**Coursework 2 Report**
**Davide Pollicino**
40401270@live.napier.ac.uk
**Edinburgh Napier University - Dana Analytics Coursework 2 (SET09120)**

## 1. Introduction

The current report will show and analyze the actions that I have taken in order to improve the **data fineness** of a given dataset and logic rules used by our model, created by the application of **Clustering, Classification and Association** algorithms, during the use of OpenRefine an Weka in order to conduct this exploratory **data mining** experiment.

For the data cleaning process, I have used **OpenRefine**, a web App tool developed by Google used to manage, modify and improve data-consistency of a given dataset. For the application of different Machine Learning algorithms that will be discussed below,I have used **Weka[1]**, a suite of machine Learning software developers using Java by the University of Waikato, widely used for data conversion, data processing, regression, clustering, classification and data visualization.

## 2. Data Preparation

Data preparation is an essential step in the field of **Data Mining,** used to fit any dataset for its future use. The operation of Data Preparation, in this coursework, was executed using **OpenRefine**,

The analysis and improvement of the **data quality** will reflect the performances of our future **models.** In this step, the key points to analyze and remove:

- Null values
- Multiple identical records
- Data conversion and integration errors

### 21. Data Cleaning

In our dataset, during the Data Cleaning process, I have cleaned and improved data consistency, updating attributes names and attributes values, in order to let the visualization of our dataset be more understandable and increase the **accuracy** of our models.

Updates apported to the attributes Names:

| Original attribute name | Updated attribute Name |
|---|---|
| '<0' | Case_no |
| 'critical/other existing credit' | checking status |
| 1169 | credit history |
| radio/tv | purpose |
| 'no known savings' | credit amount |

| | |
|---|---|
| '>=7' | employment |
| 'male single' | personal_status |
| 67 | age |
| skilled | job |
| good | class |

Updates apported to the attributes values during the data cleaning process.

| Attribute | Original | After the updated | Description |
|---|---|---|---|
| **Job** | | | |
| | yes | skilled | Assumed that yes could me “skilled” |
| **personal_status** | | | |
| | 'female dvi/del/mar' | ‘female div/sep/mar’ | Fixed the misspelling in “dep” |
| **Age** | | | |
| | 222 | 22 | Removed last digit |
| | 333 | 33 | Removed last digit |
| | 1 | 21 | Assumed that would be 21, the minimum value that contains 1 as digit after 18. |
| | 0.44 | 44 | Changed to adult integer age. |
| | 0.35 | 35 | Changed to adult integer age. |
| | 0.24 | 24 | Changed to adult integer age. |
| | `-35 | 35 | Converted to positive |
| | -29 | 29 | Converted to positive |
| | -34 | 34 | Converted to positive |
| | 6 | 26 | Assume that would be 26, the minimum value that contains 6 as a digit after 18. |

| | | | |
|---|---|---|---|
| **purpose** | | | |
| | ather | other | Spelling error |
| | business | business | Spelling error |
| | business | business | Spelling error |
| | Education | Education | Spelling and case error |
| | Radio/Tv | radio/tv | case error |
| | ‘dometic appliance’ | domestic appliance | removed quotes |
| | ‘new car’ | new car | removed quotes |
| | ‘used car’ | used car | removed quotes |
| **credit amount** | | | |
| | 5850000 | 585 | Removed zeros |
| | 5180000 | 518 | Removed zeros |
| | 7190000 | 719 | Removed zeros |
| | 13580000 | 1358 | Removed zeros |
| | 19280000 | 1928 | Removed zeros |
| | 13860000 | 1386 | Removed zeros |
| | 63610000 | 6361 | Removed zeros |
| | 11132800 | 1328 | Removed zeros |

### 2.2 Data Conversation

In order to use different algorithms like **J48** and **Apriori**, it has been necessary to convert our data from **numeric** to **nominal,** using the **unsupervised NumericToNomina filter** provided by Weka.
Even if not used, a numeric version of our dataset has been also created and attached to this report, in order to make it available for future applications of this dataset.

## 3.0 Data Analytics

Data analysis is a set of processes that involves data inspection, data cleaning and data modeling in order to extract useful information from data, such as associations, patterns, anomalies and significant structure from large amounts of structured or unstructured data.

Thanks to data analytics, we can train any model, reducing also the risk of **overfitting.** Overfitting is an internal state of a model, caused by a huge supply of data, that leads our model to **not** being able to predict accurately a given class, if a slightly different dataset will be given.

### 3.1 Classification

Classification is a **Supervised Learning** technique used to predict our nominal attribute class, called class, to analyze the generic rules to establish in this case who is highly raccomandabile for getting a loan and who is not.

The algorithm that I have used is **J48**, an algorithm used to generate a decision tree[**Figure 2]** that will allow us to predict the target class of any new dataset entry/record. (In **figure 1**, it is possible to view the **prefuse tree visualization** created by the J48 algorithm )

In order to decrease the chance to have an overfitting model, we will apply **pruning**, a technique used in Machine Learning to reduce the size of the decision tree, decreasing the space and time complexity of the classifier and increasing the accuracy.

In order to analyze the precision of our classifier, we can observe the **confusion matrix** and the accuracy estimation provided by Weka.
The confusion matrix is a table used to quantify the performance of an algorithm, showing the number of true positives (correct predictions), false positives, true negatives and false negatives.

```
  a   b   <-- classified as
651  49 |   a = good
148 152 |   b = bad
```

This image shows that there were 49 false positives and 148 false negatives, which were incorrectly classified.
The **accuracy** of our model, reported by weka is: 80.3%.

**Rule 1:**

**IF checking_status** = <0 **AND credit_history** = existing paid **AND saving_status** = <100 **AND age** = 19<=X<29 AND **credit_amount** = 2000<=X<4000: **THEN** good (7.0/1.0)

If a loan applicant has no current credit but with an existing (paid) credit history, and the age of the applicant is between 19 and 28 (inclusive) and then credit amount of the loan is between 2000 and 3999 (inclusive), the loan will be given. Out of 7 cases that apply to this rule, 1 has been incorrectly classified, which means that if machine learning was to learn from this dataset, it would predict 1 wrong case out of 7 cases.

**Rule 2:**

**IF checking_status = no checking: THEN good (394.0/46.0)**

As we can see in this rule, 88% of the clients who had no checking status got the loan. However, 46 of the 394 were incorrectly classified.

**Rule 3:**

**IF checking_status** = < 0 **AND credit_history** = existing paid **AND** (**purpose** = furniture/equipment **OR** purpose = used car **OR** purpose = business) **THEN** bad

If the loan applicant has a checking status of less or equal to and an existing paid credit history, if the purpose of the lean is an used car or a furniture acquisition or a new business lunch, they are not luckily to get a new loan.
Of course, changing the purpose of the loan could mean increasing the chances to get the loan application accepted.

**Rule 4:**

**IF checking_status** = 0 <=x<=200 **AND credit_history** = existing paid **AND credit_amount** = 0<=X<2000 **AND** **purpose** = radio/tv: good (40.0/9.0)

If the client has a checking_status equal to 0 and and a paid credit history, if the amount of credit asked for the loan is between 0 and 1999 (inclusive) and the purpose of the loan is a radio/Tv, there are high chances to get a loan. From this specific dataset, every 40 cases, 9 would have a wrong prediction.

**Rule 5:**

**IF checking_status** = <0 **ANDcredit_history** = existing paid **AND purpose** = furniture/equipment **AND** employment = 1<=X<4 AND **age** > 23 **THEN** good (13.0/1.0)

If the client has a checking stato smaller than zero, an existing paid credit history and the purpose of the loan if furniture/equipment, if the client has an employment history that goes from 1 to 4 years and the age of the loan applicant is greater than 23, it is highly probable that the loan will be accepted. According to our dataset, every 13 predictions, 1 of them will be wrong.

**Rule 6:**

**IF checking_status** = < 0 **AND credit_history** = existing paid **AND purpose** = new car **THEN** bad (31.0/9.0)

If the loan applicant has a checking status smaller than 0, an existing paid credit history and the purpose of the loan is a new car, it is highly possible that the loan will **not** be accepted. According to our dataset, every 31 predictions, 9 of them could be wrong. The final decision could always be changed by extra attributes like age and credit amount required for the loan.

### 3.2 Regression

Regression is another prediction task used to predict numerical values. Considering that the attribute that in this case we need to predict is a nominal class, I have decided to not use regression, using instead Classification, association and Clustering.

### 3.3 Association

Association is a **Rule-based machine learning** (RBML) Machine Learning technique used to **find relationships** between variables in a database.
In order to extract information from our data and create a classifier using Association, a dataset composed by nominal attributes has been given in input to the **Apriori Algorithm**

Apriori algorithm is used for frequent item set mining and association rule learning, decreasing the overall time complexity of the algorithms by decreasing the size of the candidate sets.

**Rule 1:**

**checking_status** =no checking **credit_history**=critical/other existing credit 153 ==> **class**=good 143 <conf:(0.93)> lift:(1.34) lev:(0.04) [35] conv:(4.17)

If a client without a bank account in our bank (no checking credit) and with an existing debt with another bank, could in 93% of cases get a loan according to our dataset.

**Rule 2:**

**checking_status** = no checking **purpose**=radio/tv 127 ==> **class**=good 120 <conf:(0.94)> lift:(1.35) lev:(0.03) [31] conv:(4.76)

If there is no checking status and the purpose is radio/tv then they are likely to get the loan. There is a confidence of 0.94 which means that this rule is very accurate and is true for 94% of the cases that have those attributes.

**Rule 3:**

**checking_status** = no checking **employment**=>=7 115 ==> **class**=good 107 <conf:(0.93)> lift:(1.33) lev:(0.03) [26] conv:(3.83)

Clients that don't have information about their checking status but have been working for more than 7 years are more likely to get the loan they request. This rule also has a confidence of 0.93. A confidence of 0.9 means that the rule is 90% efficient therefore if the value is 0.90 or above it means it is a very good model.

**Rule 4:**

**checking_status** =no checking **personal_status**=male single **job**=skilled 151 ==> **class**=good 139 <conf:(0.92)> lift:(1.32) lev:(0.03) [33] conv:(3.48)

Single males loan applications that have a skilled job and no checking status are very likely to get the loan in the 92% of cases.

**Rule 5:**

**checking_status** =no checking **age**=29<=X<39 151 ==> **class**=good 137 <conf:(0.91)> lift:(1.3) lev:(0.03) [31] conv:(3.02)

clients aged between 29 and 39 and that have no checking status are likely to get the loan they requested. This is true for 91% of the clients.

**Rule 6:**

**checking_status** =no checking **credit_amount**=0<=X<2000 172 ==> **class**=good 157 <conf:(0.91)> lift:(1.3) lev:(0.04) [36] conv:(3.23)

In general, clicking without any checking status that are applying for a very small credit amount that goes between 0 and 2000 are likely to get a loan in the 91% of cases.

### 3.4 Clustering

Clustering is another unsupervised learning method that divides the population or data points into a smaller number of groups such that data points in the same groups are more similar to other points in the same group (cluster).

The algorithm that I have used is SimpleKMeans. This algorithm uses a number K (by default 2 in Weka, in this case 6) that represents the number of clusters to be found. SimpleKMeans use an Euclidean distance as function in order to measure the distance between each instance and each cluster, in order to decide in which cluster insert an instance. Of course, in an initial stage, this operation is done randomly and the algorithm will stop when no more instances will be moved between clusters.

The default number of min. clusters generated by Weka using the SimpleKMeans algorithm is 2 and in order to get 6 rules, it has been necessary to modify the Weka default's settings.

Below, please find the 6 clusters made:

| | **cluster 0** | **cluster 1** | **cluster 2** | **cluster 3** | **cluster 4** | **cluster 5** |
|---|---|---|---|---|---|---|
| **checking_ status** | no checking | < 0 | no checking | no checking | <0 | 0<=X<200 |
| **credit history** | critical/other existing credit | critical/other existing credit | existing paid | existing paid | existing paid | existing paid |
| **purpose** | new car | used car | radio/tv | radio/tv | new car | radio/tv |
| **credit amount** | 0<=X<20 | 2000<=X<4000 | 0<=X<2000 | 0<=X<2000 | 2000<=X<4000 | 0<=X<2000 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | 00 | | | | | |
| **saving status** | <100 | <100 | <100 | no known savings | <100 | <100 |
| **employment** | 1<=X<4 | 4<=X<7 | >=7 | >=7 | >=7 | 1<=X<4 |
| **personal status** | female div/sep/mar | male single | male single | male single | male single | female div/sep/mar |
| **age** | 19<=X<29 | 29<=X<39 | 39<=X<49 | 29<=X<39 | 29<=X<39 | 19<=X<29 |
| **job** | skilled | high qualif/self emp/mgmt | unskilled resident | skilled | skilled | skilled |
| **class** | **bad** | **good** | **good** | **good** | **bad** | **good** |

In the **Figure 3.**, has been shown the clustering representation of the clusters created by SimpleKMeans.

```
Clustered Instances

0       200 ( 20%)
1       149 ( 15%)
2       166 ( 17%)
3       175 ( 18%)
4       134 ( 13%)
5       176 ( 18%)
```

As we can see most of the data was equally separated. Cluster 0 (#) is the biggest one. The whole dataset was 1000 instances.

We can see from the cluster 0 that a young female application with age between 19 and 29, with existing debts with the intention to get a loan for a new car,

**4 Conclusion**

Clearly,by the information provided by each rule, we can see that the fundamentals attributes used for final decision of permission or negation of a loan are credit_history and checking_status followed by the purpose and credit_amount requested. A tree has been also created, using the weka plug-in prefuse tree.

Between all the Machine Learning techniques applied in this coursework, the association technique is the one that can be used also for unsupervised machine learning algorithms, where the outcome of the result is not known.

## 5.0 Appendix and references

**[1]:**https://www.analyticsvidhya.com/learning-paths-data-science-business-analytics-business-intelligence-big-data/weka-gui-learn-machine-learning/

**Figure 1. Prefuse Tree Visualization (Classification)**

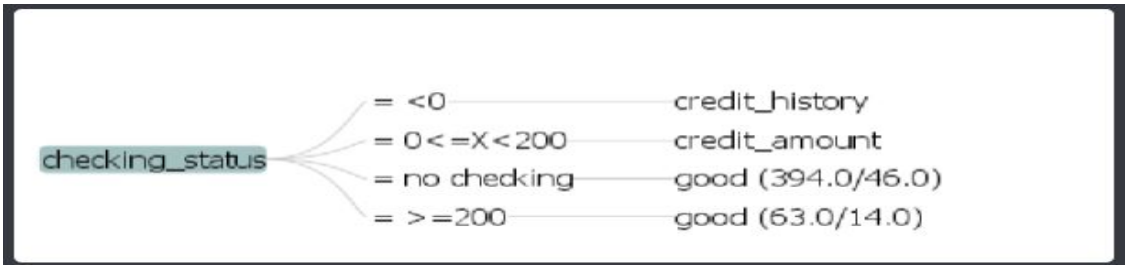


**Figure 2. Tree Visualization**

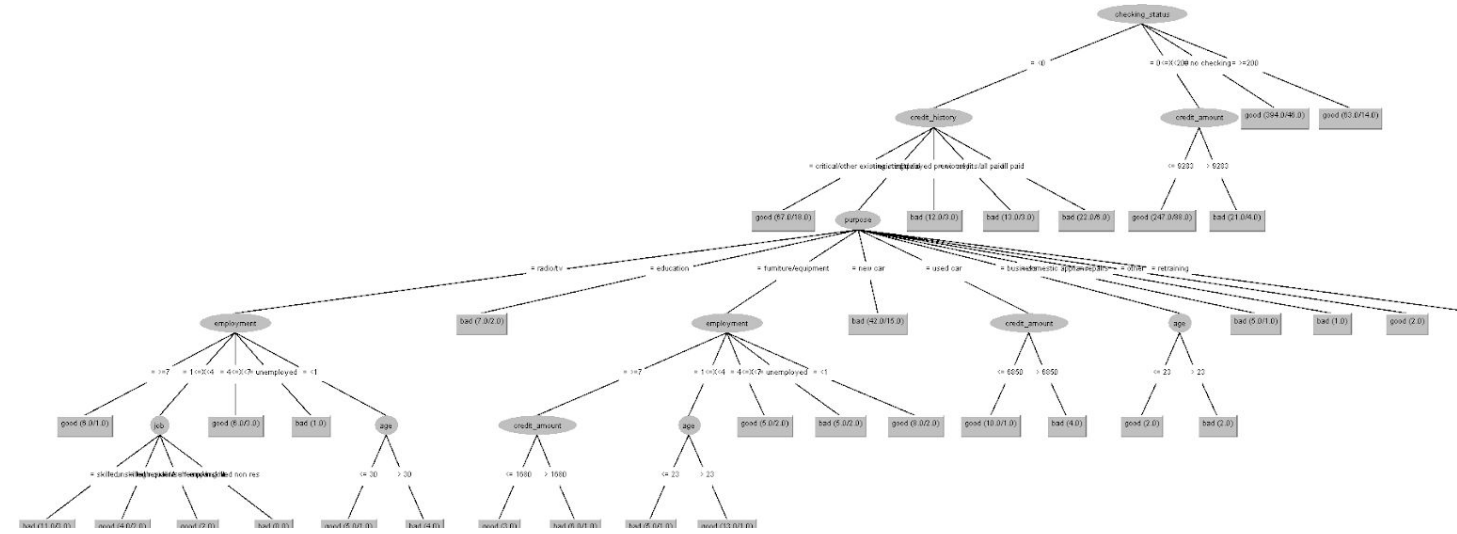

**Figure 3. Data Visualization of Clustering SimpleKMeans Algorithm**