

2014801 - DM

by Dylan Thomas

Submission date: 14-May-2024 01:14AM (UTC+0100)

Submission ID: 233525300

File name: 24885_Dylan_Thomas_2014801_-_DM_565911_1610233267.docx (7.33M)

Word count: 4733

Character count: 25720

Assignment 2: Data Mining

Author: Dylan Thomas 2014801

Data Warehousing and Data Mining (ACCA6026)

School of Applied Computing

Wales Institute of Science and Art

14/05/24 Date of Submission

Contents

1. Introduction	3
1.1 What is Data Mining?.....	3
1.2 Data Set Summary.....	3
1.2.1 Data Set Description	3
1.2.2 Data Familiarisation	7
1.2.3 Dataset Summary.....	7
1.3 Pre and Post Processing 1.3.1 Primary Pre-Processing.....	7
1.3.2 Secondary Pre-Processing.....	8
1.3.3 post-Processing	9
1.4 Analysis Methods and Results	10
1.4.1 KNN	10
1.4.2 DB Scan	12
1.4.3 Random Tree.....	15
1.5 Business Recommendations	16
1.6 Reflection	17
2. Conclusion	17
3. Reference List.....	18
Bibliography	18
Figures List	19

1. Introduction

The purpose of this assignment is to explore the well-known South German Credit Dataset created by Hofmann in the 1980's, to extract valuable patterns and trends from the data. By implementing a variety of techniques and algorithms, the aim is to uncover hidden knowledge about credit risk assessment. Through these insights, and exploring the data for actionable intelligence, an expected output of providing the business with recommendations for informed decision-making can be achieved.

1.1 What is Data Mining?

According to Koul [1] *"Data mining is the process of discovering patterns in large datasets using machine learning, statistics, and database systems to extract and transform information for further use"*.

The purpose of Data Mining is to extract valuable information from large datasets to improve predictions and decision-making. Data mining involves numerous techniques such as pattern recognition, statistical analysis, and machine learning to sift through data to identify patterns and meaningful relationships all pioneered by Samuel [2].

Implementation of Data Mining practices should empower a business to make better informed decisions for various reasons like getting ahead of the competition, driving change and innovation.

Data Mining can be utilised by the business using the following techniques:

Predictive Analysis: examining past data, data mining assists the business in predicting future patterns and actions. This enables businesses to foresee consumer demand, spot potential risks, and allocate resources effectively.

Transactional Data Analysis: Businesses can determine links between products that are frequently purchased together by looking at transactional data. Strategies like promotions and product promotions can all be improved by using this data effectively.

Fraud Detection: examining trends and abnormalities in user behaviour and financial transactions, data mining techniques can be used to identify fraudulent activity and aid reducing risks and prevention tactics.

Customer Segmentation: Businesses can divide their customers into groups according to shared characteristics like shopping patterns, demographics, or preferences by evaluating customer data. This enables customised product recommendations and targeted marketing efforts possible.

For Financial data sources like the South German Credit data set [3] Gemel explores techniques and modelling implementations for financial data specifically.

1.2 Data Set Summary

1.2.1 Data Set Description

In the field of machine learning and data mining, the South German Credit dataset is frequently

utilised, especially for classification tasks. Information about credit applicants and their corresponding credit risks is contained in it. Below is a table containing a detailed description of the dataset,

Figure 1 – Data Set

Attribute	Type	Description	Value(s)	Breakdown
Status	Categorical	Status of the applicants/debtor's bank account A11: < 0 DM A12: 0<=...<200 DM A13: ... >=200 DM/ salary assignments for at least 1 year A14: no checking account	1, 2, 3, 4	1 = 274 2 = 269 3 = 63 4 = 394 No missing values
Duration	Numerical	Duration of the loan in months	4 - 72	No missing values
Credit History	Categorical	Debtors' payment history and current credit contracts 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)	0, 1, 2, 3, 4,	0 = 40 1 = 49 2 = 530 3 = 88 4 = 293 No missing values
Purpose	Categorical	What the load is needed for A40: car(new) A41: car(old) A42: furniture/equipment A43: radio/television A44: domestic appliances A45: repairs A46: education A47: vacation A48: retraining A49: business A410: others	0, 1, 2, 3, 4, 5, 6, 7, (No Loans for Vacations) 8, 9, 10	0 = 234 1 = 103 2 = 181 3 = 280 4 = 12 5 = 22 6 = 50 7 = --- 8 = 9 9 = 97 10 = 12 No missing values
Credit Amount	Numerical	Amount in DM's loaned to the applicant	250 – 18,424	No missing values
Savings	Categorical	Debtor's savings A61: ...< 100 DM A62: 100<=...<500 DM A63: 500<=...<1000 DM A64: ...>1000 DM A65: unknow/no savings account	1, 2, 3, 4, 5	1 = 603 2 = 103 3 = 63 4 = 48 5 = 183

				No missing values
Employment duration	Categorical	Duration of the debtor's employment with current employer A71: unemployed A72: ...< 1 year A73: 1 <=...4 years A74: 4 <=...< 7 years A75: ...>= 7 years	1, 2, 3, 4, 5	1 = 62 2 = 172 3 = 339 4 = 174 5 = 253 No missing values
Instalment rate	Numerical	Instalment rate in percentage of disposable income	1, 2, 3, 4	1 = 136 2 = 231 3 = 157 4 = 476 No missing values
Personal status and sex	Categorical	Combined information about the debtor's gender, marital status A91: male divorced/separated A92: female divorced/separated/married A93: male single A94: male married/widowed A95: female single	1, 2, 3, 4, No value for single female	1 = 50 2 = 310 3 = 548 4 = 92 No missing values
Other debtors	Categorical	Is there another person who is standing as a guarantor for the debt A101: none A102: co-applicant A103: guarantor	1, 2, 3	1 = 907 2 = 41 3 = 52 No missing values
Present residence since	Numerical	The length of time in years that the debtor has resided in their current property	1, 2, 3, 4	1 = 130 2 = 308 3 = 149 4 = 413 No missing values
Property	Categorical	The debtor's most valuable assets A121: real estate A122: building society savings agreement/life insurance A123: car or another asset A124: unknown/no property	1, 2, 3, 4	1 = 282 2 = 232 3 = 332 4 = 154 No missing values
Age	Numerical	Age of the debtor at the time of the loan in years	19 - 75	No missing values
	Categorical	Instalment plans from other credit plans that the debtor may have A141: bank	1, 2, 3	1 = 139 2 = 47 3 = 814

Other instalment plans		A142: stores A143: none		No missing values
Housing	Categorical	Type of housing that the debtor lives in at the time of the loan A151: rent A152: own A153: for free	1, 2, 3	1 = 179 2 = 714 3 = 107 No missing values
Number of credits	Numerical	Number of credits including current one that the debtor has	1, 2, 3, 4	1= 633 2 = 333 3 = 28 4 = 6 No missing values
Job	Categorical	Quality of the debtor's job A171: unemployed/unskilled - non-resident A172: unskilled resident A173: skilled employee/official A174: management/self-employed/highly qualified employee/officer	1, 2, 3, 4	1 = 22 2 = 200 3 = 630 4 = 148 No missing values
People liable	Numerical	The number of people who are financially reliant on the debtor	1, 2	1= 155 2 = 845 No missing values
Telephone	Categorical	Does the debtor have a landline registered in their name A191: none A192: yes	1, 2	1 = 596 2 = 404 No missing values
Foreign worker	Categorical	Is the debtor a foreign worker A201: yes A202: no	1, 2	1 = 37 2 = 963 No missing values
Credit Risk	Cost matrix	1 = good 2 = bad	1, 2	Good = 700 Bad = 300

Overall, the South German Credit dataset proves invaluable for developing and validating credit risk evaluation models within financial decision-making frameworks. It offers a representation of the data variety and challenges encountered in real-world credit assessment situations.

1.2.2 Data Familiarisation

According to Pope, [4] Data Familiarisation involves “*immersion in the raw data (or typically a pragmatic selection from the data) to list key ideas and recurrent themes*”.

Seen as the first step of the process, with the aim of understanding the dataset and gaining initial insights to the data. Understanding the structure involves recognising data attributes and instances within the dataset. Another activity involves reviewing attributes and categorising those attributes into numerical and categorical attributes (*see Figure 1*) to select suitable analysis techniques. Exploring the distribution of the data identifies the range and variability of values per attribute, possibly identifying calculations of interest for summaries of the data like mean, median, mode, standard deviation and so on.

1.2.3 Dataset Summary

Below are some points of interest noted during the initial exploration of the dataset prior to the implementation of Processing activities:

- There's a mix of numerical and categorical attributes within the dataset, representing both personal and financial information of credit applicants.
- There are females within the dataset, however there are no single females included within the dataset. Given the data was collated in the 1980's, is this a factor affected by social norms of the time?
- Some of the attributes like **Instalment Rate**, **property** and **telephone** have an even breakdown of data, while other attributes like **Foreign Worker**, **savings** and **Personal Status + Sex** have a vast distribution with some attribute values having a one-sided breakdown.
- For the Purpose attribute, there is an anomaly where there are no values for the purpose of a loan code A47 which is the vacation reason.
- There are no missing values within the dataset, with exception of the anomalies mentioned previously.

1.3 Pre and Post Processing

1.3.1 Primary Pre-Processing

According to Garca et al [5] Data preprocessing is defined as “*a step in the data mining process that involves cleaning and transforming raw data to make it suitable for analysis.*”

The file provided was one that had been pre-prepared rather than the fully raw version of the South German Credit dataset. The following are pre-processing activities that would've been undertaken on the file:

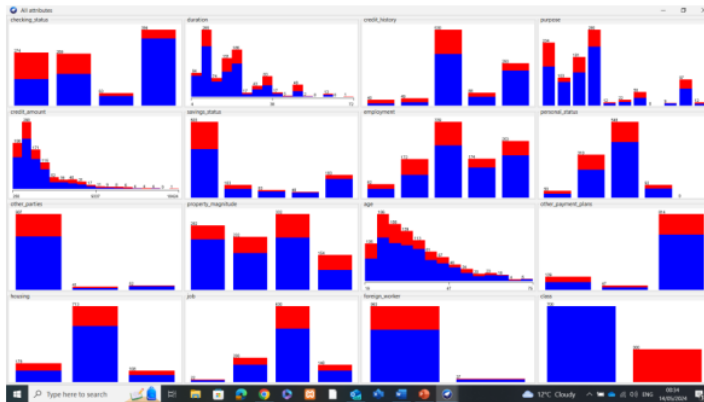
File Type Change, the file was downloaded from a zip file, saved then into a CSV file to make it usable in other software like Jupyter and Notepad++ to make it a readable format. The original document was a txt format, but for the data to be usable within the Wekka software the file type must be changed to an arff (Attribute-Relation File Format) file.

Translation, given it is a German dataset there's little surprise the headers are provided in German, so these headers must be translated to English to make it understandable. As part of the initial pre-processing the attributes/classes required translation.

Category Definition, within .arff file the data items within the dataset must be declared, and categorised into either attribute, relation, or data attributes. This is done by adding an @ symbol, followed by the classification and the data item title. By categorising the headers appropriately, it makes the data usable. The datatype is crucial for categorising numerical, date and string data format.

Data Cleansing, in this instance techniques such as mean imputation can be used, however the main activity undertaken on this dataset was removing any columns with missing values. Unnecessary on this dataset, but further cleansing activities like duplication removal and error corrections of typos or incorrect data entries could be done if applicable.

Figure 2 – Attribute visualisation.



Having landed the data in Weka, the Visualise all option allows the user to do just that. Immediately looking at this screen it allows the user to identify abnormal double peaks or trends in the data, with Purpose, Age, Credit Amount, Employment and Personal Status attracting interest.

1.3.2 Secondary Pre-Processing

Secondary Pre-processing relates to further data preparation implemented following the initial pre-processing stage but prior to the primary data mining analysis. Secondary pre-processing seeks to improve the data's applicability for certain analysis requirement or to address other issues that can arise during the modelling process. Further cleansing activities can be undertaken as part of the Secondary Pre-processing:

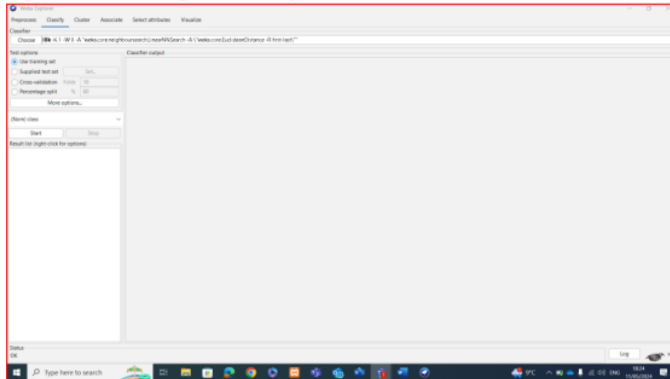
Feature Selection, identifying and choosing features that are most pertinent to the prediction task and making the necessary selections while removing features that are redundant or unnecessary. A test available within the Weka software known as Information Gain Attribute which calculates a value score for each attribute within the dataset. Those with a low value are deemed less value, within Machine Learning it is normally the variables with the largest information gain used to split the dataset. Other techniques which could be implemented include feature recursive feature elimination and univariate selection.

Dataset Splitting, prior to splitting the dataset randomising the dataset a preferable step to avoid bias within the split datasets that could skew figures. This is done within the Weka software. Once randomised the dataset can be split into Training and Test datasets, for this an 80%/20% split was

chosen. By dividing the data into a training and test dataset it helps prevent overfitting by testing on a larger training set before testing the accuracy of the model as suggested by Géron. [6]

Dataset Normalisation, ensuring features have comparable ranges or distributions by standardising the scale can enhance the convergence of optimisation methods and the effectiveness of certain models.

Figure 3 – Training Dataset Split



1.3.3 post-Processing

In data mining, the post-processing phase takes place after the main data mining process of pre-processing is complete. Its purpose is to extract significant insights or conclusions, seen as the most time intensive phase as it entails honing, analysing, and presenting the output of the selected algorithms. Post Processing activities include the following:

Model Evaluation and Refinement, metrics like the F1-Score, recall, precision, and accuracy can be used to assess how well machine learning models perform. Subsequently, the scores can be assessed to fine-tune the model hyperparameters and settings to enhance performance or fulfil certain requirements.

Additionally, combining several models or forecasts can lead to improved outcomes using techniques like model averaging, boosting, and stacking.

Result Interpretation, analysing model outputs and projections to get practical insights and make well-informed decisions is the process of interpreting model outcomes.

Emphasise the most important insights, highlight the most significant conclusions or revelations from the analysis.

Visualising the results using visual aids like charts and dashboards to present analysis results can assist comprehension and communication of findings. Summarising the key findings, results, conclusions, and takeaways for the audience at their fingertips.

Formatting Results, to assist comprehension and interpretation, the analytical results should be displayed in a structured format, such as tables, charts, or graphs. This will help to arrange the data in an understandable and easy to interpret manner. To retain visual coherence and professionalism, font and colour formatting should be consistent throughout the report.

Visualizations, creating informative visuals that both reinforce the report's narrative and successfully showcase the findings. Using the most appropriate data visualisation techniques (bar charts, scatter graphs, etc.) to make complex correlations and trends easier to understand.

1.4 Analysis Methods and Results

1.4.1 KNN

According to IBM [7] “k-nearest neighbours (KNN) algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point”.

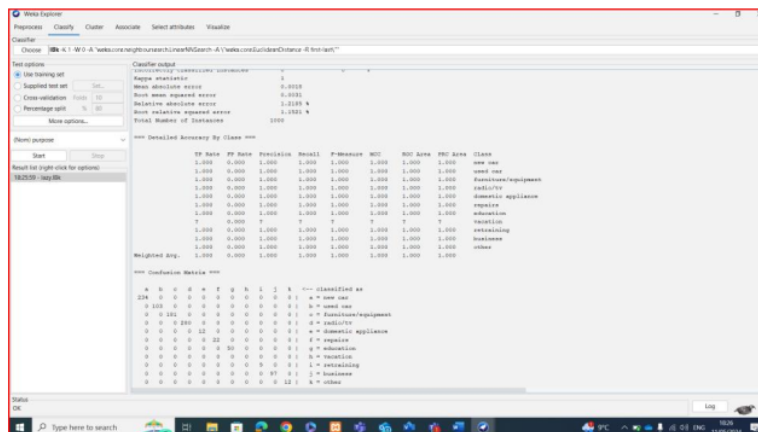
KNN was selected as a model to test the South German Credit dataset as it was recognised as one of “the top 10 data mining algorithms identified by the IEEE International Conference on Data Mining” by X.Wu et al [8] .

Simplicity, the implementation of KNN is straightforward, it involves calculating the distance between data points based on different attributes. This distance may be simply computed using a distance formula such as Euclidian.

No Training Period, KNN modelling does not involve a training period, as the data itself serves as the model that will be used for future predictions. This approach is time effective as it allows for rapid adaptation to random modelling on the existing data.

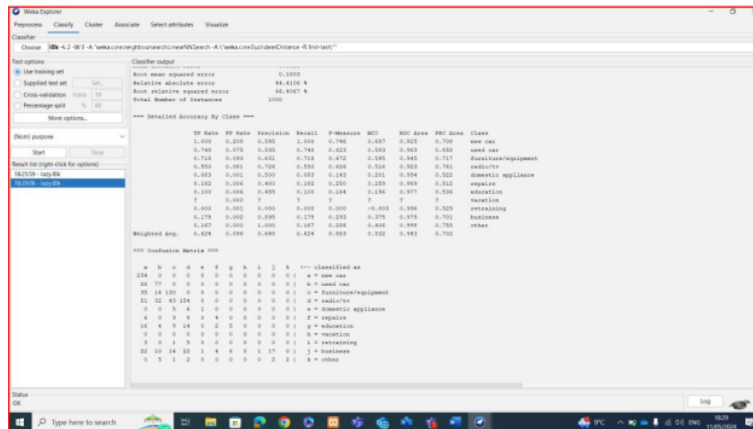
Works well with smaller dataset, Given the South German Dataset is relatively small, the negative of KNN not scaling well with larger datasets isn’t applicable.

Figure 4 – KNN Model 1



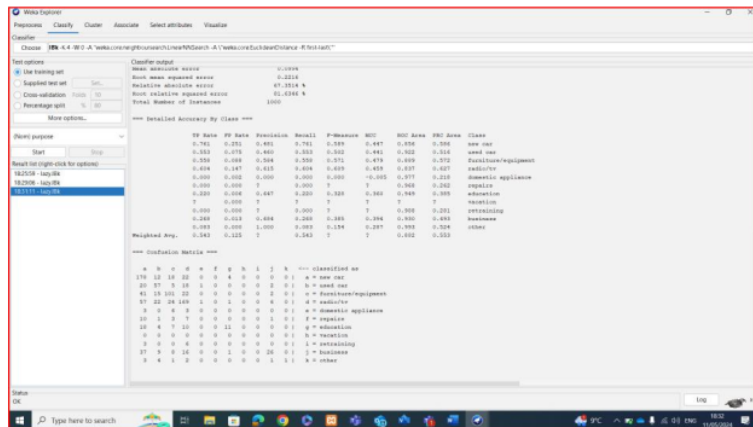
The above screenshot is of the KNN test results against a value of 1. Model 1 achieved 100% accuracy on the training data, and a very low mean absolute error (0.0018) and root mean squared error (0.0031). The high accuracy is plain to see from the screenshot with values following the straight line with few outliers.

Figure 5 – KNN Model 2



The above screenshot is of the KNN test results against a value of 2. Model 2 achieved 62.4% accuracy on the training data, and a higher mean absolute error (0.0655) and root mean squared error (0.1803). Varied performance across classes, with precision, recall, and F-measure ranging from 0 to 1.

Figure 6 – KNN Model 3



The above screenshot is of the KNN test results against a value of 4. Model 3 achieved 54.3% accuracy on the training data, and an even higher mean absolute error (0.0994) and root mean squared error (0.2216). Like Model 2, Model 3 had varied performance across classes, with precision, recall, and F-measure ranging from 0 to 1.

Overall, Model 1 still outperforms Models 2 and 3 in terms of accuracy and error metrics, demonstrating perfect classification accuracy and lower error rates. Models 2 and 3 show lower accuracy and higher error rates, indicating decreased performance compared to Model 1.

In terms of accuracy and error metrics, Model 1 outperforms the other Models, showing flawless classification accuracy and reduced error rates. In comparison, Models 2 and 3 showcase less accuracy and higher mistake rates, indicating poorer performance.

Figure 7 – KNN Visualisation (Purpose vs Credit Amount)



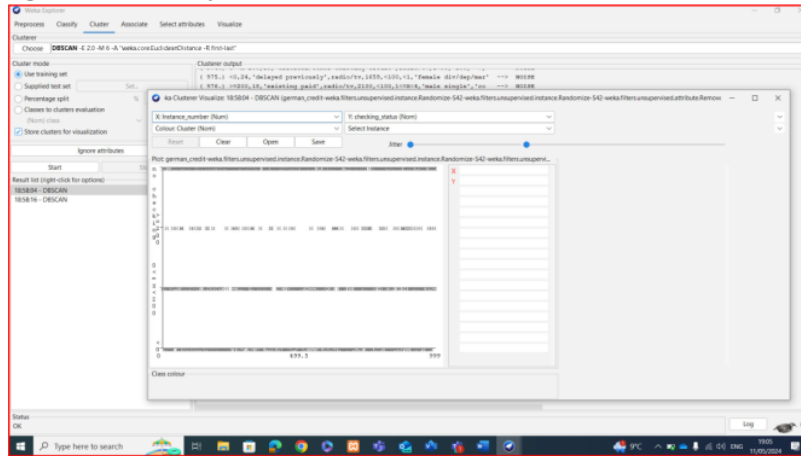
Having trained the Model, the results were visualised using the Weka software to create the above scatter graph. This visual compares the Loan Purpose against the credit amount requested. Looking at the visual at first glance, it's apparent that many loans are below 9,400 Deutschemarks, with a large amount of loans for purchasing cars and business reasons. An expected anomaly being Vacation, as the business rules don't permit loans for a vacation.

The Other purpose is an interesting visual, due to the low number of loans, and the spread of value, with the largest requested value of 18,424 Deutschemarks being requested within this reason. It would be interesting to understand the true purpose of these loans given the vast spread, and the most expensive request being far greater than the others. Repairs has an anomaly almost double the value of the next value, and the only one over 9,400, this is another loan of interest for potential further interrogation.

1.4.2 DB Scan

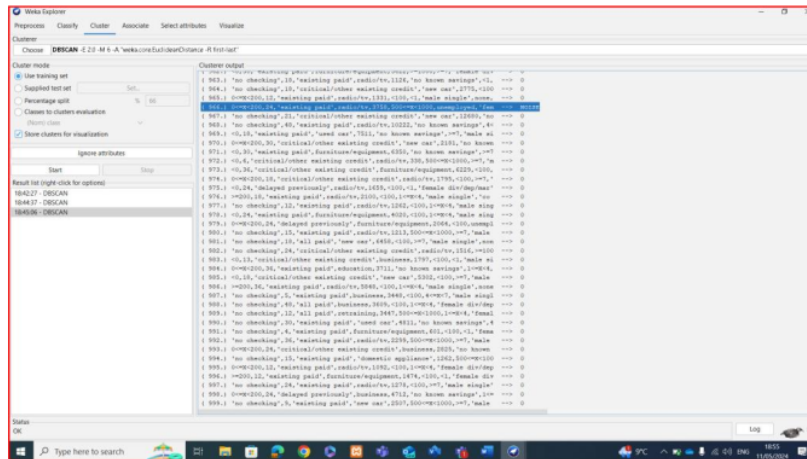
Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is an effective clustering model for application on the dataset due to its robustness to NOISE, improving data quality. The method is flexible with its cluster shapes, irregularities making it useful for complex datasets along with the determination of cluster density meaning minimal input required from the Experimenter. These benefits are only realised if the correct parameter is used, Epsilon for example which is the variable in the model yielding different results when experimenting with the figure. Due to the smaller size of the dataset, it removes the potential negative impact of the model being computationally intensive. The DBSCAN model is championed by Schubert et al [9]

Figure 8 – DB Scan Epsilon value of 0.5



The above screenshot shows the visual output from the testing of the DB Scan with an Epsilon value of 0.5, due to the whole data set resulting in unclustered values there is very little MI that can be taken from this model.

Figure 9 – DB Scan Epsilon value of 2

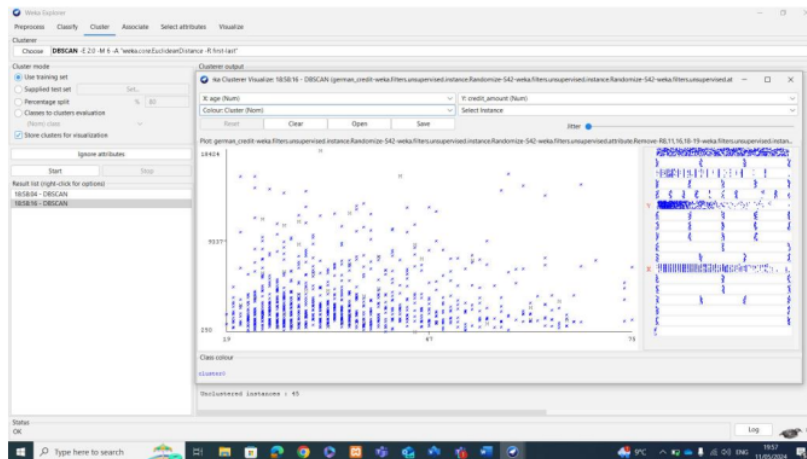


The above screenshot is the results of a DBScan with an Epsilon value of 2, resulting in 955 Clustered, unclustered 45 instances.

Within this screenshot, the last visible outlier can be seen, labelled with NOISE while the clustered values have a value of 0. This outlier is of interest as the customer is unemployed, requesting a 24-month loan to the value of 3,758 Deutschemarks. The purpose of the loan for a tv/radio must be questioned, does an unemployed individual really need to spend a large amount on media when there would be other necessities especially while unemployed? It raises questions for the other 44 instances, is there further steps that should be taken to investigate these unclustered instances?

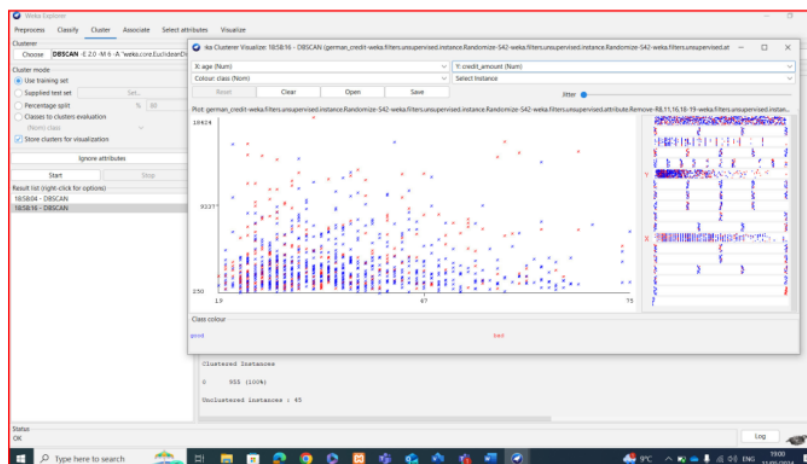
In summary, the results of both model variations show the dataset's significant clusters could not be found by the DBSCAN algorithm with the specified settings. The reporting of clustered instances is where the differences only exist.

Figure 10 – DBScan Visual



The above figure provides a visual representation of the age of those individuals taking out loans against the credit value. As expected, most of the loans are to the left of 47 years old, given in later life people should be more affluent having worked for longer and other variables like inheritance and life experience. There are some outliers to this trend with some individuals in their 70's taking out loans, but of a lower value. The data follows the expected trend of the younger applicants being more likely to fail to repay the loan within the expected timescales, though there are far more outliers to the right of 47 years old, and for low value loans which is unexpected.

Figure 11 – Class Colour Applied

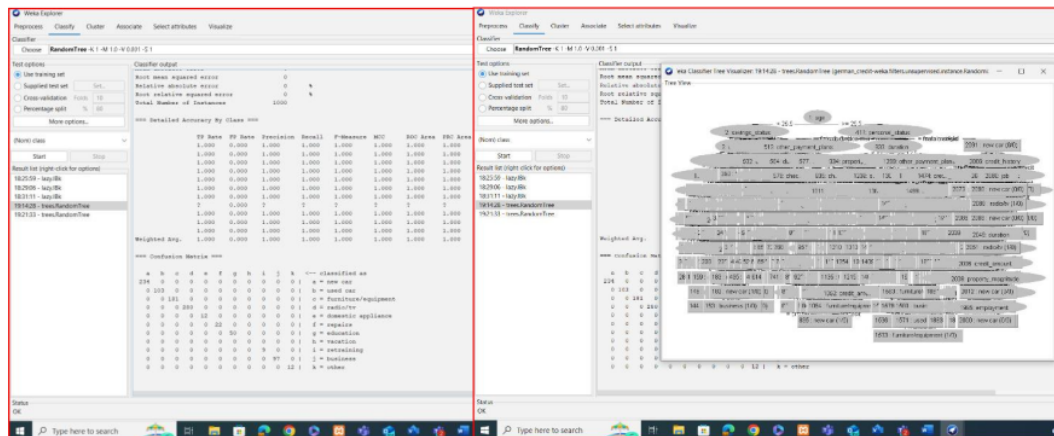


The above figure is the same visual representation as the previous figure; however, the bad loans are now visualised with a red cross rather than an M icon. This view is more uniform using the same symbols, but a differing colour for categorising.

1.4.3 Random Tree

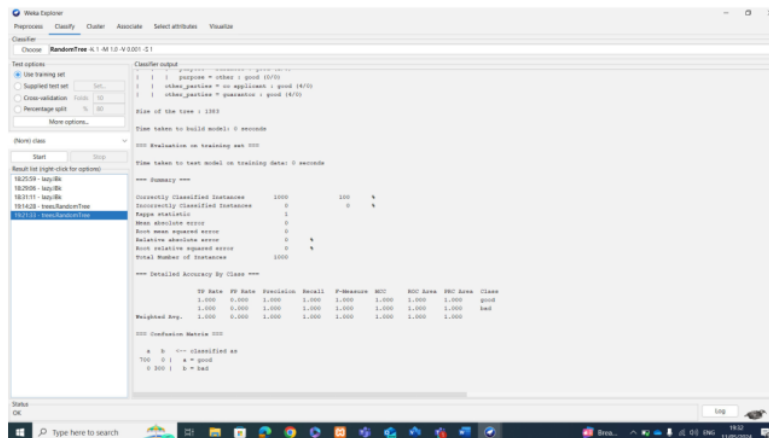
The Random Tree classifier is a machine learning algorithm that is a member of the tree-based ensemble methods. It is a variation on the Random Forest approach, which raises the model's overall accuracy and robustness by combining the predictions of several different decision trees. The Random Tree classifier compared to a single decision tree, lowers the risk of overfitting by integrating the predictions of numerous trees trained on various subsets of data. This enhances the model's ability to generalise on untested data. It is relatively efficient, and could scale if the dataset were to increase, while being robust to noisy data and outliers mitigating the impact on results.

Figure 12 – Random Tree Prior to Normalisation



The above shows the visual results of the Random Tree classifier without normalising the dataset. The output provided is interesting visually, however it holds no value as MI for the business.

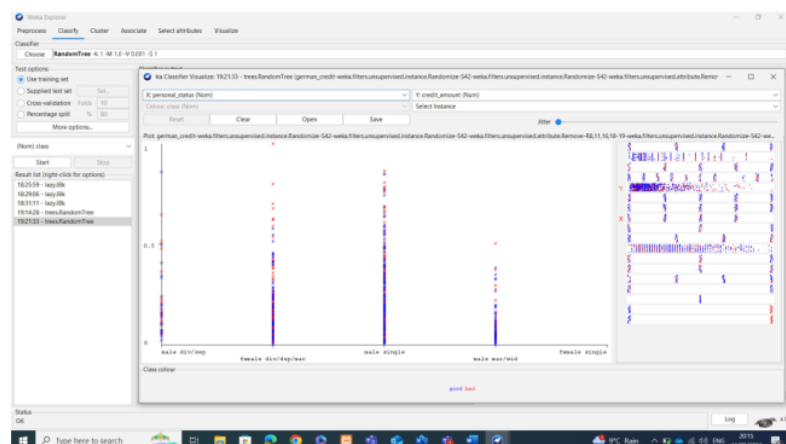
Figure 13 – Normalised Random Tree Results



The above are updated results following the normalisation of the dataset, using the option provided within Weka. Immediately the results following normalisation are more useful with the only similarity being the highly accurate confusion matrix score, along with the provision of accuracy

metrics like the F-Measure. This confusion matrix differs to the previous with values classified into good and bad values.

Figure 14 – Random Tree Visual Personal Status vs Credit Amount



The above figure is a comparison of the personal status and credit amount variables. Again highlighting the age of the dataset, the lack of single female applicants does skew overall figures. It does make for interesting reading, following an expected pattern of male divorced or oap and male married/widowed making up fewer requested loans, likely due to age and other social factors. Single males unsurprisingly seem to make most of the loan requests, having the most condensed colouration. The Female div/dep/mar figures are interesting, with highly condensed figures, however this category is home to the largest requested loan credit amount, which is c=somewhat a surprise.

1.5 Business Recommendations

Following the dataset familiarisation and the implementation of various models and their outputs, the below recommendations have been identified:

Revisit the threshold costs of loans offered, those above 9,400 deutschmarks seem to have a higher percentage of being a bad loan. To lower the percentage of Bad loans the bank may wish it to consider 9,400 deutschmarks as the ceiling figure or implement extra manual steps of considerations for loans above the suggested threshold.

Like the above recommendation, implementing **variable thresholds** with possible human interventions where necessary based on loan purpose and the employment status of the customer. The employment status has a large impact on the ability to repay a loan, as such high value loans should again have due diligence checks implemented to deeply consider whether this loan will be paid back in full on time. Similarly, the purpose should have a bearing on the value being approved.

The final recommendation, improve the accuracy and value provided by the results of the investigations, a far larger dataset with improved balance and more recent data would provide more value.

1.6 Reflection

The models implemented have provided interesting results, visuals, and provided recommendations which should allow better results for future experiments using the dataset. The accuracy of the models, depending on the variable figures used were fairly accurate, this could be due to the dataset being relatively small, with little variety in comparison to other available datasets for experimentation.

The models implemented, after experimentation with variable figures provided fairly accurate results, however this is certainly an area of improvement if this experiment were conducted again. The models used have covered both clustering and classified methods, which was important to provide a comparison. For the Random Trees model, and somewhat for the other algorithms implemented further time and exploration should've been allocated to exploration, this could've provided some further insights to the data.

If this experiment was implemented again different models should be explored, Neural Networks, in particular Convolutional Neural Networks (CNN) for other classification methods due to its flexibility and ability to capture complex patterns within the dataset. Having implemented Random Trees due to familiarity with the method, this could be superseded by Random Forest if attempted again, although the lack of missing values and dimensionality in the data made the random tree option more favourable.

Overall, the selected models provided visuals and interesting results/visuals, with various discussion points which could be interpreted. Further experimentation could have provided further discussion points though.

2. Conclusion

For this experiment on the South German Credit dataset DBSCAN, Random Forest, and K-Nearest Neighbours (KNN) were the models implemented. KNN performed well in categorising credit applicants by using its closest neighbours as a prediction. DBSCAN successfully located dense areas in the data, which was helpful in locating outliers, while Random Tree provided useful insights to the data comparing personal_status and credit_amount. Overall, these models provide a variety of methods for evaluating creditworthiness, accommodating various characteristics, and offering perspectives on risk evaluation in financial applications.

3. Reference List

Bibliography

- [1] K. Rupashi, "Overview of Data Mining. International Journal of Trend in Scientific Research and Development," 2020. [Online]. Available: <https://typeset.io/papers/overview-of-data-mining-5a7vb846qs> [Accessed 30 April 2024].
- [2] J. McCarthy, "Arthur Samuel: Pioneer in Machine Learning," 1 September 1990. [Online]. Available: <https://onlinelibrary-wiley-com.ezproxy.uwtsd.ac.uk/doi/full/10.1609/aimag.v11i3.840?sid=vendor%3Adatabase>. [Accessed 7 May 2024].
- [3] S. L. -. M. S. R. Gernel, "Data Mining Techniques and Modelling for Financial Analysis," 15 November 2022. [Online]. Available: <https://typeset.io/papers/data-mining-techniques-and-modelling-for-financial-analysis-3dbd2wkm>. [Accessed 30 April 2024].
- [4] C. Pope and S. Ziebland, "Analysing Qualitative Data," 8 January 2000. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1117368/>. [Accessed 30 April 2024].
- [5] L. J. ., H. F. García Salvador, "Data Preprocessing in Data Mining," 2015. [Online]. Available: <https://typeset.io/papers/data-preprocessing-in-data-mining-cn6pokytttd>. [Accessed 05 May 2024].
- [6] A. Geron, "Hands-on Machine Learning with Scikit-Learn, Keras + TensorFlow," 18 September 2019. [Online]. Available: <https://ebookcentral.proquest.com/lib/trinity/reader.action?docID=5892320> . [Accessed 8 May 2024].
- [7] Anonymous, "What is the KNN Algorithm?," January 2023. [Online]. Available: [https://www.ibm.com/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20\(KNN,used%20in%20machine%20learning%20today.](https://www.ibm.com/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20(KNN,used%20in%20machine%20learning%20today.) . [Accessed 7 May 2024].
- [8] X. e. a. Wu, "Top 10 Algorithms in Data Mining," 04 December 2007. [Online]. Available: <https://link.springer.com/article/10.1007/s10115-007-0114-2> . [Accessed 8 May 2024].
- [9] E. e. a. Schubert, "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN," 31 July 2017. [Online]. Available: <https://dl-acm-org.ezproxy.uwtsd.ac.uk/doi/abs/10.1145/3068335>. [Accessed 9 May 2024].

Figures List

Figure 1 – Dataset Description

Figure 2 – Attribute visualisation.

Figure 3 – Training Dataset Split

Figure 4 – KNN Model 1

Figure 5 – KNN Model 2

Figure 6 – KNN Model 3

Figure 7 – KNN Visualisation (Purpose vs Credit Amount)

Figure 8 – DB Scan Epsilon value of 0.5

Figure 9 – DB Scan Epsilon value of 2

Figure 10 – DBScan Visual

Figure 11 – Class Colour Applied

Figure 12 – Random Tree Prior to Normalisation

Figure 13 – Normalised Random Tree Results

Figure 14 – Random Tree Visual Personal Status vs Credit Amount

FINAL GRADE

GENERAL COMMENTS

74/100

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

