance         | 47      | none                    | own         |
| 76 | (0=>X200           | 36            | existing paid       | education     | 12612            | 1000=>(1500       | 1=(X)4        | 1                         | male single         | none              | 2                   | car                    | 36      | none                    | own         |
| 77 | (0=>X200           | 27            | existing paid       | business      | 1015             | (100              | 1=(X)4        | 4                         | male single         | none              | 2                   | real estate            | 21      | none                    | rent        |
| 78 | (0                 | 24            | existing paid       | radio/tv      | 1887             | (100              | 1=(X)4        | 2                         | female div/dep/m... | none              | 2                   | real estate            | 22      | none                    | own         |
| 79 | (0=>X200           | 48            | existing paid       | radio/tv      | 5981             | (100              | 1=(X)4        | 2                         | female div/dep/m... | none              | 4                   | life insurance         | 22      | none                    | rent        |
| 80 | no checking        | 12            | critical/other e... | furniture/... | 1258             | (100              | (1            | 2                         | female div/dep/m... | none              | 4                   | car                    | 35      | none                    | own         |
| 81 | no checking        | 9             | existing paid       | radio/tv      | 2753             | 1000=>(500        | >7            | 3                         | male single         | co applicant      | 4                   | life insurance         | 29      | bank                    | own         |
| 82 | =>Z=00             | 6             | delayed previo...   | radio/tv      | 683              | (100              | (1            | 2                         | female div/dep/m... | none              | 1                   | real estate            | 51      | none                    | own         |
| 83 | no checking        | 6             | existing paid       | radio/tv      | 1595             | (100              | 4=(X)7        | 3                         | male single         | none              | 2                   | life insurance         | 25      | none                    | own         |
| 84 | (0=>X200           | 9             | existing paid       | radio/tv      | 1206             | (100              | >7            | 4                         | female div/dep/m... | none              | 4                   | real estate            | 37      | none                    | rent        |
| 85 | (0=>X200           | 6             | delayed previo...   | new car       | 1209             | (100              | unemployed    | 4                         | male single         | none              | 4                   | life insurance         | 47      | none                    | own         |
| 86 | no checking        | 36            | existing paid       | radio/tv      | 2394             | no known savin... | 1=(X)4        | 4                         | female div/dep/m... | none              | 4                   | car                    | 25      | none                    | own         |
| 87 | no checking        | 6             | critical/other e... | new car       | 362              | 1000=>(500        | 1=(X)4        | 4                         | female div/dep/m... | none              | 4                   | car                    | 52      | none                    | own         |
| 88 | no checking        | 6             | existing paid       | used car      | 1236             | 5000=>(1000       | 1=(X)4        | 2                         | male single         | none              | 4                   | life insurance         | 50      | none                    | rent        |
| 89 | (0                 | 12            | critical/other e... | new car       | 691              | (100              | >7            | 4                         | male single         | none              | 3                   | life insurance         | 35      | none                    | own         |
| 90 | (0=>X200           | 8             | existing paid       | business      | 907              | (100              | (1            | 3                         | male mar/wid        | none              | 2                   | real estate            | 26      | none                    | own         |
| 91 | (0                 | 6             | critical/other e... | new car       | 3676             | (100              | 1=(X)4        | 1                         | male single         | none              | 3                   | real estate            | 37      | none                    | rent        |
| 92 | no checking        | 15            | critical/other e... | education     | 1532             | 1000=>(500        | 1=(X)4        | 4                         | female div/dep/m... | none              | 3                   | car                    | 31      | none                    | own         |
| 93 | (0=>X200           | 60            | all public services | radio/tv      | 1092             | 1000=>(500        | >7            | 3                         | female div/dep/m... | none              | 4                   | no known property      | 66      | bank                    | for free    |
| 94 | (0=>X200           | 21            | critical/other e... | business      | 3652             | (100              | 4=(X)7        | 2                         | male single         | none              | 3                   | life insurance         | 27      | none                    | own         |
| 95 | no checking        | 18            | existing paid       | used car      | 3378             | no known savin... | 1=(X)4        | 2                         | male single         | none              | 1                   | life insurance         | 31      | none                    | own         |
| 96 | (0                 | 21            | critical/other e... | new car       | 1602             | (100              | >7            | 4                         | male mar/wid        | none              | 3                   | car                    | 30      | none                    | own         |
| 97 | (0=>X200           | 36            | existing paid       | radio/tv      | 2323             | (100              | 4=(X)7        | 4                         | male single         | none              | 4                   | car                    | 24      | none                    | rent        |
| 98 | (0=>X200           | 9             | existing paid       | furniture/... | 918              | (100              | 1=(X)4        | 4                         | female div/dep/m... | none              | 1                   | life insurance         | 30      | none                    | own         |
| 99 | no checking        | 6             | existing paid       | repairs       | 660              | 5000=>(1000       | 4=(X)7        | 2                         | male mar/wid        | none              | 4                   | real estate            | 23      | none                    | rent        |
| 0  | 48                 | existing paid | education           | 7476          | (100             | 4=(X)7            | 4             | male single               | none                | 1                 | no known property   | 50                     | none    | for free                |             |

## 4. Classification/ Association: J48 Tree or Association Rules

### J48

#### Model Parameters

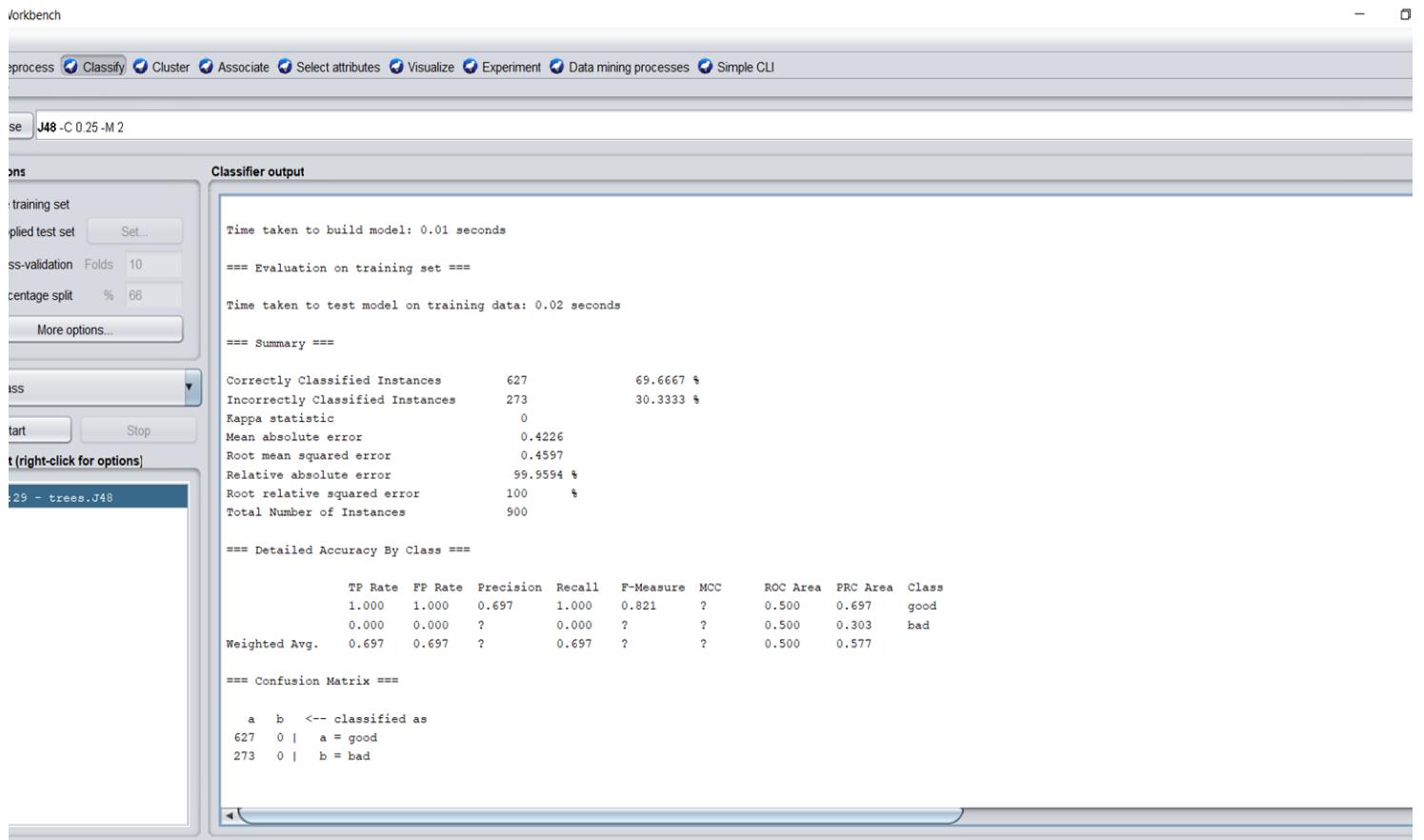




Tree

Training J48 using 900 instances and using confidence factors and number of objects as parameters also include other parameters like number of folds and seed.

Train: C = 0.25,M = 2



### Detailed Run output with confusion matrix and tree architecture:

==== Run information ====

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: german\_credit-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last-weka.filters.unsupervised.instance.Resample-S1-Z100.0-weka.filters.unsupervised.instance.Resample-S1-Z90.0-no-replacement

Instances: 900

Attributes: 21

checking\_status  
duration  
credit\_history  
purpose  
credit\_amount  
savings\_status  
employment  
installment\_commitment  
personal\_status  
other\_parties  
residence\_since  
property\_magnitude  
age  
other\_payment\_plans  
housing  
existing\_credits  
job  
num\_dependents  
own\_telephone  
foreign\_worker  
class

Test mode: evaluate on training data

==== Classifier model (full training set) ===

## J48 pruned tree

---

: good (900.0/273.0)

Number of Leaves : 1

Size of the tree : 1

Train : C=0.55,M=1

```

Program    Preprocess    Classify    Cluster    Associate    Select attributes    Visualize    Experiment    Data mining processes    Simple CLU
Classifier    Choose: J48-C 0.55-M 1

Test options
 Use training set
 Supplied test set
 Cross-validation Folds: 10
 Percentage split %
More options...

Result list (right-click for options)
Start    Stop
04:05:29 - trees,J48
04:06:54 - trees,J48
04:16:16 - Functions.MultilayerPerceptron
04:17:25 - Functions.VotedPerceptron
04:27:25 - Functions.VotedPerceptron
04:33:01 - Functions.RND
04:32:37 - Functions.RND
04:35:05 - Functions.VotedPerceptron
04:49:21 - rules.DecoR
04:24:57 - trees,J48
04:37:04 - trees,J48

Classifier output
Time taken to build model: 0.37 seconds
*** Evaluation on training set ***
Time taken to test model on training data: 0 seconds
*** Summary ***
Correctly Classified Instances 884 59.2222 %
Incorrectly Classified Instances 16 1.7778 %
Kappa statistic 0.9576
Mean absolute error 0.0202
Root mean squared error 0.1914
Relative absolute error 4.4616 %
Root relative squared error 22.0535 %
Total Number of Instances 900

*** Detailed Accuracy By Class ***
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.994 0.044 0.981 0.994 0.987 0.998 0.999 0.999 good
0.958 0.006 0.985 0.956 0.970 0.958 0.999 0.997 had
Weighted Avg. 0.982 0.033 0.982 0.982 0.982 0.998 0.999 0.999

*** Confusion Matrix ***
a b <-- classified as
623 4 | a = good
12 261 | b = bad

```

Train: C= 0.6,M=5

```

Program    Preprocess    Classify    Cluster    Associate    Select attributes    Visualize    Experiment    Data mining processes    Simple CLU    Time series forecasting
Classifier    Choose: J48-C 0.6-M 5

Test options
 Use training set
 Supplied test set
 Cross-validation Folds: 10
 Percentage split %
More options...

Result list (right-click for options)
Start    Stop
14:50:11 - trees,J48
14:54:41 - trees,J48
14:54:41 - trees,J48
14:55:11 - trees,J48
14:55:11 - trees,J48
14:59:40 - Functions.VotedPerceptron
18:50:40 - trees,J48

Classifier output
Numbers of Leaves : 211
Size of the tree : 272
Time taken to build model: 1.41 seconds
Time taken to test model on training data: 0.14 seconds
*** Evaluation on training set ***
Time taken to build model: 1.41 seconds
Time taken to test model on training data: 0.14 seconds
*** Summary ***
Correctly Classified Instances 874 87.4 %
Incorrectly Classified Instances 126 12.6 %
Kappa statistic 0.6863
Mean absolute error 0.1794
Root mean squared error 0.2397
Relative absolute error 42.7054 %
Root relative squared error 65.3419 %
Total Number of Instances 1000

*** Detailed Accuracy By Class ***
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.943 0.287 0.885 0.943 0.913 0.951 0.937 0.969 good
0.713 0.057 0.849 0.713 0.773 0.691 0.937 0.845 had
Weighted Avg. 0.874 0.218 0.872 0.874 0.871 0.891 0.937 0.938

*** Confusion Matrix ***
a b <-- classified as
660 40 | a = good
62 390 | b = bad

```

TYPE TO ENTER A CAPTION.

TYPE TO ENTER A CAPTION.

When we increase the confidence values or confidence factor(C) and the minimum number of objects(M) denoted by minNumObj the correctly classified instances increases. These are the parameters of the algorithm that have been varied.

When we took M = 1 and C = 0.55

correctly classified instances give 98.2%

incorrectly classified instances are 1.7778%

When we took M = 2 and C = 0.25

correctly classified instances give 69.66%

incorrectly classified instances are 30.33%

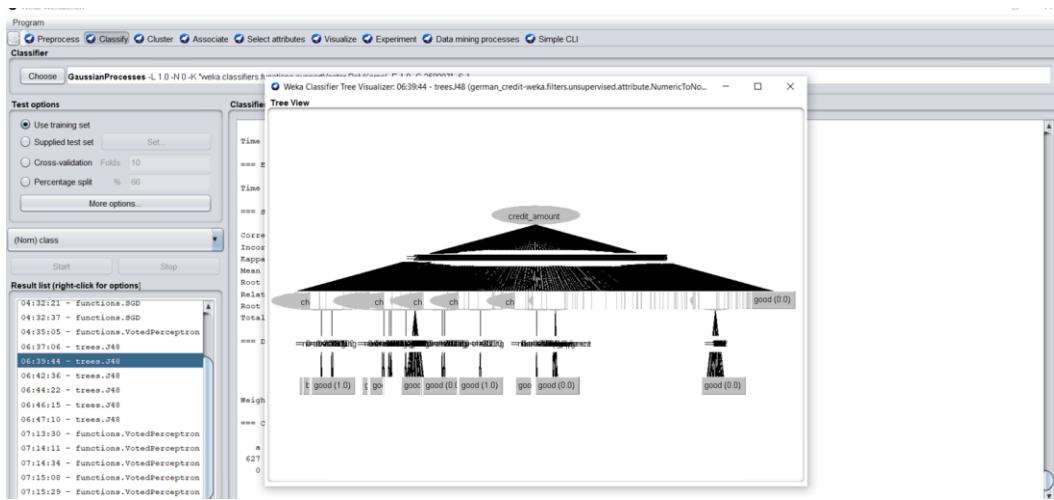
When we took M = 4 and C = 0.6

correctly classified instances give 87.4%

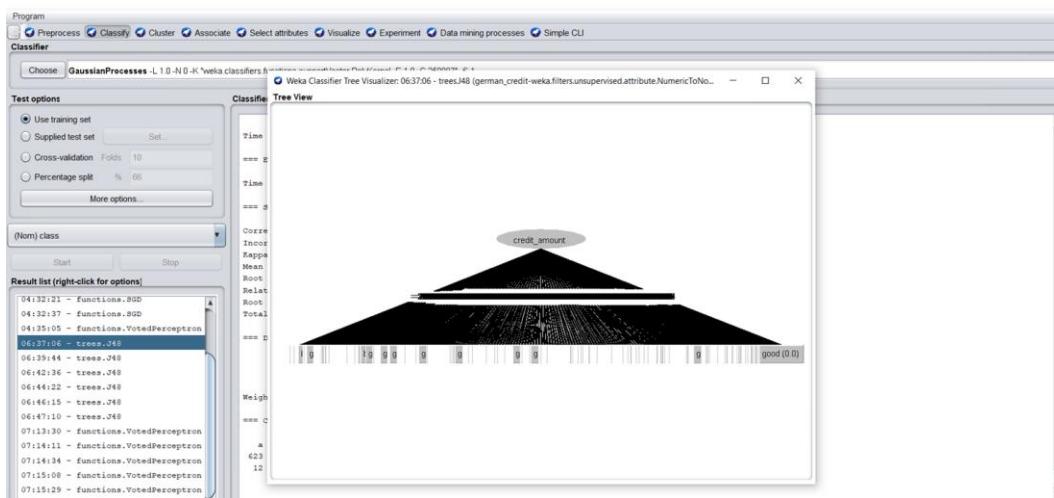
incorrectly classified instances are 12.6%

### **Visualisation of the training model**

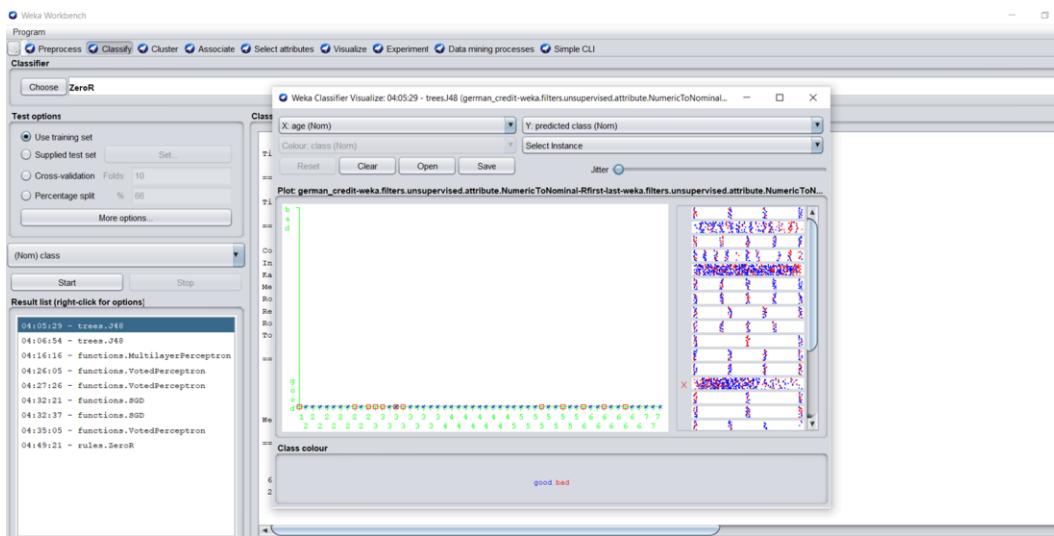
This helps to view relationship of different attributes in the model as well as the changes caused by tuning the parameters.



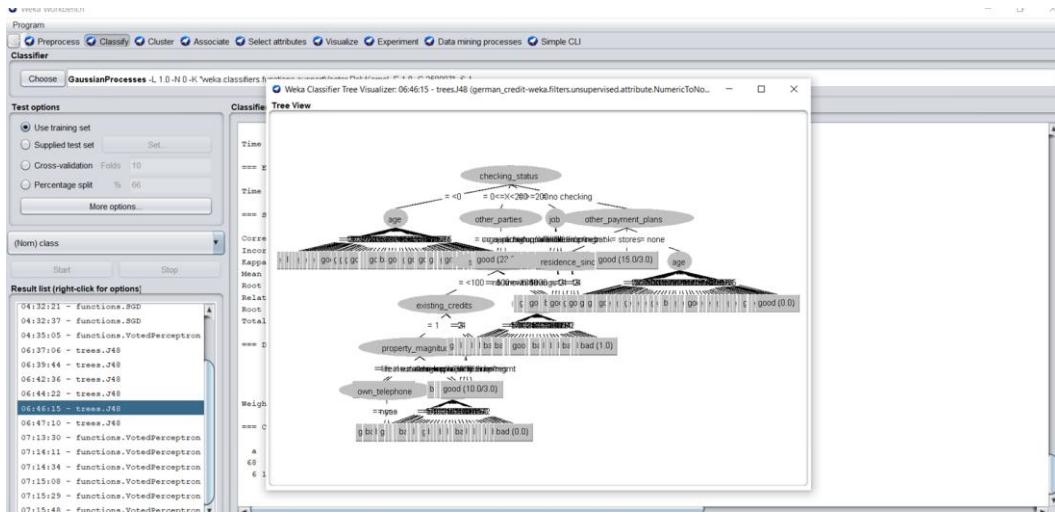
Train C=025,M=1



TRAIN:C = 0.55,M=1



**TRAIN C = 0.6, M=4**



**TRAIN C = 0.25, M=2**

## Testing the model

After training the dataset on 900 instances we test the dataset for given set of parameters.

For a C = 0.25 , M = 2 correctly classified instances are 75% and incorrectly classified instances are 25%.

Optimal performance is given when C = 0.75, M = 4 which gives classified instances are 98% and incorrectly classified instances are 2%.

For a C = 0.9 , M = 7 correctly classified instances are 87% and incorrectly classified instances are 13%.

For a C = 0.8 , M = 4 correctly classified instances are 75% and incorrectly classified instances are 25%.

```

48-C 0.75 -M 4

Classifier output

Time taken to build model: 0.34 seconds
==== Evaluation on test set ====
Time taken to test model on supplied test set: 0 seconds
==== Summary ====
Correctly Classified Instances      98      98      %
Incorrectly Classified Instances    2       2      %
Kappa statistic                     0.9452
Mean absolute error                 0.0261
Root mean squared error             0.1123
Relative absolute error              6.4961 %
Root relative squared error        25.7394 %
Total Number of Instances          100

==== Detailed Accuracy By Class ====
               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC   ROC Area  PRC Area  Class
      1.000   0.080   0.974   1.000   0.987   0.947   0.998   0.999   good
      0.920   0.000   1.000   0.920   0.958   0.947   0.998   0.993   bad
Weighted Avg.   0.980   0.060   0.981   0.980   0.947   0.998   0.998

==== Confusion Matrix ====
a b  <-- classified as
75 0 | a = good
2 23 | b = bad
  
```

### TEST C = 0.75, M =4

```

0.1-M 3

Classifier output

Time taken to build model: 0 seconds
==== Evaluation on test set ====
Time taken to test model on supplied test set: 0 seconds
==== Summary ====
Correctly Classified Instances      75      75      %
Incorrectly Classified Instances    25     25      %
Kappa statistic                     0.4017
Mean absolute error                 0.4955
Root mean squared error             0.4955
Relative absolute error              99.9457 %
Root relative squared error        99.9877 %
Total Number of Instances          100

==== Detailed Accuracy By Class ====
               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC   ROC Area  PRC Area  Class
      1.000   1.000   0.750   1.000   0.857   ?      0.500   0.750   good
      0.000   0.000   2       0.000   2       ?      0.500   0.250   bad
Weighted Avg.   0.750   0.750   2       0.750   2       ?      0.500   0.625

==== Confusion Matrix ====
a b  <-- classified as
75 0 | a = good
25 0 | b = bad
  
```

C = 0.9 , M = 7

C = 0.8 , M =4

```

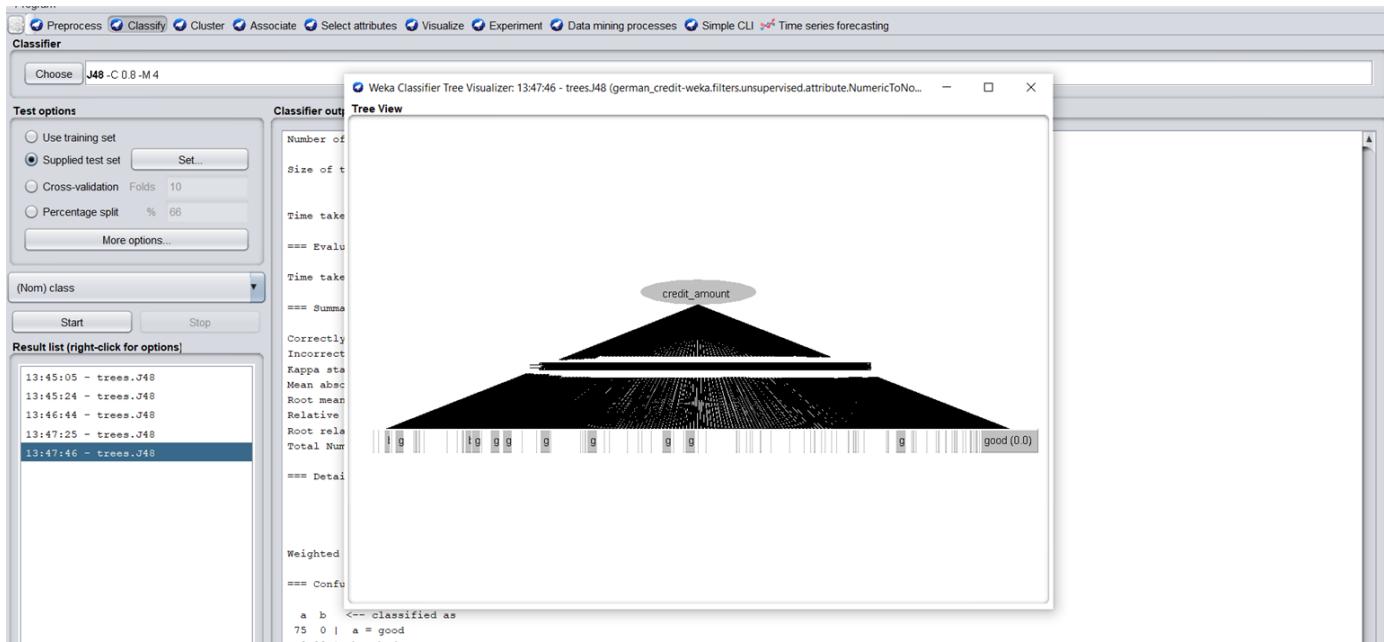
Classifier output

Time taken to build model: 0.35 seconds
==== Evaluation on test set ====
Time taken to test model on supplied test set: 0 seconds
==== Summary ====
Correctly Classified Instances      87      87      %
Incorrectly Classified Instances    13      13      %
Kappa statistic                     0.6579
Mean absolute error                 0.1989
Root mean squared error             0.3037
Relative absolute error              49.4799 %
Root relative squared error        69.6102 %
Total Number of Instances          100

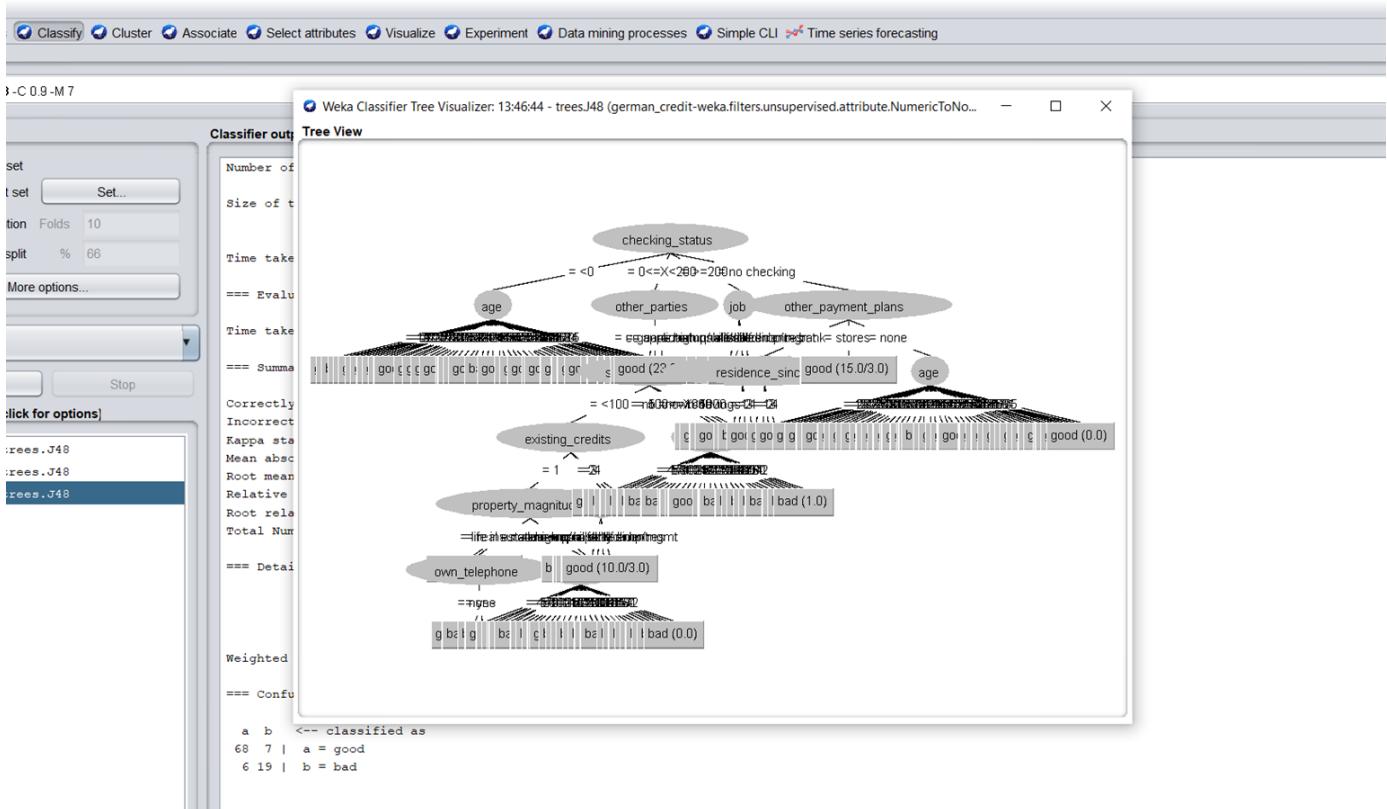
==== Detailed Accuracy By Class ====
               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC   ROC Area  PRC Area  Class
      0.907   0.240   0.919   0.907   0.913   0.658   0.928   0.574   good
      0.760   0.093   0.731   0.760   0.745   0.658   0.928   0.816   bad
Weighted Avg.   0.870   0.203   0.872   0.870   0.871   0.658   0.928   0.935

==== Confusion Matrix ====
a b  <-- classified as
87 13 | a = good
13 19 | b = bad
  
```

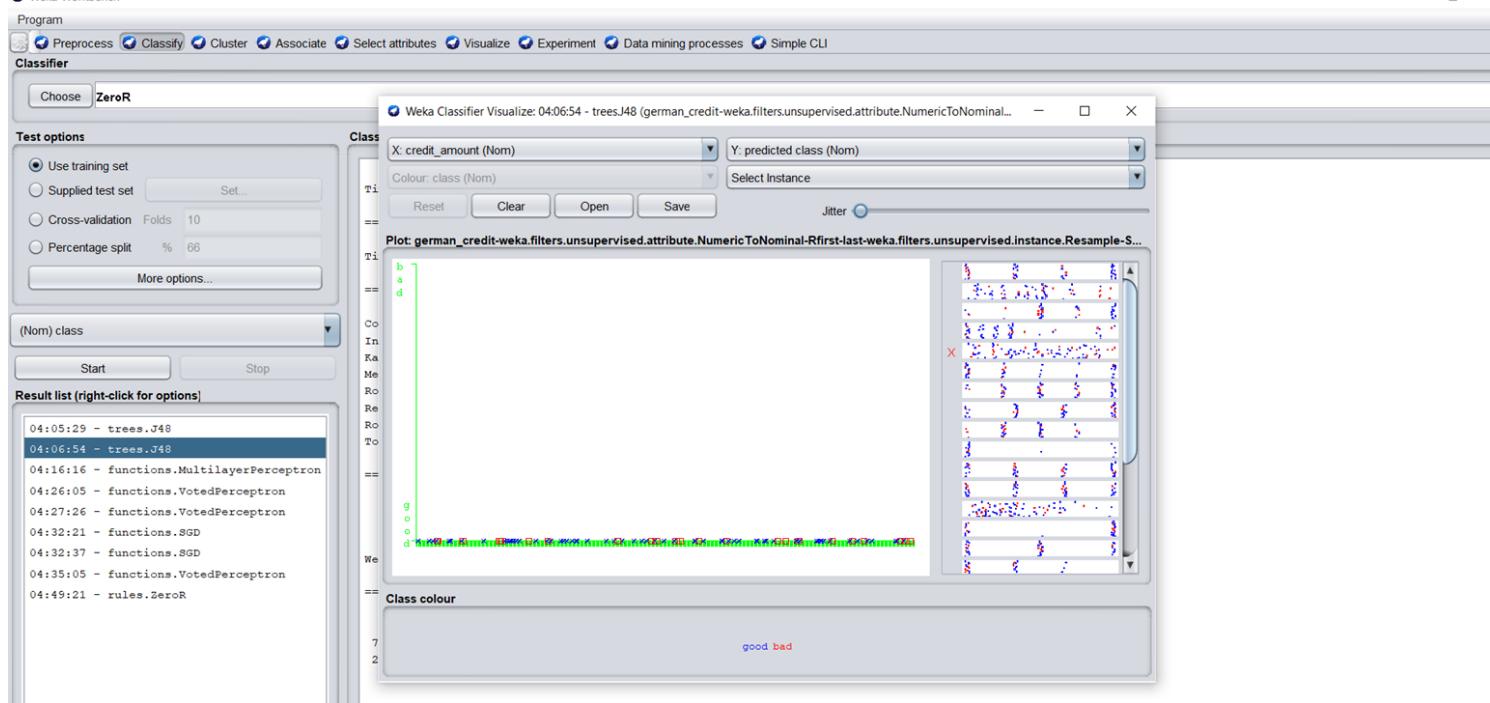
## Visualisation of the tested model



h



TYPE TO ENTER A CAPTION.



TYPE TO ENTER A CAPTION.

### Confusion matrices:

==== Confusion Matrix ===

| a   | b   | <-- classified as |
|-----|-----|-------------------|
| 660 | 40  | a = good          |
| 86  | 214 | b = bad           |

TYPE TO ENTER A CAPTION.

## ==== Confusion Matrix ====

| a  | b  | <-- classified as |
|----|----|-------------------|
| 75 | 0  | a = good          |
| 2  | 23 | b = bad           |

TYPE TO ENTER A CAPTION.

### URL link of resources

1)<https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/>

This link was helpful in understanding the different Machine Learning algorithms and draw comparisons with the algorithms used in this report.

2)[https://www.researchgate.net/figure/Explanation-of-WEKA-J48-parameters-Parameter-Description-Status\\_tbl2\\_258283784](https://www.researchgate.net/figure/Explanation-of-WEKA-J48-parameters-Parameter-Description-Status_tbl2_258283784)

This research paper was an analysis of J48 Tree and its usage in WEKA.

## 5. Classification: MLP or a similar advanced technique from Weka

Similar advanced technique used is voted perceptron.

Voted perceptron is based on the perceptron algorithm of Rousenblatt and Frank. The algorithm takes advantage of data that are linearly separable with large margins.

### Training of the model

In Voted Perceptron, the parameters are number of iterations(I) and batch size.

For the Classification model we chose an advanced technique called Voted Perceptron which comes under Perceptron Algorithm available in weka.

The optimal performance is given when  $I = 4$  and batch size = 500 correctly classified instances are 87.25% and incorrectly classified instances are 12.78%

When  $I = 2$  and batch size = 200 correctly classified instances are 80% and incorrectly classified instances are 20%

When  $I = 1$  and batch size = 40 correctly classified instances are 70% and incorrectly classified instances are 30%

```

Classifier
Choose VotedPerceptron-I2-E 10-S 1-M 10000-batch-size 200

Test options
 Use training set
 Supplied test set Set...
 Cross-validation Folds: 10
 Percentage split %: 66
More options...

(Nom) class Start Stop
Result list (right-click for options)
04:05:29 - trees.J48
04:06:14 - trees.J48
04:26:05 - functions.VotedPerceptron
04:27:16 - functions.VotedPerceptron
04:32:21 - functions.RDD
04:32:37 - functions.RDD
04:35:05 - functions.VotedPerceptron
04:37:06 - trees.J48
04:39:44 - trees.J48
04:42:36 - trees.J48
04:44:22 - trees.J48
04:46:15 - trees.J48
04:47:10 - trees.J48
07:13:30 - functions.VotedPerceptron
07:14:11 - functions.VotedPerceptron
07:14:34 - functions.VotedPerceptron
07:15:08 - functions.VotedPerceptron
07:15:29 - functions.VotedPerceptron

Classifier output
Time taken to build model: 0.6 seconds
*** Evaluation on test set ***
Time taken to test model on supplied test set: 0.07 seconds
*** Summary ***
Correctly Classified Instances      80      80 %
Incorrectly Classified Instances   20      20 %
Kappa statistic                   0.4805
Mean absolute error               0.2
Root mean squared error           0.4472
Relative absolute error            49.7655 %
Root relative squared error       102.4924 %
Total Number of Instances         100

*** Detailed Accuracy By Class ***
      TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
04:05:29  0.853  0.360  0.877  0.853  0.865  0.481  0.764  0.869  good
04:32:37  0.640  0.147  0.593  0.640  0.615  0.481  0.833  0.547  bad
Weighted Avg.  0.800  0.307  0.806  0.800  0.802  0.481  0.781  0.788

*** Confusion Matrix ***
      a   b  <-- classified as
64 11 | a = good
64 11 | b = bad

```

Train  $I = 2$ , batch-size = 200,  $I = 4$ , batch-size=400,  $I = 3$ , batch-size=400

```

Classifier
Choose VotedPerceptron-I4-E 10-S 1-M 10000-batch-size 500

Test options
 Use training set
 Supplied test set Set...
 Cross-validation Folds: 10
 Percentage split %: 66
More options...

(Nom) class Start Stop
Result list (right-click for options)
04:32:21 - functions.RDD
04:32:37 - functions.RDD
04:35:05 - functions.VotedPerceptron
04:37:06 - trees.J48
04:39:44 - trees.J48
04:42:36 - trees.J48
04:44:22 - trees.J48
04:46:15 - trees.J48
04:47:10 - trees.J48
07:13:30 - functions.VotedPerceptron
07:14:11 - functions.VotedPerceptron
07:14:34 - functions.VotedPerceptron
07:15:08 - functions.VotedPerceptron
07:15:29 - functions.VotedPerceptron

Classifier output
Time taken to build models: 2.21 seconds
*** Evaluation on training set ***
Time taken to test model on training data: 1.03 seconds
*** Summary ***
Correctly Classified Instances      785      87.2222 %
Incorrectly Classified Instances   115      12.7778 %
Kappa statistic                   0.687
Mean absolute error               0.128
Root mean squared error           0.3575
Relative absolute error            30.2833 %
Root relative squared error       77.7694 %
Total Number of Instances         900

*** Detailed Accuracy By Class ***
      TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
04:32:21  0.935  0.271  0.888  0.935  0.911  0.690  0.839  0.881  good
04:32:37  0.729  0.065  0.829  0.729  0.776  0.690  0.875  0.738  bad
Weighted Avg.  0.872  0.209  0.870  0.872  0.870  0.690  0.850  0.838

*** Confusion Matrix ***
      a   b  <-- classified as
586 41 | a = good
586 41 | b = bad

```

**TYPE TO ENTER A CAPTION.**

```

Classifier
Choose VotedPerceptron-I4-E 10-S 1-M 10000-batch-size 500

Test options
 Use training set
 Supplied test set Set...
 Cross-validation Folds: 10
 Percentage split %: 66
More options...

(Nom) class Start Stop
Result list (right-click for options)
04:32:21 - functions.RDD
04:32:37 - functions.RDD
04:35:05 - functions.VotedPerceptron
04:37:06 - trees.J48
04:39:44 - trees.J48
04:42:36 - trees.J48
04:44:22 - trees.J48
04:46:15 - trees.J48
04:47:10 - trees.J48
07:13:30 - functions.VotedPerceptron
07:14:11 - functions.VotedPerceptron
07:14:34 - functions.VotedPerceptron
07:15:08 - functions.VotedPerceptron
07:15:29 - functions.VotedPerceptron
07:15:48 - functions.VotedPerceptron

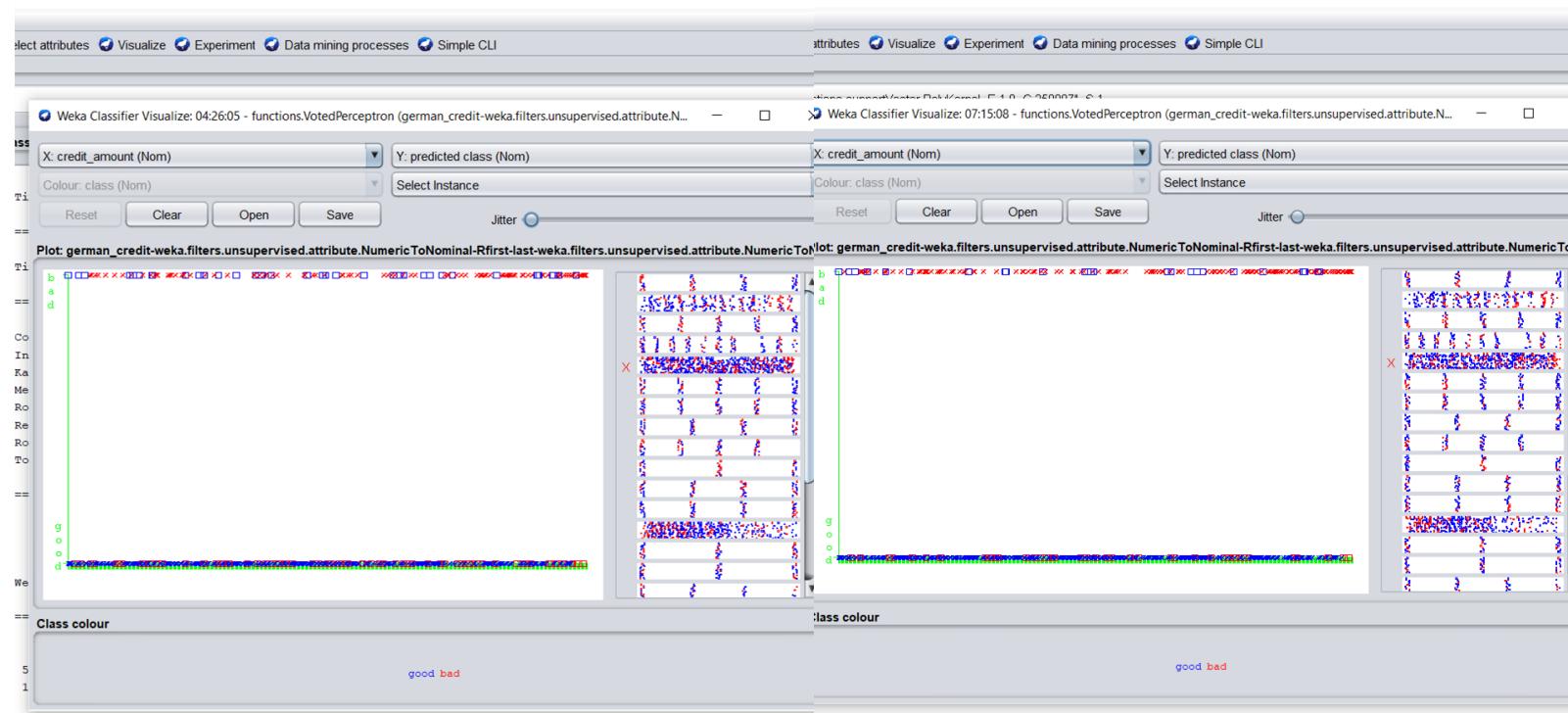
Classifier output
Time taken to build model: 2.21 seconds
*** Evaluation on training set ***
Time taken to test model on training data: 1.03 seconds
*** Summary ***
Correctly Classified Instances      785      87.2222 %
Incorrectly Classified Instances   115      12.7778 %
Kappa statistic                   0.687
Mean absolute error               0.128
Root mean squared error           0.3575
Relative absolute error            30.2833 %
Root relative squared error       77.7694 %
Total Number of Instances         900

*** Detailed Accuracy By Class ***
      TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
04:32:21  0.935  0.271  0.888  0.935  0.911  0.690  0.839  0.881  good
04:32:37  0.729  0.065  0.829  0.729  0.776  0.690  0.875  0.738  bad
Weighted Avg.  0.872  0.209  0.870  0.872  0.870  0.690  0.850  0.838

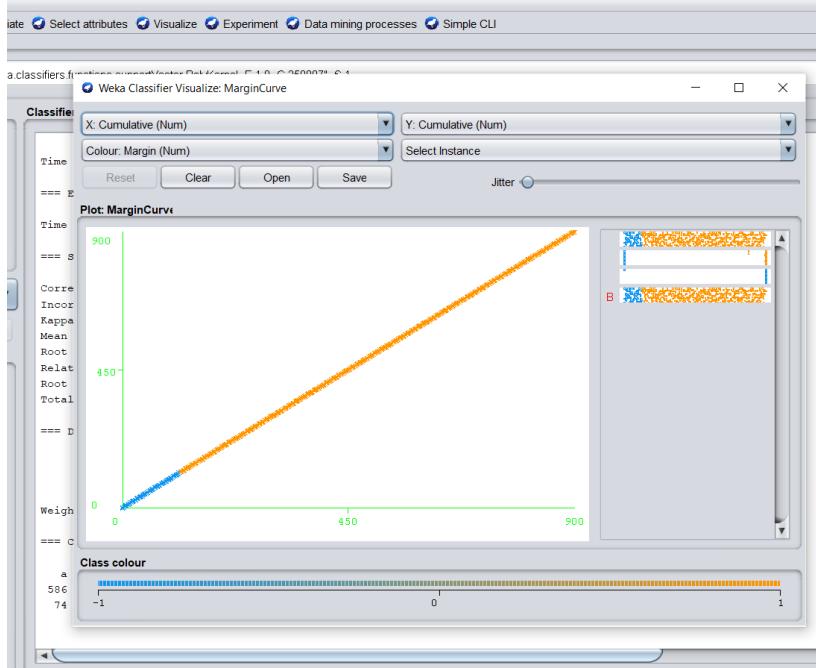
*** Confusion Matrix ***
      a   b  <-- classified as
586 41 | a = good
586 41 | b = bad

```

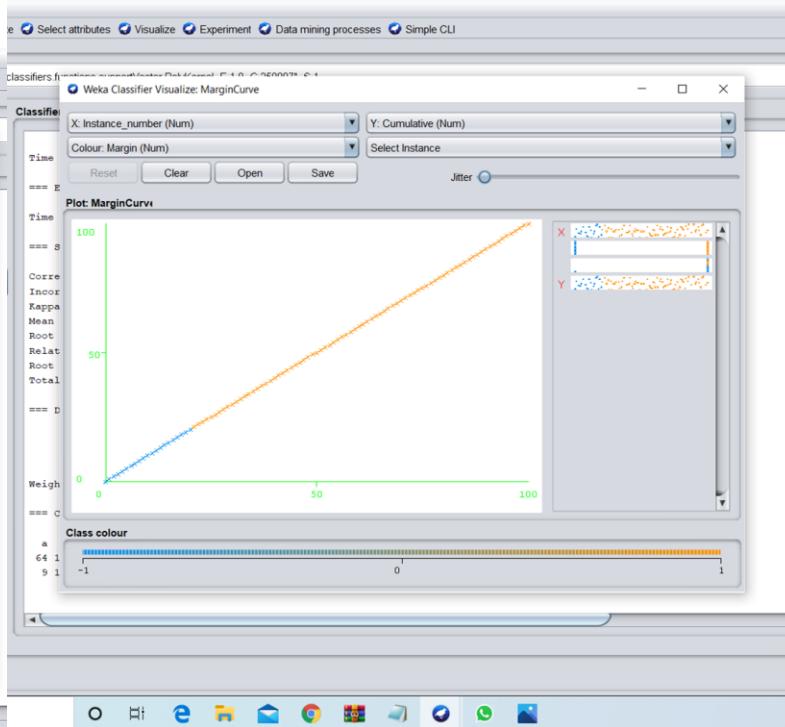
## Visualisation of training model



**TRAIN: I = 2,BATCH-SIZE=200**



TRAIN : I = 3, BATCH-SIZE = 400



TRAIN: I = 4, BATCH-SIZE = 400

## Testing the model

After training the model we test the model using the test data.

Optimal performance, when I = 5 and batch size = 400  
correctly classified instances are 86%  
incorrectly classified instances are 14%

When I = 2 and batch size = 400  
correctly classified instances are 81%  
incorrectly classified instances are 19%

When I = 4 and batch size = 300  
correctly classified instances are 85%  
incorrectly classified instances are 15%.

==== Run information ====

Scheme: weka.classifiers.functions.VotedPerceptron -I 1 -E 1.0 -S 1 -M 10000

Relation: german\_credit-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last-weka.filters.unsupervised.instance.Resample-S1-Z100.0-weka.filters.unsupervised.instance.Resample-S1-Z90.0-no-replacement

Instances: 100

Attributes: 21

- checking\_status
- duration
- credit\_history
- purpose
- credit\_amount
- savings\_status
- employment
- installment\_commitment
- personal\_status
- other\_parties
- residence\_since
- property\_magnitude
- age
- other\_payment\_plans
- housing
- existing\_credits
- job
- num\_dependents
- own\_telephone
- foreign\_worker

class

Test mode: evaluate on training data

==== Classifier model (full training set) ====

VotedPerceptron: Number of perceptrons=268

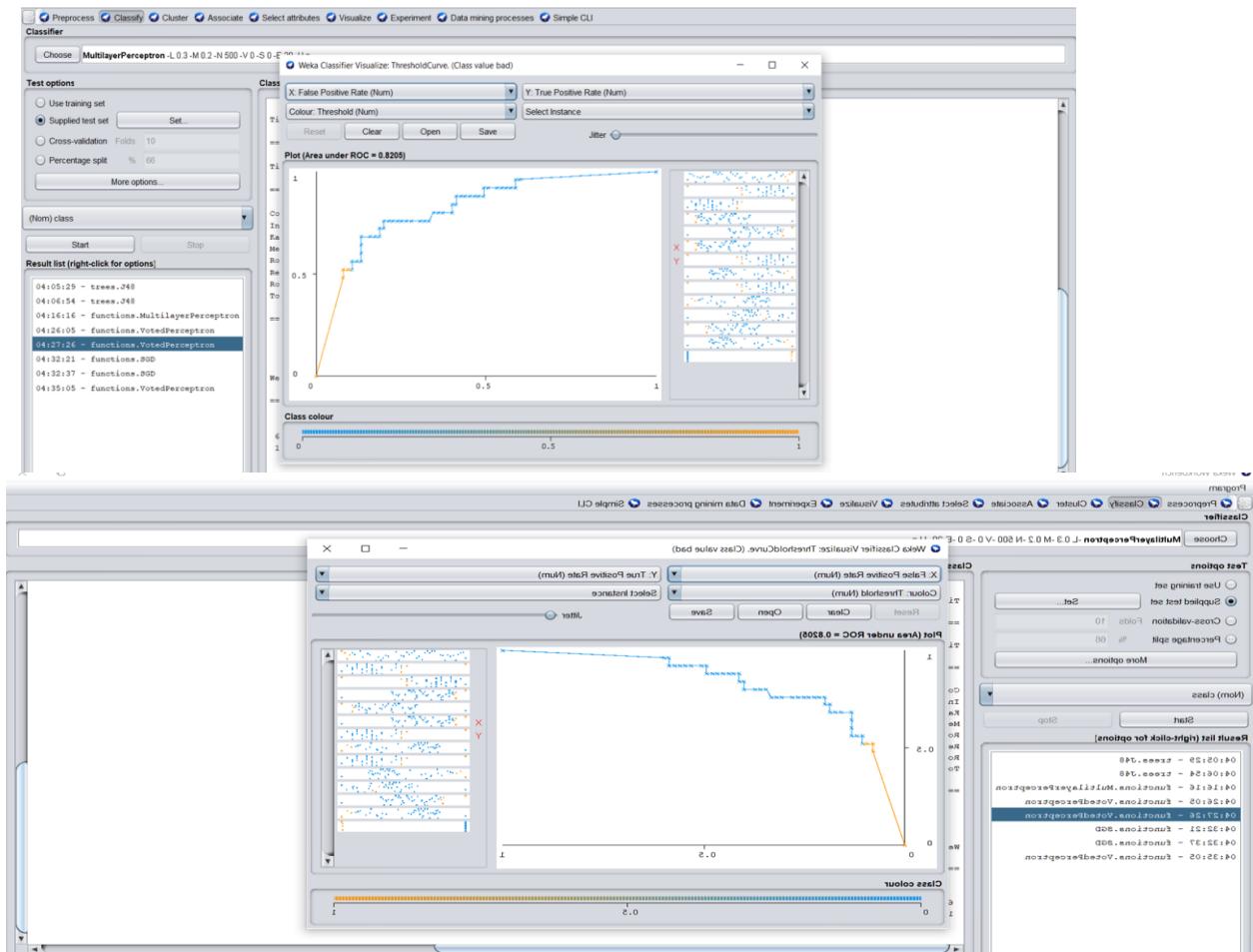
Time taken to build model: 0.24 seconds

## Visualisation of tested model

### Confusion matrices:-



$I = 5$  and batch size = 400



$I = 2$  and batch size = 400

```

==== Confusion Matrix ====
a   b   <-- classified as a   b   <-- classified as
586  41 |   a = good      67  8 |   a = good
74 199 |   b = bad       7 18 |   b = bad

```

TYPE TO ENTER A CAPTION.

TYPE TO ENTER A CAPTION.

### **URL links to online resources**

2)<https://en.wikipedia.org/wiki/Perceptron>

A reference to Voted Perceptron, an advanced technique from WEKA.

2)[https://www.researchgate.net/publication/289318021\\_Real-time\\_training\\_of\\_Voted\\_Perceptron\\_for\\_classification\\_of\\_EEG\\_data](https://www.researchgate.net/publication/289318021_Real-time_training_of_Voted_Perceptron_for_classification_of_EEG_data)

This link is to a research paper based on Voted Perceptron that was used to gain better insight in the topic.

## **6. Clustering: K-Means or DBSCAN**

Clustering technique used for this dataset is K-Means Algorithm. K-Means Algorithm is an example of unsupervised learning. It starts with a group of selected centroids which are used to form clusters of points and finally we perform the iterations.

The parameters that are being varied to produce different outcomes are Maximum number of iterations(maxIterations) and number of clusters (numClusters).

When maxIterations = 500 and numClusters = 2

then the clustered instances are divided into 65% and 30%

When maxIterations = 600 and numClusters = 3 then the clustered instances are divided into 54%, 29% and 18%

When maxIterations = 400 and numClusters = 3 then the clustered instances are divided into 47%, 40% and 13%

I tried several configurations for k-means clustering. Changing the distance function between euclidean and Manhattan did not change clusters very much. The initialization method had a very big effect on cluster formation. When choosing random, resulting 2 clusters did not correspond appropriate clusters, but choosing farthest first, k-means++ or canopy did give the same clusters corresponding class with very high accuracy. The following results are produced.

The screenshot shows the Weka Workbench interface with the SimpleKMeans classifier selected. The command line at the top is:

```
Choose SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.ManhattanDistance" -R first-last -I 500 -num-slots 1 -S 10
```

The Cluster mode section shows "Use training set" selected. The Cluster output window displays the following data:

|                        | cluster 1     | cluster 2                    | cluster 3       | cluster 4       |
|------------------------|---------------|------------------------------|-----------------|-----------------|
| checking_status        | no checking   | no checking                  | no checking     | no checking     |
| duration               | 24            | 12                           | 24              | 24              |
| credit_history         | existing paid | existing paid critical/other | existing credit | existing credit |
| purpose                | radio/tv      | radio/tv                     | new car         | new car         |
| credit_amount          | 1386          | 1386                         | 1169            | 1169            |
| savings_status         | <100          | <100                         | <100            | <100            |
| employment             | 1<=X<4        | 1<=X<4                       | >=7             | >=7             |
| installment_commitment | 4             | 4                            | 4               | 4               |
| personal_status        | male single   | male single                  | male single     | male single     |
| other_parties          | none          | none                         | none            | none            |
| residence_since        | 4             | 4                            | 4               | 4               |
| property_magnitude     | car           | car                          | life insurance  | 36              |
| age                    | 26            | 26                           | none            | none            |
| other_payment_plans    | none          | none                         | yes             | yes             |
| housing                | own           | own                          | own             | own             |
| existing_credits       | 1             | 1                            | 2               | 2               |
| job                    | skilled       | skilled                      | skilled         | skilled         |
| num_dependents         | 1             | 1                            | 1               | 1               |
| own_telephone          | none          | none                         | yes             | yes             |
| foreign_worker         | yes           | yes                          | yes             | yes             |
| class                  | good          | good                         | good            | good            |

Time taken to build model (full training data) : 0.03 seconds  
 === Model and evaluation on training set ===  
 Clustered Instances  
 0 633 ( 70%)  
 1 267 ( 30%)

## Manhattan and Euclidean distance

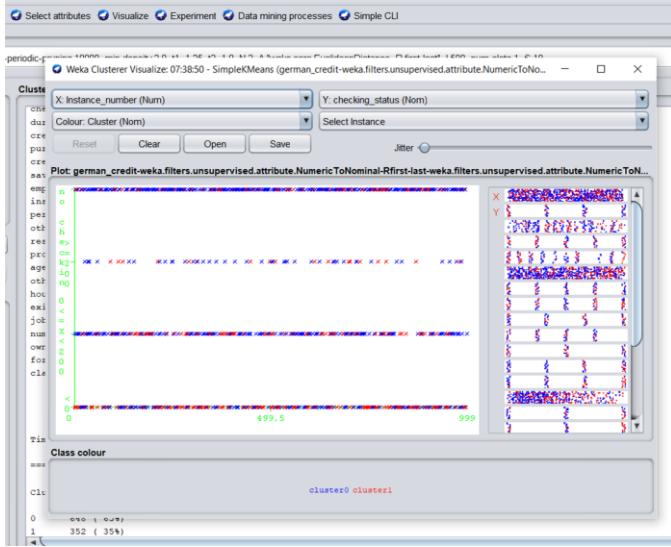
The screenshot shows the Weka Workbench interface with the SimpleKMeans classifier selected. The command line at the top is:

```
Choose SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance" -R first-last -I 500 -num-slots 1 -S 10
```

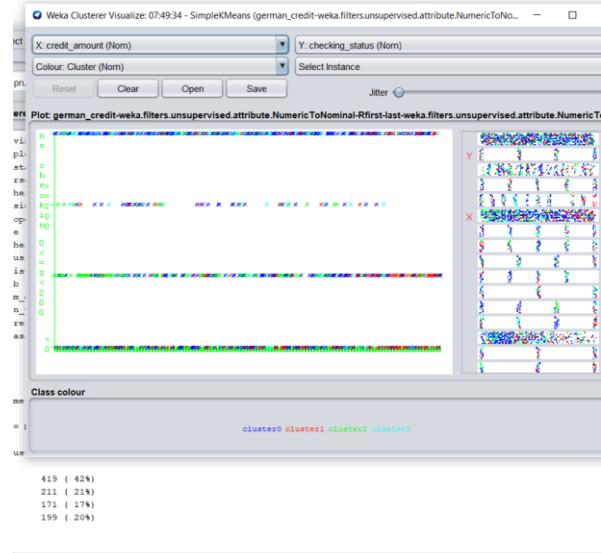
The Cluster mode section shows "Use training set" selected. The Cluster output window displays the following data:

|                        | cluster 1     | cluster 2                    | cluster 3       | cluster 4       |
|------------------------|---------------|------------------------------|-----------------|-----------------|
| checking_status        | no checking   | no checking                  | no checking     | no checking     |
| duration               | 24            | 12                           | 24              | 24              |
| credit_history         | existing paid | existing paid critical/other | existing credit | existing credit |
| purpose                | radio/tv      | radio/tv                     | new car         | new car         |
| credit_amount          | 1386          | 1386                         | 1169            | 1169            |
| savings_status         | <100          | <100                         | <100            | <100            |
| employment             | 1<=X<4        | 1<=X<4                       | >=7             | >=7             |
| installment_commitment | 4             | 4                            | 4               | 4               |
| personal_status        | male single   | male single                  | male single     | male single     |
| other_parties          | none          | none                         | none            | none            |
| residence_since        | 4             | 4                            | 4               | 4               |
| property_magnitude     | car           | car                          | life insurance  | 36              |
| age                    | 26            | 26                           | none            | none            |
| other_payment_plans    | none          | none                         | own             | own             |
| housing                | own           | own                          | 1               | 2               |
| existing_credits       | 1             | 1                            | 1               | 1               |
| job                    | skilled       | skilled                      | skilled         | skilled         |
| num_dependents         | 1             | 1                            | 1               | 1               |
| own_telephone          | none          | none                         | yes             | yes             |
| foreign_worker         | yes           | yes                          | yes             | yes             |
| class                  | good          | good                         | good            | good            |

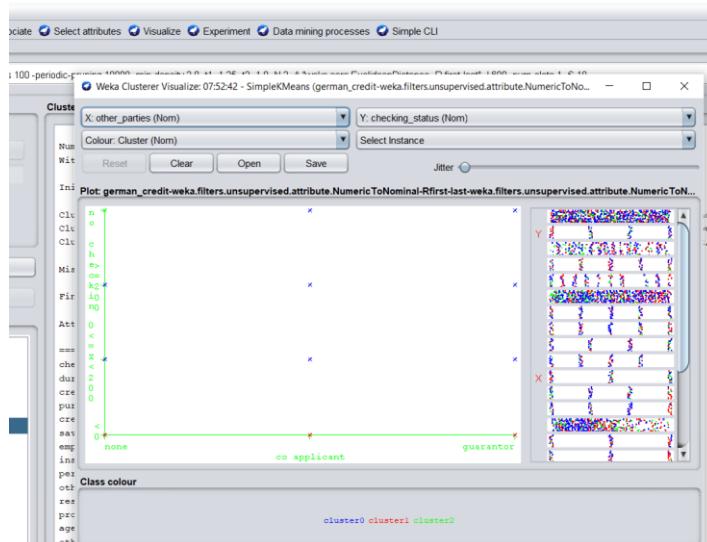
Time taken to build model (full training data) : 0.01 seconds  
 === Model and evaluation on training set ===  
 Clustered Instances  
 0 633 ( 70%)  
 1 267 ( 30%)



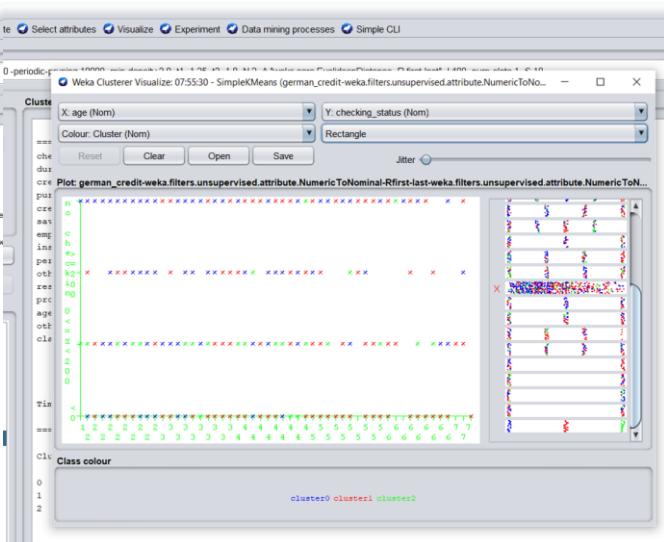
maxIterations = 500 and numClusters = 2



maxIterations = 400 and numClusters = 3



maxIterations = 600 and numClusters = 3



The screen shots given below consist of following values of parameters:

1) maxIterations = 500 and numClusters = 2

2) maxIterations = 400 and numClusters = 3    3) maxIterations = 600 and numClusters = 3

Choose SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance" -R first-last" -I 500 -num-slots 1 -S 10

**Cluster mode**

- Use training set
- Supplied test set Set...
- Percentage split % 66
- Classes to clusters evaluation (Nom) class
- Store clusters for visualization

Ignore attributes Start Stop

**Result list (right-click for options)**

- 07:38:50 - SimpleKMeans

**Clusterer output**

|                        | no checking   | no checking   | <0                |
|------------------------|---------------|---------------|-------------------|
| duration               | 24            | 12            | 24                |
| credit_history         | existing paid | existing paid | existing paid     |
| purpose                | radio/tv      | radio/tv      | new car           |
| credit_amount          | 1258          | 1258          | 727               |
| savings_status         | <100          | <100          | <100              |
| employment             | 1=<X<4        | 1=<X<4        | >=7               |
| installment_commitment | 4             | 4             | 4                 |
| personal_status        | male single   | male single   | male single       |
| other_parties          | none          | none          | none              |
| residence_since        | 4             | 2             | 4                 |
| property_magnitude     | car           | real estate   | no known property |
| age                    | 27            | 27            | 36                |
| other_payment_plans    | none          | none          | none              |
| housing                | own           | own           | own               |
| existing_credits       | 1             | 1             | 1                 |
| job                    | skilled       | skilled       | skilled           |
| num_dependents         | 1             | 1             | 1                 |
| own_telephone          | none          | none          | yes               |
| foreign_worker         | yes           | yes           | yes               |
| class                  | good          | good          | good              |

Time taken to build model (full training data) : 0.03 seconds  
 === Model and evaluation on training set ===

Choose SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance" -R first-last" -I 1600 -num-slots 1 -S 10

**Cluster mode**

- Use training set
- Supplied test set Set...
- Percentage split % 66
- Classes to clusters evaluation (Nom) class
- Store clusters for visualization

Ignore attributes Start Stop

**Result list (right-click for options)**

- 07:38:50 - SimpleKMeans
- 07:47:36 - SimpleKMeans
- 07:49:34 - SimpleKMeans
- 07:51:26 - SimpleKMeans
- 07:52:04 - SimpleKMeans
- 07:52:42 - SimpleKMeans

**Clusterer output**

|                        | 1258        | 1393        | 727               | 1154        |
|------------------------|-------------|-------------|-------------------|-------------|
| credit_amount          |             |             |                   |             |
| savings_status         | <100        | <100        | <100              | <100        |
| employment             | 1=<X<4      | 1=<X<4      | >=7               | >=7         |
| installment_commitment | 4           | 4           | 4                 | 4           |
| personal_status        | male single | male single | male single       | male single |
| other_parties          | none        | none        | none              | none        |
| residence_since        | 4           | 2           | 4                 | 4           |
| property_magnitude     | car         | car         | no known property | real estate |
| age                    | 27          | 27          | 36                | 23          |
| other_payment_plans    | none        | none        | none              | none        |
| housing                | own         | own         | own               | own         |
| existing_credits       | 1           | 1           | 1                 | 1           |
| job                    | skilled     | skilled     | skilled           | unskilled   |
| num_dependents         | 1           | 1           | 1                 | 1           |
| own_telephone          | none        | none        | yes               | none        |
| foreign_worker         | yes         | yes         | yes               | yes         |
| class                  | good        | good        | good              | good        |

Time taken to build model (full training data) : 0.01 seconds  
 === Model and evaluation on training set ===

**Clustered Instances**

|   | 0          | 1          | 2 |
|---|------------|------------|---|
| 0 | 537 ( 54%) |            |   |
| 1 |            | 287 ( 29%) |   |
| 2 |            | 176 ( 18%) |   |

TYPE TO ENTER A CAPTION

## Additional output of the clustering process

For the above results produced

Choose SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 1.25 -t2 1.0 -N 4 -A "weka.core.EuclideanDistance" -R first-last" -I 500 -num-slots 1 -S 10

**Cluster mode**

- Use training set
- Supplied test set Set...
- Percentage split % 66
- Classes to clusters evaluation (Nom) class
- Store clusters for visualization

Ignore attributes Start Stop

**Result list (right-click for options)**

- 07:38:50 - SimpleKMeans
- 07:47:36 - SimpleKMeans
- 07:49:34 - SimpleKMeans

**Clusterer output**

Missing values globally replaced with mean/mode

Final cluster centroids:

| Attribute              | Full Data (1000.0) | Cluster# 0 (419.0) | 1 (211.0)                 | 2 (171.0)          | (15)       |
|------------------------|--------------------|--------------------|---------------------------|--------------------|------------|
| checking_status        | no checking        | no checking        | <0                        | 0=<X<200           | no chec    |
| duration               | 24                 | 24                 | 24                        | 12                 |            |
| credit_history         | existing paid      | existing paid      | existing paid             | existing paid      |            |
| purpose                | radio/tv           | new car            | used car                  | radio/tv           |            |
| credit_amount          | 1258               | 1275               | 1199                      | 709                |            |
| savings_status         | <100               | <100               | <100                      | <100               |            |
| employment             | 1=<X<4             | 1=<X<4             | >7                        | <1                 |            |
| installment_commitment | 4                  | 4                  | 4                         | 4                  |            |
| personal_status        | male single        | female div/dep/mar | male single               | male single        | male si    |
| other_parties          | none               | none               | none                      | none               |            |
| residence_since        | 4                  | 2                  | 4                         | 3                  |            |
| property_magnitude     | car                | car                | no known property         | real estate        | life insur |
| age                    | 27                 | 25                 | 35                        | 23                 |            |
| other_payment_plans    | none               | none               | none                      | none               |            |
| housing                | own                | own                | own                       | own                |            |
| existing_credits       | 1                  | 1                  | 1                         | 1                  |            |
| job                    | skilled            | skilled            | high qualif/self emp/mgmt | unskilled resident | ski        |
| num_dependents         | 1                  | 1                  | 1                         | 1                  |            |
| own_telephone          | none               | none               | yes                       | none               |            |

Preprocess Classify Cluster Associate Select attributes Visualize Experiment Data mining processes Simple CLI

Choose SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 1.25 -t2 1.0 -N 3 -A "weka.core.EuclideanDistance" -R first-last" -I 400 -num-slots 1 -S 10

**Cluster mode**

- Use training set
- Supplied test set Set...
- Percentage split % 66
- Classes to clusters evaluation (Nom) class
- Store clusters for visualization

Ignore attributes Start Stop

**Result list (right-click for options)**

- 07:38:50 - SimpleKMeans
- 07:47:36 - SimpleKMeans
- 07:49:34 - SimpleKMeans
- 07:51:26 - SimpleKMeans
- 07:52:04 - SimpleKMeans
- 07:52:42 - SimpleKMeans
- 07:55:10 - SimpleKMeans
- 07:55:30 - SimpleKMeans

**Clusterer output**

(1000.0) (472.0) (399.0) (129.0)

|                        |               |                    |               |               |
|------------------------|---------------|--------------------|---------------|---------------|
| checking_status        | no checking   | no checking        | <0            | 0=<X<200      |
| duration               | 24            | 12                 | 24            | 12            |
| credit_history         | existing paid | existing paid      | existing paid | existing paid |
| purpose                | radio/tv      | new car            | radio/tv      | radio/tv      |
| credit_amount          | 1258          | 1258               | 717           | 1424          |
| savings_status         | <100          | <100               | <100          | <100          |
| employment             | 1=<X<4        | 1=<X<4             | >7            | <1            |
| installment_commitment | 4             | 4                  | 4             | 4             |
| personal_status        | male single   | female div/dep/mar | male single   | male mar/wid  |
| other_parties          | none          | none               | none          | none          |
| residence_since        | 4             | 2                  | 4             | 3             |
| property_magnitude     | car           | car                | car           | real estate   |
| age                    | 27            | 27                 | 36            | 26            |
| other_payment_plans    | none          | none               | none          | none          |
| class                  | good          | good               | good          | good          |

Time taken to build model (full training data) : 0.01 seconds

Model and evaluation on training set

Clustered Instances

|   |            |
|---|------------|
| 0 | 472 ( 47%) |
| 1 | 399 ( 40%) |
| 2 | 129 ( 13%) |

Choose SimpleKMeans -init 0 -max-candidates 1000 -periodic-pruning 10000 -min-density 2.0 -t1 1.25 -t2 1.0 -N 3 -A "weka.core.EuclideanDistance" -R first-last" -I 400 -num-slots 1 -S 10

Type to enter a caption

**Cluster mode**

- Use training set
- Supplied test set Set...
- Percentage split % 66
- Classes to clusters evaluation (Nom) class
- Store clusters for visualization

Ignore attributes Start Stop

**Result list (right-click for options)**

- 07:38:50 - SimpleKMeans
- 07:47:36 - SimpleKMeans
- 07:49:34 - SimpleKMeans
- 07:51:26 - SimpleKMeans
- 07:52:04 - SimpleKMeans
- 07:52:42 - SimpleKMeans
- 07:55:10 - SimpleKMeans
- 07:55:30 - SimpleKMeans

**Clusterer output**

Number of iterations: 4  
Within cluster sum of squared errors: 7679.0

Initial starting points (random):

Cluster 0: 'no checking',36,'critical/other existing credit','new car',7855,<100,1=<X<4,4,'female div/dep/max',none,2,'real estate',25,stores,bad  
Cluster 1: <0,24,'critical/other existing credit','used car',6615,<100,unemployed,2,'male single',none,4,'no known property',75,none,good  
Cluster 2: 0=<X<200,12,'existing paid',radio/tv,1155,<100,>7,3,'male mar/wid',guarantor,3,'real estate',40,bank,good

Missing values globally replaced with mean/mode

Final cluster centroids:

| Attribute              | Full Data (1000.0) | Cluster# 0 (472.0) | 1 (399.0)     | 2 (129.0)     |
|------------------------|--------------------|--------------------|---------------|---------------|
| checking_status        | no checking        | no checking        | <0            | 0=<X<200      |
| duration               | 24                 | 12                 | 24            | 12            |
| credit_history         | existing paid      | existing paid      | existing paid | existing paid |
| purpose                | radio/tv           | new car            | radio/tv      | radio/tv      |
| credit_amount          | 1258               | 1258               | 717           | 1424          |
| savings_status         | <100               | <100               | <100          | <100          |
| employment             | 1=<X<4             | 1=<X<4             | >7            | <1            |
| installment_commitment | 4                  | 4                  | 4             | 4             |
| personal_status        | male single        | female div/dep/max | male single   | male mar/wid  |

Url links your online resources

- 1) <http://facweb.cs.depaul.edu/mobasher/classes/ect584/WEKA/k-means.html>

This link references to K-Means Clustering in WEKA. Illustrations have been used to explain the concept in depth

- 2) <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>

This article is an explanation of the K-Means Clustering Algorithm with examples of the same.

## Evaluate the clusters using the “classes to clusters evaluation

| Attribute              | Full Data     | Clusters#          |                   |               |
|------------------------|---------------|--------------------|-------------------|---------------|
| checking_status        | <0            | 0 (419.0)          | 1 (211.0)         | 2 (171.0)     |
| duration               | 24            | 24                 | 12                | 12            |
| credit_history         | existing paid | existing paid      | existing paid     | existing paid |
| purpose                | radio/tv      | new car            | used car          | radio/tv      |
| credit_amount          | 1258          | 1275               | 1199              | 709           |
| savings_status         | <100          | <100               | <100              | <100          |
| employment             | 1<=4          | 1<=4               | >=5               | <1            |
| installment_commitment | 4             | 4                  | 4                 | 4             |
| personal_status        | male single   | female div/dep/mar | male single       | male single   |
| other_parties          | none          | none               | none              | none          |
| residence_since        | 1             | 2                  | 4                 | 2             |
| property_magnitude     | car           | car                | no known property | real estate   |
| age                    | 27            | 25                 | 35                | 23            |
| other_payment_plans    | none          | none               | none              | none          |
| duration_in_month      | none          | none               | none              | none          |

maxIterations = 100 and numClusters = 2

| Attribute           | Cluster 0 | Cluster 1         |
|---------------------|-----------|-------------------|
| property_magnitude  | car       | real estate       |
| other_payment_plans | none      | no known property |
| housing             | own       | own               |
| existing_credits    | 1         | 1                 |
| job                 | skilled   | skilled           |

maxIterations = 600 and numClusters = 2

TYPE TO ENTER A CAPTION.

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning  
10000 -min-density 2.0 -t1 1.25 -t2 1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I  
500 -num-slots 1 -S 10

Relation: german\_credit-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last-  
weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last-  
weka.filters.unsupervised.instance.Resample-S1-Z100.0-  
weka.filters.unsupervised.instance.Resample-S1-Z90.0-no-replacement

Instances: 900

Attributes: 21

checking\_status  
duration  
credit\_history  
purpose  
credit\_amount  
savings\_status  
employment  
installment\_commitment  
personal\_status  
other\_parties  
residence\_since  
property\_magnitude  
age  
other\_payment\_plans  
housing  
existing\_credits  
job

num\_dependents  
own\_telephone  
foreign\_worker

Ignored:

class

Test mode: Classes to clusters evaluation on training data

==== Clustering model (full training set) ===

kMeans

=====

Number of iterations: 4

Within cluster sum of squared errors: 8123.0

Initial starting points (random):

Cluster 0: <0,12,'existing paid',education,684,<100,1<=X<4,4,'male single',none,4,car,40,none,rent,1,'unskilled resident',2,none,yes

Cluster 1: 0<=X<200,18,'critical/other existing credit',furniture/equipment,7374,<100,unemployed,4,'male single',none,4,'life insurance',40,stores,own,2,'high qualif/self emp/mgmt',1,yes,yes

Missing values globally replaced with mean/mode

Final cluster centroids:

| Attribute                   | Full Data     | Cluster#                     |                |
|-----------------------------|---------------|------------------------------|----------------|
|                             |               | 0                            | 1              |
|                             | (900.0)       | (630.0)                      | (270.0)        |
| <hr/>                       |               |                              |                |
| <hr/>                       |               |                              |                |
| checking_status<br>checking | no checking   | no checking                  | no             |
| duration                    | 24            | 12                           | 24             |
| credit_history<br>credit    | existing paid | existing paid critical/other | existing       |
| purpose                     | radio/tv      | radio/tv                     | new car        |
| credit_amount               | 1386          | 1386                         | 1169           |
| savings_status              | <100          | <100                         | <100           |
| employment                  | 1<=X<4        | 1<=X<4                       | >=7            |
| installment_commitment      | 4             | 4                            | 4              |
| personal_status<br>single   | male single   | male single                  | male           |
| other_parties               | none          | none                         | none           |
| residence_since             | 4             | 4                            | 4              |
| property_magnitude          | car           | real estate                  | life insurance |
| age                         | 26            | 26                           | 36             |
| other_payment_plans         | none          | none                         | none           |
| housing                     | own           | own                          | own            |
| existing_credits            | 1             | 1                            | 2              |
| job                         | skilled       | skilled                      | skilled        |
| num_dependents              | 1             | 1                            | 1              |

|                |      |      |     |
|----------------|------|------|-----|
| own_telephone  | none | none | yes |
| foreign_worker | yes  | yes  | yes |

## 7.TimeSeries Forecasting

In Time Series Forecasting prediction about the future is done using a method called Extrapolation using time series data. We are basically analysing the existing trends and predicting if the current scenario will continue in the future.

In this dataset the Time Series Forecasting uses 7 attributes:-

- 1) Duration
- 2) Credit\_Amount
- 3) Installment\_Commitment

4) Residence\_Since

5) Age

6) Existing\_Credits

7) Num\_dependents

Different regression methods are used for time series forecasting

1) Linear Regression

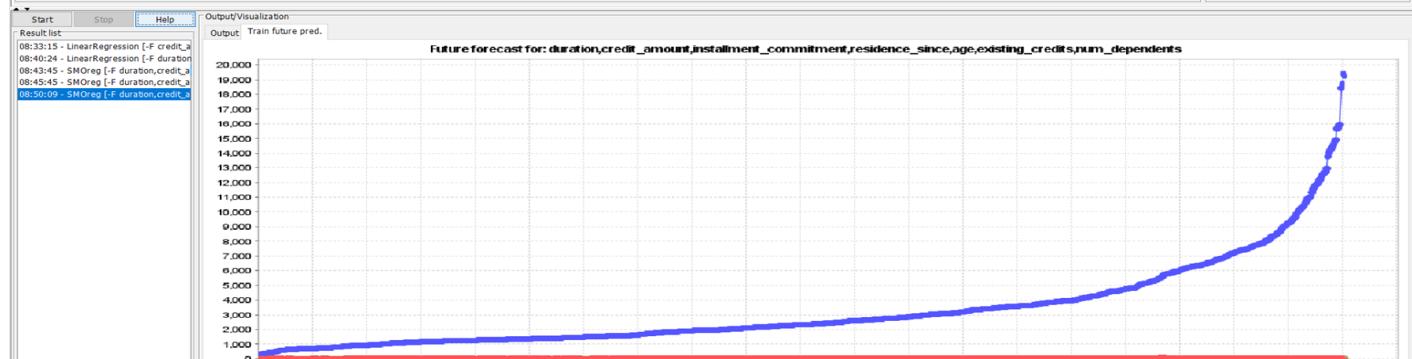
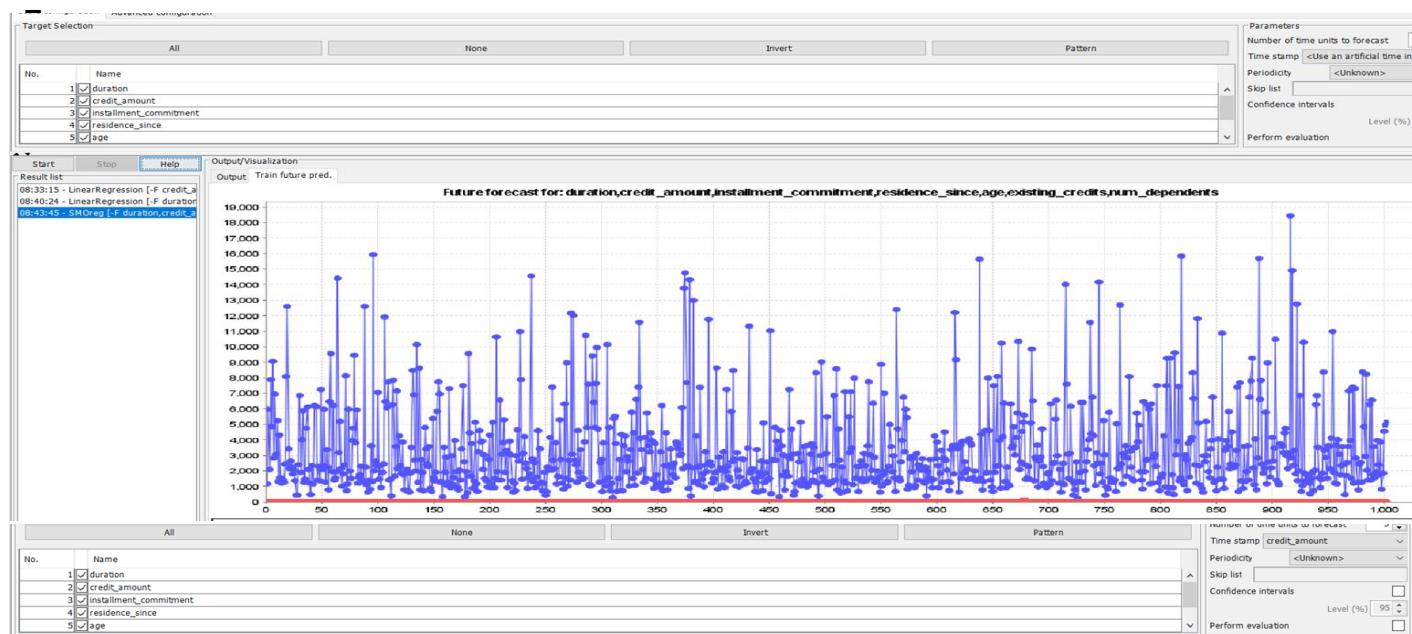
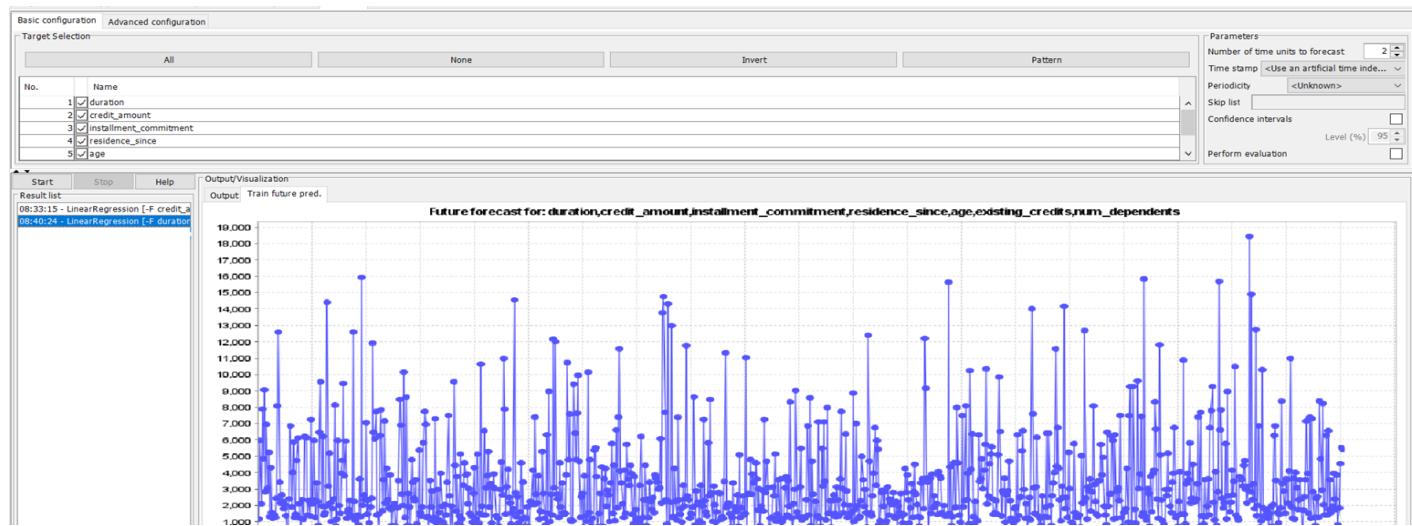
2) SMOreg

3) Gaussian Process

4) Additive Regression

Only the attribute Credit\_Amount can be used to predict a foul transaction.

## Diagrams and Visualization historical Values and predictions:1)Linear Regression



## 2)SMOREG

## Results

The screenshot displays three vertically stacked software interfaces for time series analysis, likely from the Frama framework. Each interface shows a list of results, configuration options, and a detailed output section.

**Top Interface:**

- Result list:** Shows five selected models: 08:33:15 - LinearRegression [-F credit\_a] and 08:40:24 - LinearRegression [-F duration].
- Configuration:** Set to "Train future pred." with a time stamp of <Unknown>, periodicity of <Unknown>, and confidence intervals at 95%.
- Output:** Displays a table of predicted values for various rows (976 to 997) across columns including duration, credit\_amount, installment\_commitment, residence\_since, age, and predicted values (57, 64, 42, 47, 25, 49, 33, 28, 26, 30, 25, 33, 64, 29, 48, 37, 34, 23, 30, 50, 31, 40).

**Middle Interface:**

- Result list:** Shows five selected models: 08:33:15 - LinearRegression [-F credit\_a] and 08:39:41 - LinearRegression [-F credit\_a].
- Configuration:** Set to "Train future pred." with a time stamp of <Unknown>, periodicity of <Unknown>, and confidence intervals at 95%.
- Output:** Displays a table of predicted values for various rows (976 to 988) across columns including duration, credit\_amount, installment\_commitment, residence\_since, age, and predicted values (57, 64, 42, 47, 25, 49, 33, 28, 26, 30, 25, 33, 64, 29, 48, 37, 34, 23, 30, 50, 31, 40).

**Bottom Interface:**

- Result list:** Shows five selected models: 08:33:15 - LinearRegression [-F credit\_a], 08:40:24 - LinearRegression [-F duration], 08:43:45 - SMReg [-F duration,credit\_a], 08:45:45 - SMReg [-F duration,credit\_a], 08:50:09 - SMReg [-F duration,credit\_a], 08:53:54 - GaussianProcesses [-F duration], 08:55:10 - AdditiveRegression [-F duration], 08:56:04 - AdditiveRegression [-F duration], 08:59:38 - AdditiveRegression [-F duration], 09:01:01 - LinearRegression [-F duration], 09:02:27 - GaussianProcesses [-F duration], 09:03:42 - LinearRegression [-F credit\_a], 09:04:02 - LinearRegression [-F duration], 09:04:29 - LinearRegression [-F duration], and 09:04:47 - LinearRegression [-F duration].
- Configuration:** Set to "Train future pred." with a time stamp of <None>, periodicity of <Unknown>, and confidence intervals at 95%.
- Output:** Displays a detailed evaluation section for the selected models, including statistics for duration, credit\_amount, and existing\_credits. For duration, the statistics are: N=988, Mean absolute error=7.0231, Root mean squared error=9.226. For credit\_amount, the statistics are: N=988, Mean absolute error=24.1828, Root mean squared error=76.1078. For existing\_credits, the statistics are: N=988, Mean absolute error=0.5061.

### 1) LINEAR REGRESSION AND ITS PREDICTIONS

Preprocess Classification Forecast

Number of time units to forecast: 2

Time stamp: credit\_amount

Periodicity: Daily

Skip list: <Unknown>

Confidence intervals:

Level (%): 95

Perform evaluation:

**Result list:**

- 08:33:15 - LinearRegression [-F credit\_a]
- 08:40:24 - LinearRegression [-F duration]
- 08:43:45 - SMOrreg [-F duration,credit\_a]
- 08:45:45 - SMOrreg [-F duration,credit\_a]
- 08:50:09 - SMOrreg [-F duration,credit\_a]
- 08:53:54 - GaussianProcesses [-F duration]
- 08:55:10 - AdditiveRegression [-F duration]

**Start Stop Help**

**Output/Visualization**

**Output: Train future pred.**

| No. | Name   | 17987 .3874 | 39    | 11760 | 2 | 3  | 32 | 1 | 1 |
|-----|--|-------------|-------|-------|---|----|----|---|---|
| 3   | <input checked="" type="checkbox"/> installment_commitment | 18005 .5796 | 45    | 11816 | 2 | 4  | 29 | 2 | 1 |
| 4   | <input checked="" type="checkbox"/> residence_since        | 18023 .7718 | 24    | 11938 | 2 | 3  | 39 | 2 | 2 |
| 5   | <input checked="" type="checkbox"/> age                    | 18041 .964  | 30    | 11998 | 1 | 1  | 34 | 1 | 1 |
| 6   | <input checked="" type="checkbox"/> existing_credits       | 18060 .1562 | 48    | 12169 | 4 | 4  | 36 | 1 | 1 |
| 7   | <input checked="" type="checkbox"/> num_dependents         | 18078 .3483 | 48    | 12204 | 2 | 2  | 48 | 1 | 1 |
|     |  | 18096 .5405 | 36    | 12389 | 1 | 4  | 37 | 1 | 1 |
|     |  | 18114 .7327 | 24    | 12579 | 4 | 2  | 44 | 1 | 1 |
|     |  | 18132 .9249 | 36    | 12612 | 1 | 4  | 47 | 1 | 2 |
|     |  | 18151 .1171 | 21    | 12680 | 4 | 4  | 30 | 1 | 1 |
|     |  | 18169 .3093 | 48    | 12749 | 4 | 1  | 37 | 1 | 1 |
|     |  | 18187 .5015 | 18    | 12976 | 3 | 4  | 38 | 1 | 1 |
|     |  | 18205 .6937 | 60    | 13756 | 2 | 4  | 63 | 1 | 1 |
|     |  | 18223 .8859 | 60    | 14027 | 4 | 2  | 27 | 1 | 1 |
|     |  | 18242 .0781 | 39    | 14179 | 4 | 4  | 30 | 2 | 1 |
|     |  | 18260 .2703 | 36    | 14318 | 4 | 2  | 57 | 1 | 1 |
|     |  | 18278 .4625 | 48    | 14421 | 2 | 2  | 25 | 1 | 1 |
|     |  | 18296 .6547 | 6     | 14555 | 1 | 2  | 23 | 1 | 1 |
|     |  | 18314 .8468 | 60    | 14782 | 3 | 4  | 60 | 2 | 1 |
|     |  | 18333 .039  | 6     | 14896 | 1 | 4  | 68 | 1 | 1 |
|     |  | 18351 .2312 | 60    | 15653 | 2 | 4  | 21 | 2 | 1 |
|     |  | 18369 .4234 | 48    | 15672 | 2 | 2  | 23 | 1 | 1 |
|     |  | 18387 .6156 | 36    | 15857 | 2 | 3  | 43 | 1 | 1 |
|     |  | 18405 .8078 | 54    | 15945 | 3 | 4  | 58 | 1 | 1 |
|     |  | 18424 .     | 10424 | 1     | 2 | 22 | 1  | 1 |   |

## ADDITIVE REGRESSION AND SMOREG

Preprocess Classification Forecast

Number of time units to forecast: 4

Time stamp: <Unknown>

Periodicity:

Skip list: <Unknown>

Confidence intervals:

Level (%): 95

Perform evaluation:

**Result list:**

- 08:33:15 - LinearRegression [-F credit\_a]
- 08:40:24 - LinearRegression [-F duration]
- 08:43:45 - SMOrreg [-F duration,credit\_a]
- 08:45:45 - SMOrreg [-F duration,credit\_a]
- 08:50:09 - SMOrreg [-F duration,credit\_a]
- 08:53:54 - GaussianProcesses [-F duration]
- 08:55:10 - AdditiveRegression [-F duration]

**Start Stop Help**

**Output/Visualization**

**Output: Train future pred.**

| No. | Name   | 18023 .7718  | 24       | 11938       | 2       | 3       | 39       | 2       | 2       |
|-----|--|--------------|----------|-------------|---------|---------|----------|---------|---------|
| 3   | <input checked="" type="checkbox"/> installment_commitment | 18041 .964   | 30       | 11956       | 1       | 1       | 34       | 2       | 2       |
| 4   | <input checked="" type="checkbox"/> residence_since        | 18060 .1562  | 48       | 11959       | 4       | 4       | 36       | 1       | 2       |
| 5   | <input checked="" type="checkbox"/> age                    | 18078 .3483  | 48       | 12204       | 2       | 2       | 49       | 1       | 2       |
| 6   | <input checked="" type="checkbox"/> existing_credits       | 18096 .5405  | 36       | 12389       | 1       | 4       | 37       | 1       | 1       |
| 7   | <input checked="" type="checkbox"/> num_dependents         | 18114 .7327  | 24       | 12579       | 4       | 2       | 44       | 1       | 1       |
|     |  | 18132 .9249  | 36       | 12612       | 1       | 4       | 47       | 1       | 1       |
|     |  | 18151 .1171  | 21       | 12680       | 4       | 1       | 30       | 1       | 1       |
|     |  | 18169 .3093  | 48       | 12749       | 4       | 1       | 37       | 1       | 1       |
|     |  | 18187 .5015  | 18       | 12976       | 3       | 4       | 38       | 1       | 1       |
|     |  | 18205 .6937  | 60       | 13756       | 2       | 4       | 63       | 1       | 1       |
|     |  | 18223 .8859  | 60       | 14027       | 4       | 2       | 27       | 1       | 1       |
|     |  | 18242 .0781  | 39       | 14179       | 4       | 4       | 30       | 2       | 1       |
|     |  | 18260 .2703  | 36       | 14318       | 4       | 2       | 57       | 1       | 1       |
|     |  | 18278 .4625  | 48       | 14421       | 2       | 2       | 25       | 1       | 1       |
|     |  | 18296 .6547  | 6        | 14555       | 1       | 2       | 23       | 1       | 1       |
|     |  | 18314 .8468  | 60       | 14782       | 3       | 4       | 60       | 2       | 1       |
|     |  | 18333 .039   | 6        | 14896       | 1       | 4       | 68       | 1       | 1       |
|     |  | 18351 .2312  | 60       | 15653       | 2       | 4       | 21       | 2       | 1       |
|     |  | 18369 .4234  | 48       | 15672       | 2       | 2       | 23       | 1       | 1       |
|     |  | 18387 .6156  | 36       | 15857       | 2       | 3       | 43       | 1       | 1       |
|     |  | 18405 .8078  | 54       | 15945       | 3       | 4       | 58       | 1       | 1       |
|     |  | 18424 .      | 10424    | 1           | 2       | 32      | 1        | 1       |         |
|     |  | 18442 .1922* | 6 .7562  | 15792 .0208 | 2 .5142 | 2 .4516 | 55 .0649 | 1 .6197 | 1 .1167 |
|     |  | 18460 .3844* | 26 .7459 | 15792 .0208 | 2 .5142 | 1 .8847 | 56 .8424 | 1 .4181 | 1 .1167 |
|     |  | 18478 .5764* | 29 .8548 | 15792 .0208 | 2 .0946 | 2 .2716 | 22 .4422 | 1 .1746 | 1 .0222 |
|     |  | 18496 .7680* | 39 .0468 | 15792 .0208 | 2 .0896 | 1 .9492 | 26 .5462 | 1 .3104 | 1 .0282 |

**Preprocess Classify Cluster Associate Select attributes Visualize Forecast**

Basic configuration Advanced configuration

**Target Selection:**

All None Invert Pattern

**Output/Visualization**

**Output: Train future pred.**

| No.   | Name   | 976      | 24         | 1258    | 3       | 3        | 57      | 1       | 1 |
|-------|--|----------|------------|---------|---------|----------|---------|---------|---|
| 1     | <input checked="" type="checkbox"/> duration               | 977      | 6          | 753     | 2       | 3        | 64      | 1       | 1 |
| 2     | <input checked="" type="checkbox"/> credit_amount          | 978      | 18         | 2427    | 4       | 2        | 42      | 2       | 1 |
| 3     | <input checked="" type="checkbox"/> installment_commitment | 979      | 24         | 2538    | 4       | 4        | 47      | 2       | 2 |
| 4     | <input checked="" type="checkbox"/> residence_since        | 980      | 15         | 1264    | 2       | 2        | 25      | 1       | 1 |
| 5     | <input checked="" type="checkbox"/> age                    | 981      | 30         | 8386    | 2       | 2        | 49      | 1       | 1 |
|       |  | 982      | 48         | 4844    | 3       | 2        | 33      | 1       | 1 |
|       |  | 983      | 21         | 2923    | 1       | 1        | 28      | 1       | 1 |
|       |  | 984      | 36         | 8229    | 2       | 2        | 26      | 1       | 2 |
|       |  | 985      | 24         | 2028    | 2       | 2        | 30      | 2       | 1 |
|       |  | 986      | 15         | 1633    | 4       | 3        | 25      | 2       | 1 |
|       |  | 987      | 42         | 629     | 2       | 1        | 33      | 2       | 1 |
|       |  | 988      | 13         | 1409    | 2       | 4        | 64      | 1       | 1 |
|       |  | 989      | 24         | 6579    | 4       | 2        | 29      | 1       | 1 |
|       |  | 990      | 24         | 1743    | 4       | 2        | 48      | 2       | 1 |
|       |  | 991      | 12         | 3565    | 2       | 1        | 37      | 2       | 2 |
|       |  | 992      | 15         | 1569    | 4       | 4        | 34      | 1       | 2 |
|       |  | 993      | 18         | 1936    | 2       | 4        | 23      | 2       | 1 |
|       |  | 994      | 36         | 3959    | 4       | 3        | 30      | 1       | 1 |
|       |  | 995      | 12         | 2390    | 4       | 3        | 50      | 1       | 1 |
|       |  | 996      | 12         | 1736    | 3       | 4        | 31      | 1       | 1 |
|       |  | 997      | 30         | 3857    | 4       | 4        | 40      | 1       | 1 |
|       |  | 998      | 12         | 804     | 4       | 4        | 38      | 1       | 1 |
|       |  | 999      | 45         | 15      | 4       | 4        | 23      | 1       | 1 |
|       |  | 1000     | 45         | 4576    | 3       | 4        | 27      | 1       | 1 |
| 1001* |  | 33 .6613 | 4961 .4472 | 1 .7471 | 2 .5222 | 25 .8716 | 1 .22   | 1 .0011 |   |
| 1002* |  | 21 .104  | 5162 .6715 | 0 .726  | 3 .4155 | 43 .0355 | 0 .7277 | 1 .0011 |   |

**Type to enter a caption**

## Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable *causes* the other (for example, higher SAT scores do not *cause* higher college grades), but that there is some significant association between the two variables. A [scatterplot](#) can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the [correlation coefficient](#), which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

A linear regression line has an equation of the form  $Y = a + bX$ , where  $X$  is the explanatory variable and  $Y$  is the dependent variable. The slope of the line is  $b$ , and  $a$  is the intercept (the value of  $y$  when  $x = 0$ ).

### Least-Squares Regression

The most common method for fitting a regression line is the method of least-squares. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values.

#### Example

The dataset "Televisions, Physicians, and Life Expectancy" contains, among other variables, the number of people per television set and the number of people per physician for 40 countries. Since both variables probably reflect the level of wealth in each country, it is reasonable to assume that there is some positive association between them. After removing 8 countries with missing values from the dataset, the remaining 32 countries have a correlation coefficient of 0.852 for number of people per television set and number of people per physician. The  $r^2$  value is 0.726 (the square of the correlation coefficient), indicating that 72.6% of the variation in one variable may be explained by the other. (Note: see [correlation](#) for more detail.) Suppose we choose to consider number of people per television set as the explanatory variable, and number of people per physician as the dependent variable. Using the MINITAB "REGRESS" command gives the following results:

The regression equation is People.Phys. = 1019 + 56.2 People.Tel.

To view the fit of the model to the observed data, one may plot the computed regression line over the actual data points to evaluate the results. For this example, the plot appears to the right, with number of individuals per television set (the explanatory variable) on the x-axis and number of individuals per physician (the dependent variable) on the y-axis. While most of the data points are clustered towards the lower left corner of the plot (indicating relatively few individuals per television set and per physician), there are a few points which lie far away from the main cluster of the data. These points are known as *outliers*, and depending on their location may have a major impact on the regression line (see below).



### REGRESSION EQUATION

### URL Link to Online Resources

- 1) <https://machinelearningmastery.com/time-series-forecasting/>

A detailed article explaining the concept of Time Series Forecasting and more information on Time Series as a whole.

## 8. Research Publication Summary and relevance / potential relevance to your work

- **Publication and researchers**

### **Citation**

1Gurram Sai Kumar. " Credit Card Fraud Detection System Based On Machine Learning Techniques." IOSR Journal of Computer Engineering (IOSR-JCE) 21.3 (2019): 45-52.

Publication - IOSR Journal of Computer Engineering (IOSR-JCE)

Researchers - Gurram Sai Kumar, Madala Vekaiah Naidu, Dr. Mandugula Sujatha

- **Dataset**

German Credit Card dataset obtained from the UCI (University of California, Irvine) machine learning repository

- **Technique (mention any adaptions)**

In this research, ensemble models were the majority concerned.

Bagging is an ensemble algorithm that is used to improve factors such as stability and accuracy of a machine learning algorithm. Random Forest is another ensemble algorithm that helps to identify relevant predictor variables to make feature selection easier.

eXtreme Gradient Boosting (XGBoost) is a kind of GBM model that follows the principle of gradient boosting. The differences in modelling details that exist are that XGBoost uses a more regularized model formalization to control over-fitting which helps achieve better performance.

Light Gradient Boosting Machine (LightGBM) is a gradient boosting framework that works upon tree-based algorithms. Given its highly efficient and scalable behaviour it is capable of supporting many different GBM Algorithms. It is several times faster than most existing implementations of gradient boosting trees which is backed by its fully greedy tree-growth method, histogram-based memory and computation optimization.

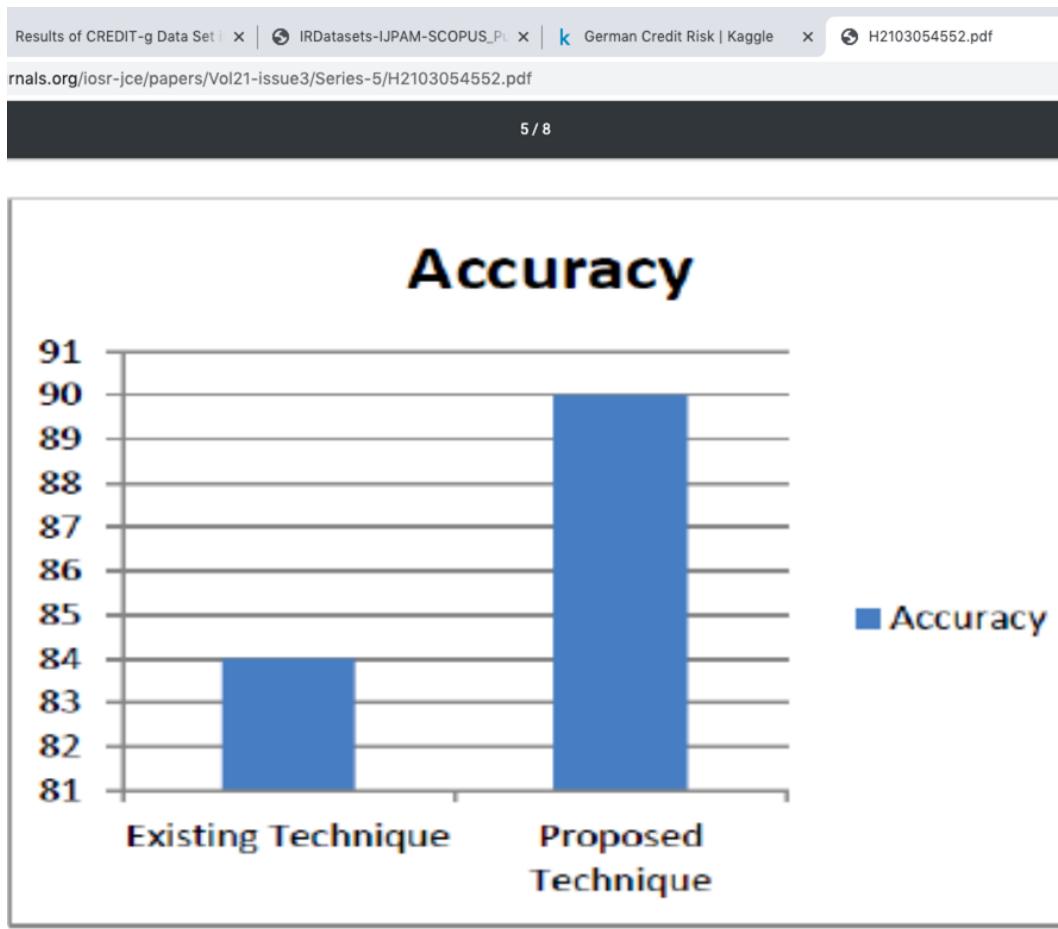
Adaptive Boosting(AdaBoost) has been used as part of the implementation method in order to boost the performance of decision tree and has been implemented in WEKA(Waikato Environment for Knowledge Analysis).This boosting algorithm can be applied to any classifier's learning algorithm.

- **Major Findings**

Two different forms of experimental results have been provided, which are:-

- 1) Experimental results of decision tree without any boosting techniques
- 1) Experimental results of decision tree together with AdaBoost.

The proposed algorithm is a clear cut enhancement of the already available algorithms. Application of boosting algorithms in combination with these algorithms gives much higher accuracy. On analysis of various other features such as Sensitivity and Specificity , it can clearly



|                    | KNN  | Random Tree                            | Proposed Algorithm              |
|--------------------|--|--|---------------------------------|
| <b>Accuracy</b>    | 0.9691   | 0.9432                                 | 0.9824                          |
| <b>Sensitivity</b> | 0.8835   | 0                                      | 0.9767                          |
| <b>Specificity</b> | 0.9711   | 0                                      | 0.9824                          |
| <b>Limitations</b> | Cannot detect the fraud at the time of transaction | Not suitable for Randomness in dataset | Not Applied for non-Linear data |

TABLE: COMPARISION OF ALGORITHMS

TYPE TO ENTER A CAPTION.

be derived that the proposed algorithm does a better job than any existing technique. The only limitation encountered in the proposed algorithm is its non applicability to Linear data.

- **Relevance / potential relevance to your work**

The research paper gives an insight into making the existing algorithms better by addition of boosting algorithms. Since, we have used the basic techniques such as J48 tree, Voted Perceptrons, K-Means and Time Series Forecasting, we have only looked into the basic existing techniques that are possible to solve the problem of Credit Card Fraud. Adding of Boosting Algorithms to the techniques analysed in this report would be a beneficial task. We can further investigate into ways to enhance the current models and come up with the best approach after analysing it on various factors of accuracy , sensitivity and specificity.

## **9. Division of Labor**

