

Data_Mining_Ashleigh_Simons

by Ashleigh Simons

Submission date: 13-May-2024 03:56PM (UTC+0100)

Submission ID: 233459784

File name: 23605_Ashleigh_Simons_Data_Mining_Ashleigh_Simons_565911_1817325174.docx (835.08K)

Word count: 3766

Character count: 19385

Module: Data Warehousing and Data
Mining

Module Code: ACCA6026

Student Name: Ashleigh Simons

Student Number: 2003948

Task One

The Purpose and Practice of Data Mining in Businesses

Data mining involves “searching and analysing large batches of raw data” [1]. This data is required to be collected, adjusted so that it is usable, then analysed in a number of separate ways to understand what the data is presenting. This data can show new insights, and potentially show relationships that may suggest reasoning for actions to be taken. There are a number of ways that businesses can make use of data mining and the raw data that they have collected to analyse. This will potentially outline strategies that can be used and taken to gain an advantage and understand what could or might happen next.

The first of many ways businesses can use data mining is for “pattern recognition” [2]. When analysing large sets of data, patterns may identify relationships within the data that are not easily noticeable at first. This, in the context of a business, may provide some valuable insights that the business can take advantage of. One way they can take advantage of this is by “predicting future outcomes” [3]. This implies that the business would be able to make informed decisions for adapting a process, implementing something new, or eliminating a process that is found to be negative towards the business operations. Predicting future outcomes also improves “Risk Management” [4]. If a business knows what could become a threat to the business, then actions can be taken to try and prevent or reduce the damage of the risk. It also allows the business to know the risks of implementing something new. After using these aspects of data mining, “process optimisation” [5] can be used. This is the form of understanding something that the business does and improving it to produce a more accurate desired and beneficial outcome.

Data Set Description

Within the South German Credit Data there is dataset about customers of a German bank that was collected from 1973 – 1975. Within this dataset there is a number of attributes that are listed in German, their variable names, and the purpose of the content. It is aimed that using this data, to data mine and analyse will show factors which could suggest whether a customer may fully pay back a loan they have or if they will fail to repay their loan in time.

The attributes that can be found within the dataset when translated to English consist of Checking account status, credit duration in months, credit history, purpose of credit, credit amount, savings account, employment duration, instalment rate, sex and marital status, another debtors/guarantor, current residence, property, age, instalment plans, housing, number of credits the debtor has, job, number of people liable, telephone, foreign worker, and credit risk.

Within this dataset there also consists of one thousand records of data across the different attributes and using the brief of the data of these 1000 records, 300 of them consist of records with bad credits, and 700 records of good credits with predictor variables. This could suggest that the data is unbalanced as there is not an even amount records for both good and bad, this could impact the results found during the sample runs and results phase. There are also no missing values present within the dataset, so there will be no need to populate the dataset with data to complete the dataset and eliminate records deemed not usable because of a lack of data present.

Pre-Processing

Data preprocessing is the method used when “transforming raw data into an understandable process” [7]. This must be completed to ensure that the data is of quality and to make sure that it is usable when applying data mining algorithms. Whilst preprocessing, it comes with the requirements of making sure that the data is “Accurate, Complete, Consistent, Timeless, Believable, and Interpretable” [8]. Having the presence of these aspects we can make sure that the data we use is correct and complete without any gaps or errors. It will also make sure that the data is similar within the columns and rows, and that it has been updated if the data is being required to be used regularly. Finally, it should also be from trusted sources and understandable so that the data can be used correctly and with confidence that the future findings will be reliable.

This process can be split into four sections, the first being data cleaning. Data cleaning is the process of “fixing, removing” [9] raw data. This is vital as it is an opportunity to get rid of any data that may be corrupted, incorrectly formatted, or duplicated within the dataset which if left included, it could impact the results concluded from the data which would make the findings less dependable. Data Integration refers to the collection and combination of “multiple sources” [10]. If we had two files that needed to be brought together to be used, this would be included within this stage. There would have to be a considerable amount of care taken to combine two files and depending on what is included within them not all the data may be able to be used. The creation of additional attributes would be required within the data and having the correct data populated within those attributes within the data would be required, requiring checks to be made afterwards to understand if this has been completed correctly.

Data Reduction is the next section, and this is required to “reduce the size of the original data” [11]. This is sometimes required when a data set is exceptionally large, the reduction aims to improve the efficiency of the software being used to mine the data. This data could for example be split in half to take less time to run but will not see much difference with the data being only one thousand records. This would be most effective if the data had millions of records, that can be split up and then run separately so that the data is being kept accurate and can be used

more efficiently, rather than the amount of time being longer and trying to process all of the data at the same time and potentially experiencing issues.

Data Transformation is also one of the sections of these and is typically involved at the same time as data cleaning. This can involve changing the “format, structure, and values” [12]. If the raw data is provided in a number of file types and formats, then it is required to be change into something that is usable to the software that is being used for the data mining. Within this case it will be Weka 3.8.6 which requires the file types to be .ARFF files. This means that the files provided will need to be converted. In addition to this, the file format and structure will need to be change.

```
@relation german_credit
@attribute 'checking_status' { '<0', '0<=X<200', '>=200', 'no checking' }
@attribute 'duration' real
@attribute 'credit_history' { 'no credits/all paid', 'all paid', 'existing paid', 'delayed previously', 'critical/
@attribute 'purpose' { 'new car', 'used car', furniture/equipment, radio/tv, 'domestic appliance', repairs, educat
retraining, business, other }
@attribute 'credit_amount' real
@attribute 'savings_status' { '<100', '100<=X<500', '500<=X<1000', '>=1000', 'no known savings' }
@attribute 'employment' { 'unemployed', '<1', '1<=X<4', '4<=X<7', '>=7' }
@attribute 'installment_commitment' real
@attribute 'personal_status' { 'male div/sep', 'female div/dep/mar', 'male single', 'male mar/wid', 'female single
@attribute 'other_parties' { none, 'co applicant', guarantor }
@attribute 'residence_since' real
@attribute 'property_magnitude' { 'real estate', 'life insurance', car, 'no known property' }
@attribute 'age' real
@attribute 'other_payment_plans' { bank, stores, none }
@attribute 'housing' { rent, own, 'for free' }
@attribute 'existing_credits' real
@attribute 'job' { 'unemp/unskilled non res', 'unskilled resident', skilled, 'high qualif/self emp/mgmt' }
@attribute 'num_dependents' real
@attribute 'own_telephone' { none, yes }
@attribute 'foreign_worker' { yes, no }
@attribute 'class' { good, bad }
@data
'<0',6,'critical/other existing credit',radio/tv,1169,'no known savings','>=7',4,'male single',none,4,'real
estate',67,none,own,2,skilled,1,yes,yes,good
'0<=X<200',48,'existing paid',radio/tv,5951,'<100', '1<=X<4',2,'female div/dep/mar',none,2,'real
estate',22,none,own,1,skilled,1,none,yes,bad
'no checking',12,'critical/other existing credit',education,2096,'<100', '4<=X<7',2,'male single',none,3,'real
estate',49,none,own,1,'unskilled resident',2,none,yes,good
'<0',42,'existing paid',furniture/equipment,7882,'<100', '4<=X<7',2,'male single',guarantor,4,'life insurance',45,n
free',1,skilled,2,none,yes,good
'<0',24,'delayed previously', 'new car',4870,'<100', '1<=X<4',3,'male single',none,4,'no known property',53,none,'fo
free',2,skilled,2,none,yes,bad
'no checking',36,'existing paid',education,9055,'no known savings', '1<=X<4',2,'male single',none,4,'no known prope
free',1,'unskilled resident',2,yes,yes,good
'no checking',24,'existing paid',furniture/equipment,2835,'500<=X<1000', '>=7',3,'male single',none,4,'life
```

Figure 1 – Correct Data File Structure and Content.

Above you can see it starts with the relation, which is the name of the data set, then the attributes of the data followed by its specifications, and then the data separated by commas, which populates each of the attributes [13]. This structure is now fully complete, and we know that it already contains no missing values and is accurate to the specifications of each attribute that the data can be opened and used now within our software.

Post-Processing

Noe that the data is ready to be opened in our software we can by selecting the file that has the prepared data and structure. To begin this process, we can click on the 'Explorer' application on Weka, and then open the file in the location that it is stored in. Once this is complete, we can see the following environment.

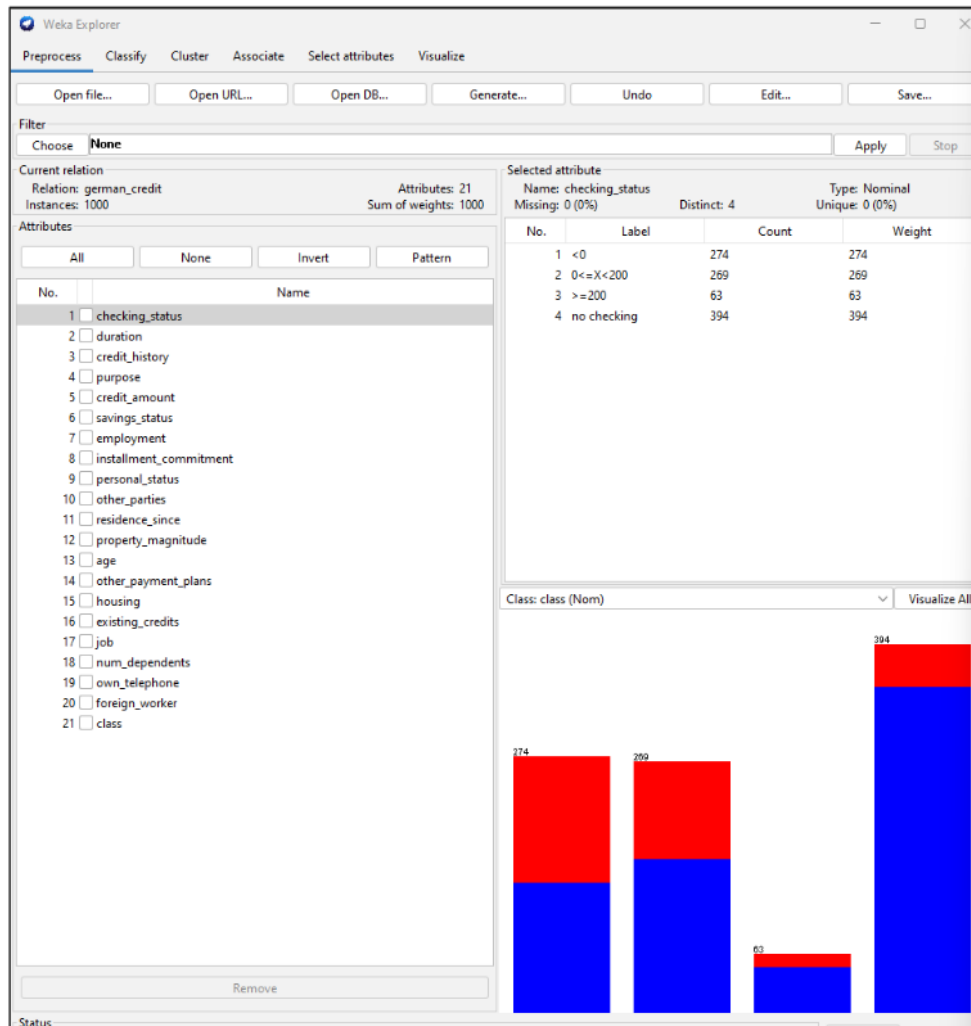


Figure 2 – Explorer Screen in Weka with Loaded in Dataset.

Above you can see the screen where we can start to understand more about the data. In the main body we can see the current relation, attributes, selected attributes, and class. These provide us with some information about the data that has been selected.

Attributes

All

None

Invert

Pattern

No.	Name
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
16	☐ existing_credits
17	☒ job
18	☐ num_dependents
19	☐ own_telephone
20	☐ foreign_worker
21	☐ class

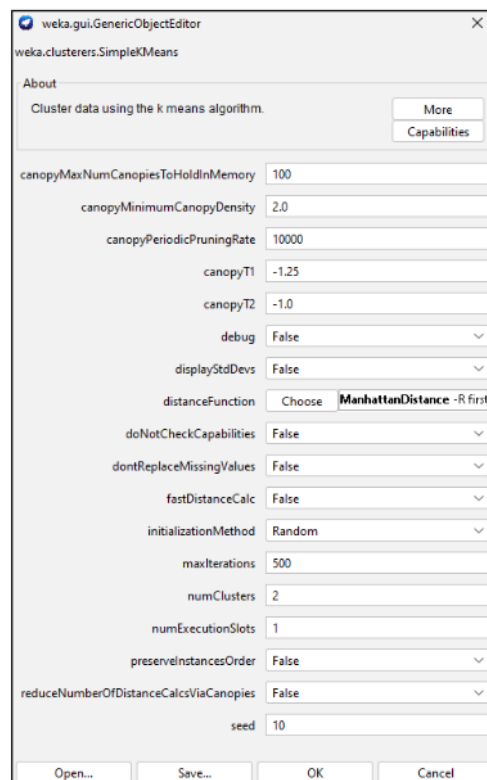
Figure 4 – Selected Attributes

Furthermore, now that we have these selected attributes, we can remove the remaining unselected attributes as these will not be used within the next phase. The reasoning for why the chosen attributes have been selected and why they could provide valuable insight and information is based on their potential insights into the type of loans that are being paid back and their amounts. If an individual works, what type of jobs they are working in may also affect their ability to pay their loans back. The insight into their accounts both checking, and savings may indicate how much money the individual has, and this could affect their ability to pay back their loan. Knowing about their personal status and other parties may also affect their ability to pay the loan back or someone else being able to if they are not themselves. All these selected attributes once classified and clustered may suggest reasons and provide us with some patterns about customers and their likelihood of being able to pay their loans and what type of loans.

Sample Runs and Results

Within this section it will show the methods used for analysis and the finds of the analysis that has been conducted. To begin this process Simple K Means will be the first cluster method that will be used. The benefit to using this method over other methods is that it is simple to use and typically easy to implement. This method will allow us to select several clusters, which will then be randomly selected amongst the data. These clusters will be known as the centroids and allow data to cluster around that centroid. There are some weaknesses to this method starting with the requirement to be provided with the number of clusters (K). For this reason, the number of clusters will be varied and stated to suggest possible reasons for the data appearing the way that it might. There will also be the use of a percentage split, this will allow the data to be trained and then assessed on. Hopefully, producing more accurate results.

SimpleKMeans Algorithm 2 Clusters



The screenshot shows the 'weka.gui.GenericObjectEditor' window for the 'weka.clusterers.SimpleKMeans' class. The 'About' tab is active, displaying the description: 'Cluster data using the k means algorithm.' Below this, various configuration parameters are listed with their current values:

- canopyMaxNumCanopiesToHoldInMemory: 100
- canopyMinimumCanopyDensity: 2.0
- canopyPeriodicPruningRate: 10000
- canopyT1: -1.25
- canopyT2: -1.0
- debug: False
- displayStdDevs: False
- distanceFunction: Choose ManhattanDistance -R first
- doNotCheckCapabilities: False
- dontReplaceMissingValues: False
- fastDistanceCalc: False
- initializationMethod: Random
- maxIterations: 500
- numClusters: 2
- numExecutionSlots: 1
- preserveInstancesOrder: False
- reduceNumberOfDistanceCalcsViaCanopies: False
- seed: 10

At the bottom of the window are four buttons: 'Open...', 'Save...', 'OK', and 'Cancel'.

This is the first Cluster, as you can see to the left it shows the object editor which allows us to change details about the algorithm that is going to be used. This cluster data will be being the K means algorithm and we can see there is two clusters using a random method, this also has a maximum number of iterations of 500 which is the training data, and the rest of the data will be used as testing data.

We can also see that the distance function is using the Manhattan Distance, and this is the function for calculating which centroid the data will cluster to depending on the closest distance horizontally and vertically to the centroid.

From starting this, we received information in the cluster output that can be seen on the next page.

```

Cluster output

=== Run information ===

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.ManhattanD
Relation:    german_credit-weka.filters.unsupervised.attribute.Remove-R16,18-21
Instances:   1000
Attributes:  16
             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
             job
Test mode:   split 50% train, remainder test

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

KMeans
=====

Number of iterations: 5
Sum of within cluster distances: 5416.0051691029605

Initial starting points (random):
Cluster 0: 'no checking',36,'critical/other existing credit','new car',7855,<100,1<=X<4,4,'female div/den/mar',none,2,'real estate',25,stores,own,skilled

```

Figure 6 - Cluster Output Part 1

```

Cluster output

Cluster 0: 'no checking',36,'critical/other existing credit','new car',7855,<100,1<=X<4,4,'female div/den/mar',none,2,'real estate',25,stores,own,skilled
Cluster 1: <0,24,'critical/other existing credit','used car',6615,<100,unemployed,2,'male single',none,4,'no known property',75,none,'for free','high qualif/sel:

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute          Full Data          Cluster#          0          1
                   (1000.0)          (615.0)          (385.0)
=====
checking_status     no checking         no checking         <0
duration            18                 18                 18
credit_history       existing paid        existing paid        existing paid
purpose             radio/tv            radio/tv            new car
credit_amount        2319.5              2323               2303
savings_status       <100                <100                <100
employment           1<=X<4              1<=X<4              >=7
installment_commitment 3                   3                   4
personal_status      male single          male single          male single
other_parties        none                 none                 none
residence_since      3                   2                   4
property_magnitude   car                  car no known property
age                  33                  31                  38
other_payment_plans  none                 none                 none
housing              own                  own                  own
job                  skilled              skilled              skilled

Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on test split ===

KMeans
=====

```

Figure 7 - Cluster Output Part 2

```

Clusterer output
Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on test split ===

kMeans
=====

Number of iterations: 3
Sum of within cluster distances: 2855.6865516150624

Initial starting points (random):

Cluster 0: >=200,12,'delayed previously','new car',2247,<100,1<=X<4,2,'female div/dep/mar',none,2,car,36,stores,own,skilled
Cluster 1: 0<=X<200,42,'critical/other existing credit',business,5954,<100,4<=X<7,2,'female div/dep/mar',none,1,'real estate',41,bank,own,'unskilled resident'

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute                                Full Data                                Cluster#                                0                                1
                                      ($00.0)                                (360.0)                                (140.0)
-----
checking_status                        no checking                            no checking                            0<=X<200
duration                             18                                    18                                    20.5
credit_history                        existing paid                            existing paid critical/other existing credit
purpose                             radio/tv                                radio/tv                                radio/tv
credit_amount                         2509                                    2333                                    3108
savings_status                       <100                                    <100                                    <100
employment                           1<=X<4                                1<=X<4                                4<=X<7
installment_commitment                3                                        3                                        3
personal_status                       male single                            male single                            male single
other_parties                         none                                    none                                    none
residence_since                       3                                        3                                        3
property_magnitude                    car                                    car                                    real estate
age                                   34                                    33                                    36
other_payment_plans                   none                                    none                                    none
housing                               own                                    own                                    own
job                                    skilled                                skilled                                unskilled resident

Time taken to build model (percentage split) : 0 seconds

Clustered Instances

0    341 ( 68%)
1    159 ( 32%)

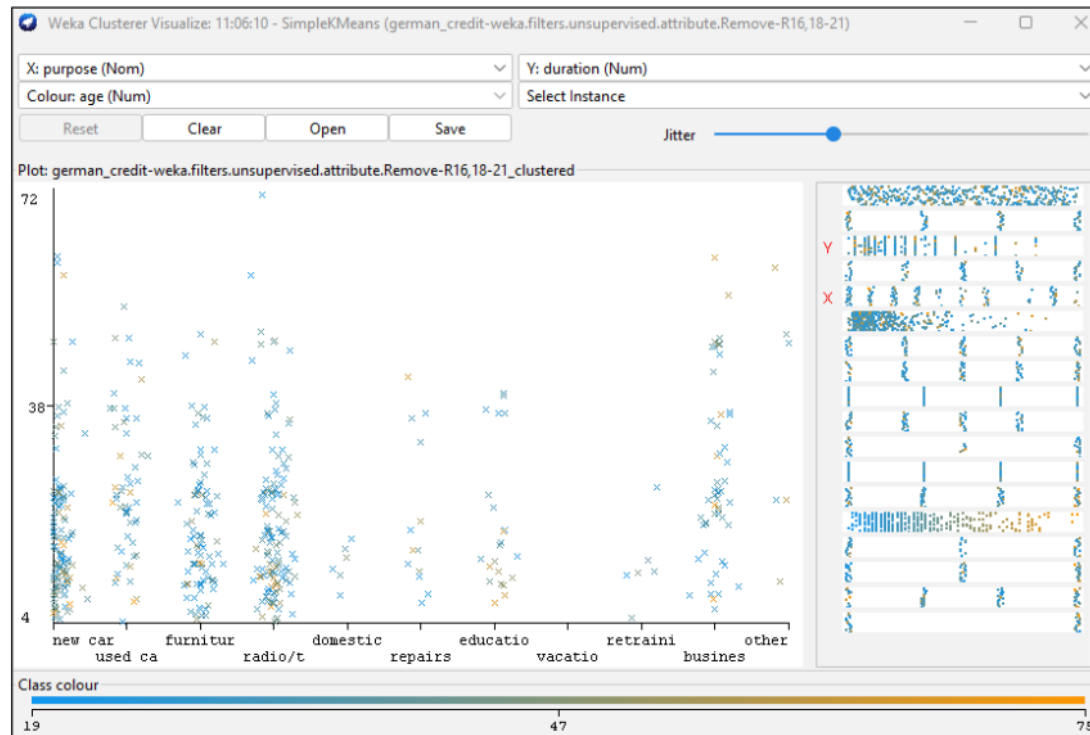
```

Figure 8 - Cluster Output Part 3

From the previous three cluster output screenshots you can see that the data and the attributes were used. This was split 50/50 for training and testing. There was five iterations of the training set and 3 iterations for the testing set. These were then split into two clusters with a number of different attributes for each of the clusters within the training and testing.

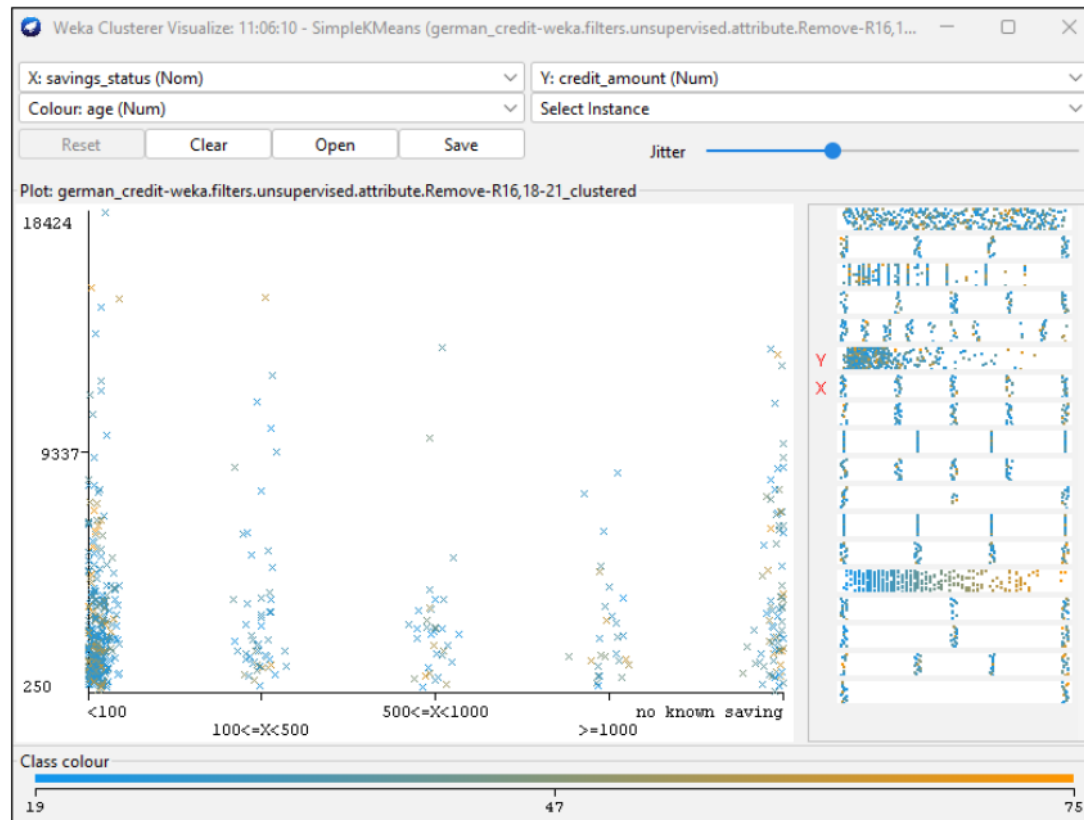
To visualise this data better we can click on the cluster within the results list and select 'Visualise cluster'

Loan Purpose over Time Duration for Different Ages



The image above shows the types of loans that are being taken out and the time periods that they are for starting at 4 months to 72 months. These loans have been taken out for a variety of varied reasons and by customers of a variety of different ages. We can see from this that there are less loans being made by people over the ages of forty-seven, with the majority of points being a shade of blue. There are also a larger number of loans being taken out for longer than 38 months by people under the age of forty-seven. A high majority

Savings Status and Credit Amounts for Different Ages



From the arrangement above we can see that there is dominant number of people getting loans of less than 9337, and more so less than 4500 while they have a credit amount of less than one hundred. This can be seen to be people under the ages of forty-seven majority and less instances where the customer is above forty-seven. To see some further examples of these arrangements. (See Appendix 1.1 - 1.3)

DBSCAN Algorithm Cluster

Density based clustering is another algorithm that can be used on Weka. This algorithm is used to identify clusters without the need of giving the number of clusters that are desired to be used. This is done automatically and based on the density of the data within the dataset that we are using. This is good as we will not know for sure the number of clusters to use so this algorithm will do that to what it needs.

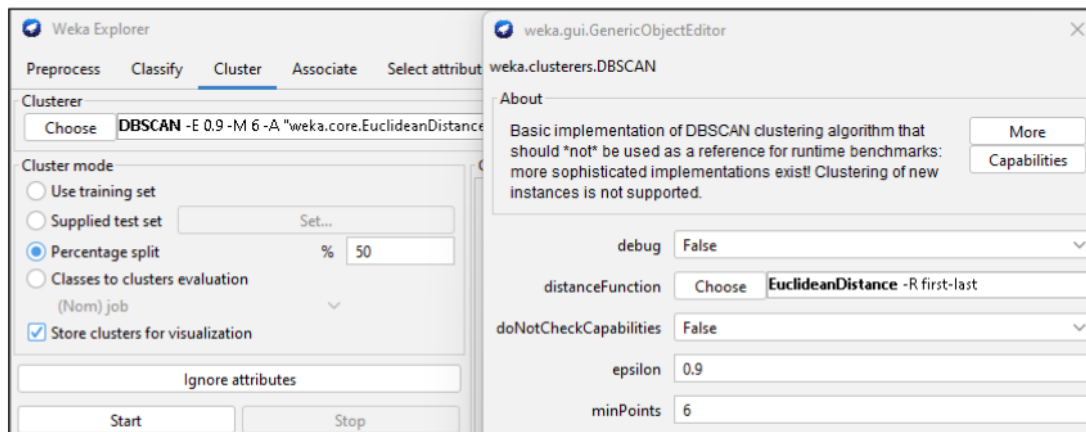


Figure 9 – DBSCAN Object Editor

Above you can see the selections that have been made before starting this algorithm. The distance function being used is the Euclidean Distance and this is the measure of a straight-line between two different points which then identifies a cluster based on the density of points. When these points are within a dense region or distance, they will then be considered parts of the same cluster.

```

Clusterer output
=== Run information ===

Scheme:      weka.clusterers.DBSCAN -E 0.9 -M 6 -A "weka.core.EuclideanDistance -R first-last"
Relation:    german_credit-weka.filters.unsupervised.attribute.Remove-R16,18-21
Instances:   1000
Attributes:  16
              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
              job

Test mode:   split 50% train, remainder test

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

DBSCAN clustering results
=====

Clustered DataObjects: 1000
Number of attributes: 16
Epsilon: 0.9; minPoints: 6
Distance-type:
Number of generated clusters: 0
Elapsed time: .12

( 0.) <0,6,'critical/other existing credit',radio/tv,1169,'no known savings --> NOISE
( 1.) 0<=X<200,48,'existing paid',radio/tv,5951,<100,1<=X<4,2,'female div/d --> NOISE
( 2.) 'no checking',12,'critical/other existing credit',education,2096,<100 --> NOISE
( 3.) <0,42,'existing paid',furniture/equipment,7882,<100,4<=X<7,2,'male si --> NOISE

```

Figure 10 - Cluster Output Part 1

```

(479.) 'no checking',6,'existing paid',radio/tv,1346,100<=X<500,>=7,2,'male --> NOISE
(480.) <0,45,'existing paid',radio/tv,1845,<100,1<=X<4,4,'male single',none, --> NOISE
(481.) 'no checking',48,'critical/other existing credit','used car',2751,'no --> NOISE
(482.) 'no checking',18,'critical/other existing credit',furniture/equipment --> NOISE
(483.) 'no checking',12,'existing paid',business,1076,<100,1<=X<4,2,'male ma --> NOISE
(484.) 0<=X<200,30,'delayed previously',business,2503,100<=X<500,>=7,4,'male --> NOISE
(485.) 'no checking',30,'all paid','used car',7485,'no known savings',unempl --> NOISE
(486.) <0,24,'existing paid','used car',2910,<100,4<=X<7,2,'male single',non --> NOISE
(487.) 'no checking',36,'existing paid',furniture/equipment,10974,<100,unemp --> NOISE
(488.) 'no checking',6,'existing paid',furniture/equipment,1766,<100,1<=X<4, --> NOISE
(489.) 'no checking',15,'existing paid',radio/tv,1213,500<=X<1000,>=7,4,'mal --> NOISE
(490.) <0,18,'existing paid','used car',7511,'no known savings',>=7,1,'male --> NOISE
(491.) <0,24,'all paid',furniture/equipment,3349,500<=X<1000,<1,4,'male sing --> NOISE
(492.) 0<=X<200,60,'delayed previously',radio/tv,9157,'no known savings',1<= --> NOISE
(493.) 0<=X<200,18,'existing paid',repairs,3872,<100,unemployed,2,'female di --> NOISE
(494.) 'no checking',12,'critical/other existing credit',business,2292,<100, --> NOISE
(495.) 'no checking',24,'critical/other existing credit',repairs,5507,<100,> --> NOISE
(496.) 'no checking',12,'critical/other existing credit',business,1412,<100, --> NOISE
(497.) 'no checking',15,'critical/other existing credit','used car',3368,>=1 --> NOISE
(498.) 'no checking',36,'critical/other existing credit',furniture/equipment --> NOISE
(499.) >=200,6,'delayed previously',radio/tv,683,<100,<1,2,'female div/dep/m --> NOISE

Time taken to build model (percentage split) : 0.03 seconds

Clustered Instances

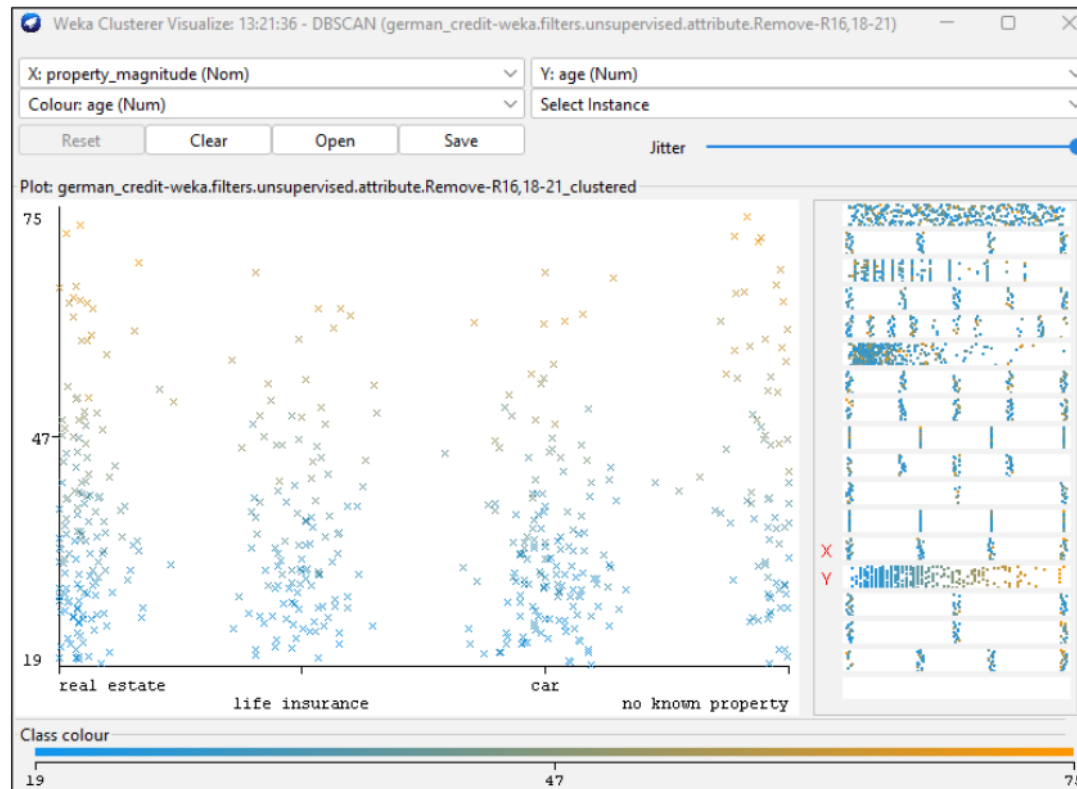
Unclustered instances : 500

```

Figure 11 - Cluster Output Part 2

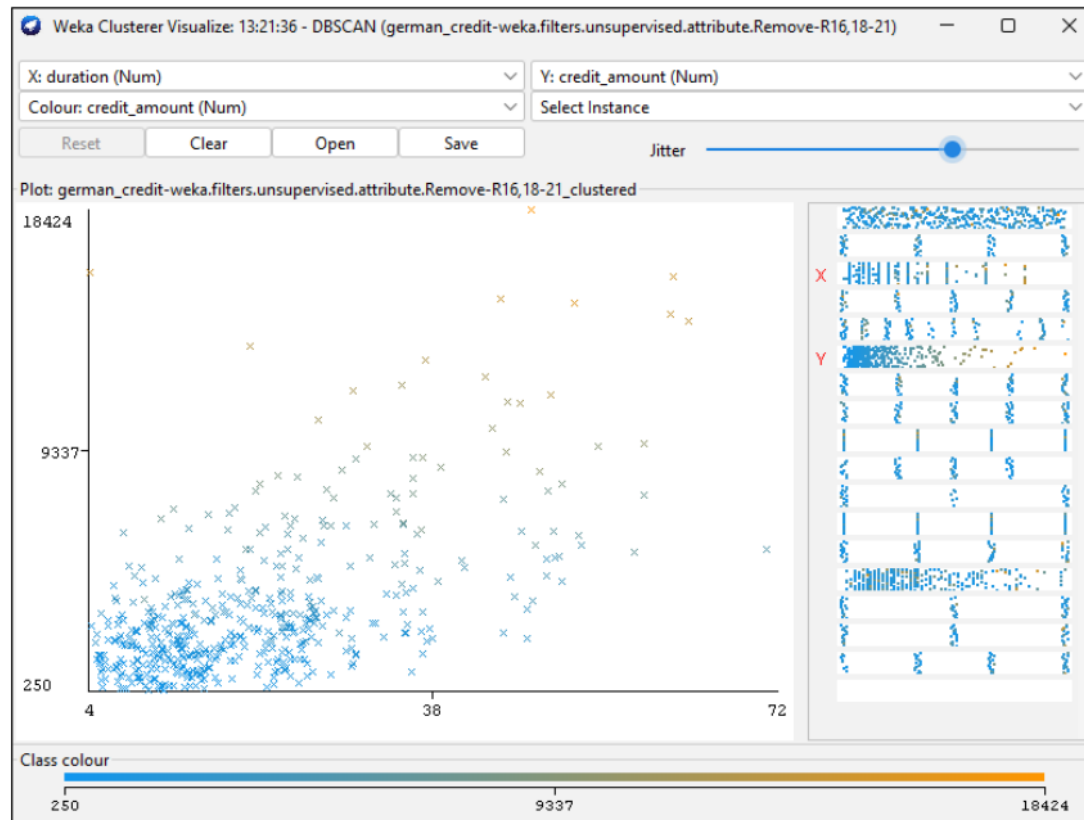
From these two output screenshots we can see that the dataset was split 50/50 with five hundred of the instances being used for training and then testing. We can now see these cluster arrangements by clicking on the recent DBSCAN in the results list and Visualise cluster arrangement. The next page will highlight one of these clusters.

Property Magnitude and Customer Age



The image above highlights the different ages of the customers compared to the property magnitude. We can see that customers ages 19-47 are more likely to get a loan for real estate, life insurance, and cars. As the ages rise of the customers there seems to be less customers for each of these areas. However, the occurrence of people younger and older than forty-seven for no known property appear to be even if not just more so for the older customers.

Credit Amount for Durations



We can see from the visualisation above that the lower loans taken out by people are for less time than the larger loans. For example, loans for less than five thousand will be for around 30-38 months or less on the majority, where there is a lot of loans that appear above 9000 that are longer than 38 months.

Business recommendations

This section will outline potential reasons as to why people may or may not be able to pay back the loans that they owe. The first reason as to why people may be able to pay their loan back is the amounts owed and the time duration they must pay the money back. This should suggest a likelihood of paying the money back as it was found that many smaller amount loans were for a shorter duration of time, this would suggest that the individuals would be able to pay this amount back and the customers on a longer duration loan would have the time and opportunity to save the amount required or pay off the amount owed over the time that they have. This could however be a risk as it was discovered that some individuals on long term loans have a lot less money within their savings suggesting that even though these people are on a longer loan duration they will have less money being earned suggesting a number of customers may not be able to afford each of the payments over the time of the loan.

Furthermore, the type of reasons that people are getting loans could also affect the likelihood of the customer paying the loan back in time. For example, it was found that a lot of the short loans were for vehicles which were new and used cars, as well as furniture and radios/televisions. With some of the longer duration loans being because of business loans. This could suggest that these loans might be able to be paid off in time, since having a vehicle may provide the person to go to their work of employment, they would be making more money to have more money to go towards the repayment of the loan. This could also suggest the business loans might be able to be paid off since they would hopefully be making more money than the loan is costing them with their businesses.

Another reason people might be able to pay back their loans is because of their relationship status. It was found that single males and females that were div/ dep/ mar had the majority of loans suggesting that if the individuals are paying the loan themselves, they might have less money outcoming in other aspects of their lives. However, this could be a negative and suggest that they might have less money to pay back the loan in time as they are paying the loan on their own and not with the help of a guarantor or co applicant to help cover then payments.

Reflection

The use of Weka for the analytical process, particularly through its clustering algorithms, proves highly efficient in creating clusters that may outline factors contributing to customers loan repaying abilities. Before getting to the stage of using these effective and efficient features, the preprocessing and postprocessing stages require particular diligence and investigation to provide the software with the required data that is understandable and usable. Providing a better outcome to more reliable and accurate results, suggesting how important this stage can be before any analysis is conducted. Then the task of finding the correct algorithms and parameters to train and assess the data to produce results. Providing opportunities for the analysis to be perfected through the addition of other clusters, distance functions, and seeds to see the differences in results.

If this process were to be repeated, more focus would be placed on the classifiers, rather than just the clusters. Although investigated but not stated this could have provided further insight into the data and outline potential reasons for the loans to be repaid intime or late by customers. Additionally, further testing and adaptation of the dataset and its training and testing data would be beneficial. As well as other ventures into other attributes and their potential impacts.

References

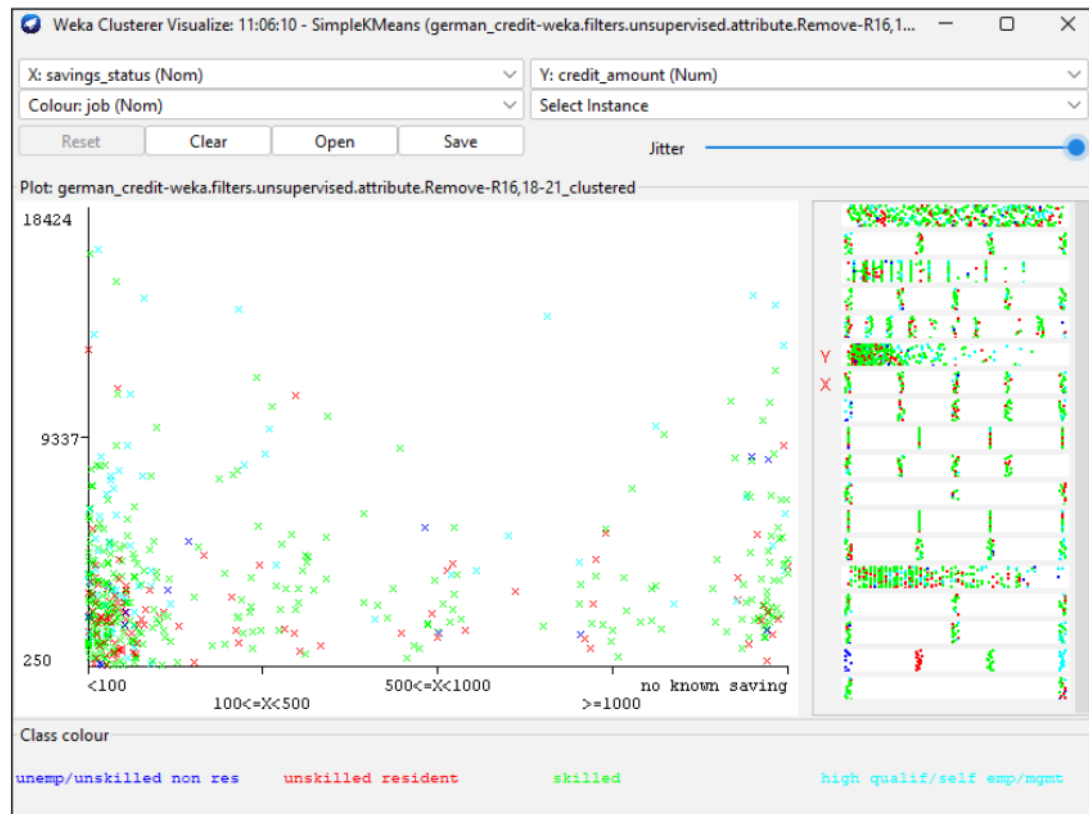
- [1] A. Twin, "What is data mining? how it works, benefits, techniques, and examples," Investopedia, [ONLINE], Available: <https://www.investopedia.com/terms/d/datamining.asp>
- [2] R. Jones, "Benefits of Data Mining", Edq.com, [ONLINE], Available: <https://www.edq.com/blog/3-positive-side-effects-of-data-mining-in-washington/>
- [3] A. S. Gillis, C. Stedman, and A. Hughes, "What is data mining? Definition from TechTarget," Business Analytics, [ONLINE], Available: <https://www.techtarget.com/searchbusinessanalytics/definition/data-mining#:~:text=Data%20mining%20is%20the%20process,make%20more%20informed%20business%20decisions.>
- [4] Simplilearn, "What is data mining? A complete guide: Simplilearn," Simplilearn.com, [ONLINE], Available: <https://www.simplilearn.com/what-is-data-mining-article>
- [5] COCUS, "Process mining: Optimize business processes with your data," COCUS, [ONLINE], Available: <https://www.cocus.com/en/process-mining-optimize-business-processes/#:~:text=Process%20mining%20belongs%20to%20the,of%20monitoring%20and%20optimizing%20them.>
- [6] R, "German credit data.," R, [ONLINE], Available: <https://search.r-project.org/CRAN/refmans/rchallenge/html/german.html>
- [7] S. Anunaya, "Data preprocessing in Data Mining: A Hands On Guide," Analytics Vidhya, [ONLINE], Available: <https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/>
- [8] S. Anunaya, "Data preprocessing in Data Mining: A Hands On Guide," Analytics Vidhya, [ONLINE], Available: <https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/>
- [9] Tableau, "Guide to data cleaning: Definition, benefits, components, and how to clean your data," Tableau, [ONLINE], Available: <https://www.tableau.com/learn/articles/what-is-data-cleaning>
- [10] Qlik, "What is Data Integration? definition, examples & use cases," Qlik.com, [ONLINE], Available: <https://www.qlik.com/us/data-integration>
- [11] Unacademy, "Data reduction techniques," Unacademy, <https://unacademy.com/content/csir-ugc/study-material/mathematical-sciences/data-reduction-techniques/#:~:text=Data%20reduction%20is%20a%20method,on%20the%20smaller%20data%20set.> (accessed May 4, 2024).

[12] Javatpoint, "Data transformation in data mining - javatpoint," www.javatpoint.com, [ONLINE], Available: <https://www.javatpoint.com/data-transformation-in-data-mining>

[13] ARFF-format, "The ARFF Data Format," The University of Edinburgh, [ONLINE], Available: <https://www.inf.ed.ac.uk/teaching/courses/dme/html/arff.html>

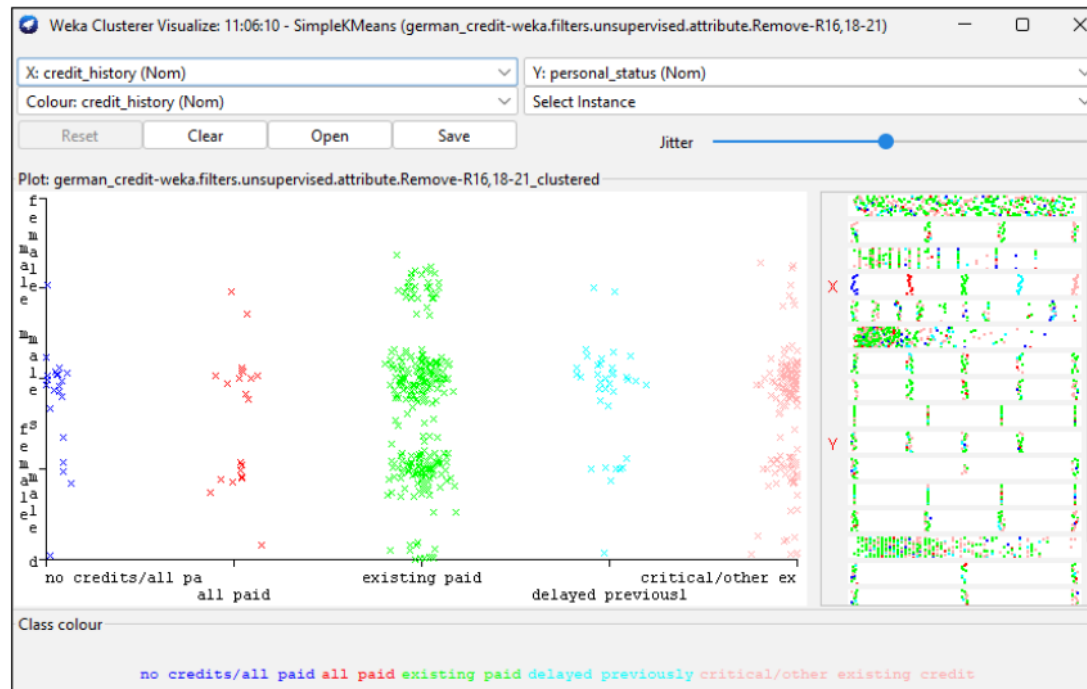
Appendix

1.1 Savings Status and Credit Amounts for Different Job Status



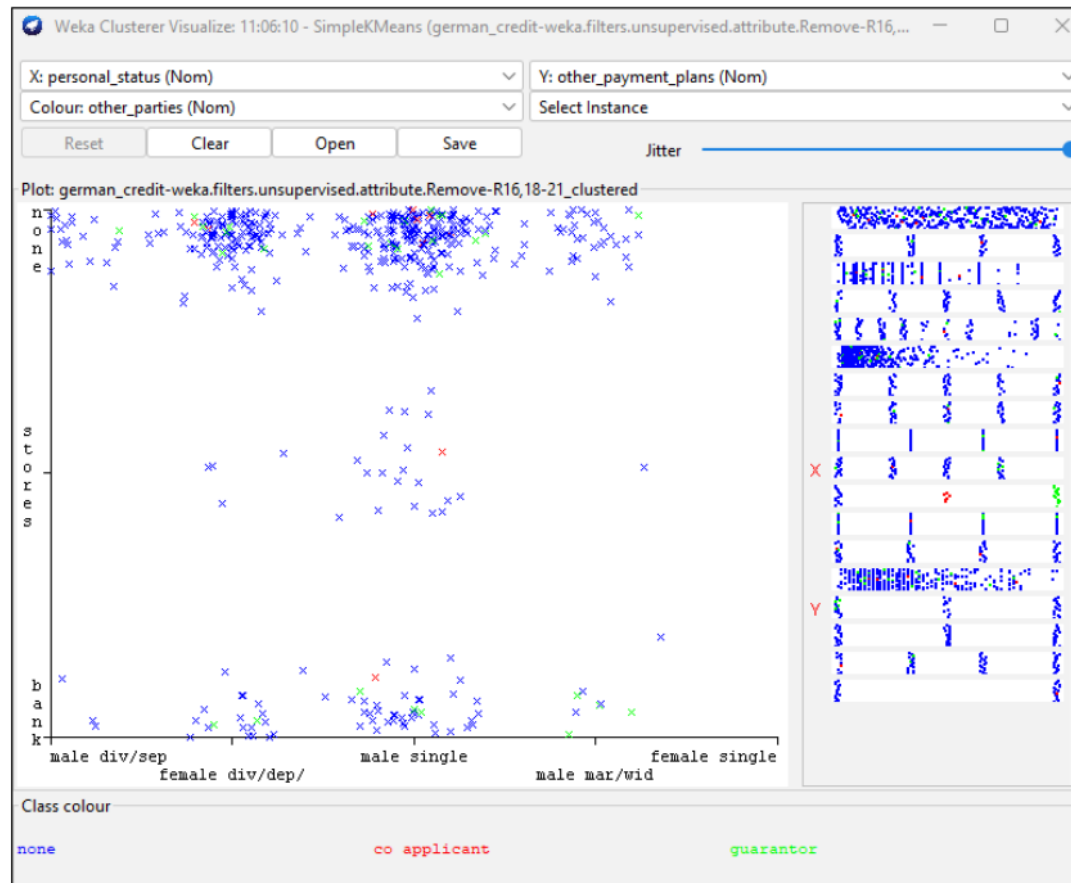
By changing the class colour to job status, we can see that a lot of the loans being taken out are from skilled and high qualified/self-employed individuals. There are several people that are unskilled and even less of people that are unemployed. The loan amount being taken out over 9337 can be seen to be taken out by high qualified/self-employed individuals.

1.2 Credit History and Personal Status for Credit History



We can see from the cluster assignment above that there is a large number of loans that are by single males and females that are div/dep/mar. These also a number of loans that are while having existing paid credit and critical/ other existing credit.

1.3 Personal Status and Other Payment Plans with Other Parties



From the above we can see that a lot of the loans being taken out are when they have no other payment plans, these are majority from the single males and female div/dep/mar. We can also see that there is considerably less with co applicant and guarantors. However, these are seen more through the bank loans and with 'none' payment plans.

Data_Mining_Ashleigh_Simons

GRADEMARK REPORT

FINAL GRADE

GENERAL COMMENTS

56/100

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

