Data Analytics

Classification:

IF checking\_status = <0 AND credit\_history = existing paid AND purpose = radio/tv AND employment = >=7 THE good (6.0/1.0)

6/1 means that out of 6 cases, one is incorrectly classified.

1. **Introduction**

Briefly introduce the aims of this coursework.

1. **Data Preparation**
   1. **Data Cleaning**

Describe the way the data is cleaned by the use of OpenRefine.

**2.2. Data conversion**

Describe how the cleaned data is converted to data sets, which can be analysed by

algorithms.

1. **Data Analytics**
   1. **Classification**

What it is

Pruning

What the j48

Rules (5-7)

Accuracy of algorithm

Confusion matrix

* 1. **Regression**
  2. **Association**

What it is

Where it is used

What is apriori

Parameters changed (a rule that has a confidence of over 0.9)

Rules (6)

* 1. **Clustering**

What it is

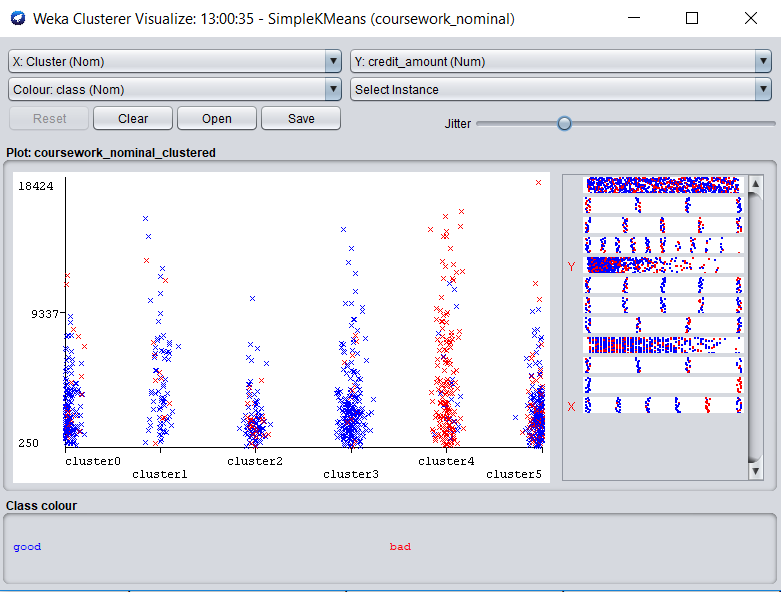
Algorithm simplekmeans

Table with clusters

Size of the clusters

How many cases fall into that

Visualisation:



|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Original Data** | **Changed Data** | **Reason** |
| purpose |  |  |  |
|  | ather | other | Spelling mistake |
|  | business/busness | business | Spelling mistake |
|  | Education | education | Spelling mistake |
|  | Radio/Tv | radio/tv | Spelling mistake |
| credit\_amount |  |  |  |
|  | 111328000 | 8582 | Calculated the mean for the purpose other and replaced with that value |
|  | 19280000 | 1928 | Removed all zeros after looking at the credit\_amount values in that Purpose |
|  | 13580000 | 8582 | Calculated the mean for the purpose other and replaced with that value |
|  | 13860000 | 1386 | Removed all zeros after looking at the credit\_amount values in that Purpose |
|  | 63610000 | 6361 | Removed all zeros after looking at the credit\_amount values in that Purpose |
|  | 5180000 | 5180 | Removed all zeros after looking at the credit\_amount values in that Purpose |
|  | 5850000 | 5850 | Removed all zeros after looking at the credit\_amount values in that Purpose |
|  | 7190000 | 7190 | Removed all zeros after looking at the credit\_amount values in that Purpose |
| personal\_status |  |  |  |
|  | Female div/sep/mar | Female div/sep/mar | Changed dep to sep to correspond with similar value male div/sep |
| age |  |  |  |
|  | -29 | 29 | Removed negative |
|  | -34 | 34 | Removed negative |
|  | -35 | 35 | Removed negative |
|  | 0.24 | 24 | Removed decimal |
|  | 0.35 | 35 | Removed decimal |
|  | 0.44 | 44 | Removed decimal |
|  | 1 | 19 | Looked at similar  values for this row to  assume |
|  | 6 | 26 | Looked at similar  values for this row to  assume |
|  | 222 | 22 | Removed the last  digit |
|  | 333 | 33 | Removed the last  digit |
| job |  |  |  |
|  | yes | skilled | Looked at similar  values for this row to  assume |
| age | to nominal | in increments of 10  19<=X<29 etc… | because data needs to be nominal for association and some classification algorithms |
| changed credit amount to nominal in increments of 2000 |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |