

DWDM Assignment 2 (Nathan Kyle Hudson)

by Kyle Hudson

Submission date: 21-May-2024 02:28PM (UTC+0100)

Submission ID: 234555342

File name: 49924_Kyle_Hudson_DWDM_Assignment_2__Nathan_Kyle_Hudson__565911_2145246748.docx
(3.11M)

Word count: 4532

Character count: 26078

Data Warehousing and Data Mining

Data Mining

ACCA6026

By
Nathan Kyle Hudson

2017102

Table of Contents

1. Introduction	4
1.2 Data Mining Purpose and Practices	4
1.3 How Data Mining can be used in Business	4
2. South German Credit Dataset.....	5
3. Processing	6
3.1 Primary Pre-processing.....	6
3.2 Secondary Pre-processing.....	8
3.3 Post-processing.....	9
4. Sample Runs	10
4.1 Random Forest.....	11
4.2 Random Trees.....	13
4.3 Naïve Bayes.....	15
4.4 Logistic	17
4.5 Bayes Net.....	19
4.6 KNN.....	21
5. Results	22
5.1 Random Forest.....	23
5.2 Random Trees.....	25
5.3 Naïve Bayes.....	27
5.4 Logistic	29
5.5 Bayes Net.....	31
5.6 KNN.....	33
5.7 Classification Results	34
6. Business Recommendations	35
7. Reflection	35
7.1 Does it work well?	36
7.2 Advantages and Disadvantages	36
7.3 Different Approaches (if any).....	36
7.4 Conclusion.....	36
8. Bibliography	38
9. Appendix	40
9.1 Full list of Attributes	40

Table of Figures

Figure 1. South German Credit dataset	5
Figure 2. Attributes named in dataset	7
Figure 3. Filtered and sorted dataset	7
Figure 4. PCA analysis of dataset	8
Figure 5. Normalisation and Standardisation of dataset	9
Figure 6. Visualisation of South German dataset	10
Figure 7. Random Forest training set	11
Figure 8. Random Forest normalised training set	12
Figure 9. Random Trees training set	13
Figure 10. Random Trees normalised training set	14
Figure 11. Naive Bayes training set	15
Figure 12. Naive Bayes normalised training set	16
Figure 13. Logistic training set	17
Figure 14. Logistic normalised training set	18
Figure 15. Bayes Net training set	19
Figure 16. Bayes Net normalised training set	20
Figure 17. KNN (IBK) training set	21
Figure 18. KNN (IBK) normalised training set	22
Figure 19. Random Forest 10-fold cross validation	23
Figure 20. Random Forest 80/20% percentage split	24
Figure 21. Random Trees 10-fold cross validation	25
Figure 22. Random Tree 80/20% percentage split	26
Figure 23. Naive Bayes 10-fold cross validation	27
Figure 24. Naive Bayes 80/20% percentage split	28
Figure 25. Logistic 10-fold cross validation	29
Figure 26. Logistic 80/20% percentage split	30
Figure 27. Bayes Net 10-fold cross validation	31
Figure 28. Bayes Net 80/20% percentage split	32
Figure 29. KNN (IBk) 10-fold cross validation	33
Figure 30. KNN (IBk) 80/20% percentage split	34
Figure 31. Classification results	34
Figure 32. Full list of Attributes	40

1. Introduction

This report will analyse the UCI Machine Learning Repository's South German Credit Data Set to identify loan repayment predictors. The analysis includes data pre-processing, analysis, and visualisation, as well as German-to-English column heading translation. Data mining's purpose, dataset characteristics, pre- and post-processing steps, analysis methods, and business recommendations are covered in the report. Analysis accuracy, efficiency, and improvements are also discussed.

1.2 Data Mining Purpose and Practices

Data mining uses statistical and computational methods to find patterns, correlations, and insights in large datasets, [1]. Data mining primarily extracts useful information from raw data and organises it for use. It uses statistics, machine learning, database management, and AI. Learning practical skills to improve decision-making, operations, and organisational effectiveness is the main goal. This data can help companies make smart decisions, predict trends, and gain an edge, [2], [3].

Data mining has many steps. Data is collected from CRM, ERP, social media, and public databases. Data is carefully collected to ensure accuracy and relevance at this stage. To ensure dataset accuracy and consistency, data cleaning includes error correction, duplicate removal, and missing value resolution using imputation, normalisation, standardisation, and outlier detection, [4]. Data integration uses Extract, Transform, and Load (ETL) to combine data from different departments and formats into a single dataset, [5]. The data selection phase selects only project-relevant data for analysis. Data transformation uses aggregation, discretization, and feature selection or extraction to format or structure data. To find patterns in prepared data, data mining uses advanced algorithms like classification, clustering, regression, association rule learning, and anomaly detection, [6]. Pattern evaluation uses precision, recall, and F-score to assess pattern importance and usefulness. Knowledge representation uses dashboards, graphs, and reports to present extracted insights clearly and practically. Stakeholders can easily interpret and use data insights. [7].

1.3 How Data Mining can be used in Business

Data mining is essential in many industries. CRM analyses customer interaction data to identify the most profitable customers, forecast their needs, customise services, and improve retention rates by clustering for targeted marketing strategies, [8], [9].

Marketing data mining improves efficiency by identifying effective channels and strategies, predicting customer responses to campaigns, and maximising ROI. Understanding trends and patterns helps predict demand, manage inventory, and set prices to maximise revenue, [10].

Data mining is used in healthcare to predict disease outbreaks, customise treatments, and improve patient care through early detection and resource allocation, [12].

2. South German Credit Dataset

[illegible]

Figure 1. South German Credit dataset

By downloading the South German Credit data set and exporting it to an excel file, the complete data can be viewed, this is seen in Figure 1 above. The data set includes parameters such as age, job, and credit history, as well as additional attributes including an individual's living condition, current credits and more.

The dataset has 1000 data points. This sample facilitates the creation of strong prediction models that can differentiate between borrowers who are likely to repay their debts and those who have a higher probability of defaulting. Additionally, it provides a substantial amount of data for study.

Upon examining the data set, it is evident that there are missing values. The missing values decreases the quality of the dataset, guaranteeing an undependable basis for constructing correct models.

The complete list of attributes is available in the Appendix of this report.

3. Processing

This section includes the primary, secondary, and post-processing of the South German dataset, with a description of each process.

3.1 Primary Pre-processing

Primary pre-processing in Data Mining encompasses the early steps of preparing raw data for machine learning tasks. It primarily covers operations such as cleaning, basic transformation and data loading, [14].

The first initial pre-processing of the dataset is to the dataset download the English descriptions, which was provided, as the current version is in German. This means that a thorough examination can be conducted on the characteristics that are included.

Once this has been done, data loading starts and is accomplished by utilising different file formats such as CSV, or ARFF, enabling users to import data from a wide range of sources.

[illegible]

Figure 2. Attributes named in dataset

There were two methods that are employed to perform initial preprocessing on the South German dataset. One approach involved opening the file in Excel. After opening the file, the procedure was to manually edit and add attribute names. Figure 2 above provides a visual representation of this.

[illegible]

Figure 3. Filtered and sorted dataset

In addition, utilisation of the filter tool in Excel to arrange each attribute and identify any erroneous or absent values. Manually processing the data was done to ensure its suitability for use in Weka. After completing this task, successful elimination of any inaccurate information was removed from the dataset.

Weka data pre-processing was the second method. In Weka, primary pre-processing involves data cleaning tasks like resolving missing values, eliminating duplicates, and addressing outliers that may have been missed in Excel. Weka supports mean, median, and mode imputation for missing values. It supports advanced algorithms like k-nearest neighbours imputation.

The outlier identification and removal filters in Weka also improve data quality by detecting and fixing anomalies that could affect model training. Additionally, users can select attributes, discretize data, and normalise data to prepare it for machine learning tasks, [15].

3.2 Secondary Pre-processing

Secondary pre-processing in Weka encompasses sophisticated procedures for preparing data, which are carried out after the initial steps of data cleaning and basic processing. These strategies encompass feature selection, which involves identifying and preserving the most pertinent qualities while eliminating redundant or irrelevant ones. This process reduces the complexity of the data and enhances the performance of the model, [15].



Figure 4. PCA analysis of dataset

Feature transformation, such as Principal Component Analysis (PCA), converts data into more appropriate formats or representations, improving interpretability and efficacy. This can be seen in Figure 4 above, [16].

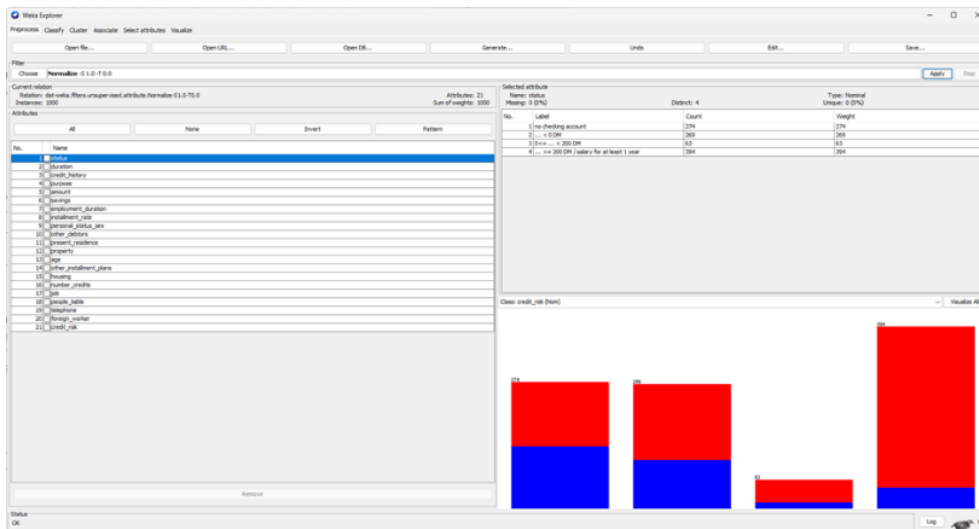


Figure 5. Normalisation and Standardisation of dataset

Figure 5 shows normalisation or standardisation, which is a process that adjusts the scales or distributions of characteristics to make them more similar. This helps prevent features with larger scales from overpowering those with lower scales, ultimately enhancing the efficiency of specific algorithms.

Resampling ensures a more equitable class distribution, improving prediction accuracy for underrepresented classes.

These steps optimise the dataset, improve model efficiency and accuracy, lower computational complexity, improve interpretability, and ensure robustness and generalisation to unfamiliar data, [17].

3.3 Post-processing

Weka has post-processing techniques for machine learning model output after training and evaluation. The goal is to ensure reliable, interpretable, and effective findings. Model evaluation uses accuracy and ROC curves to assess performance. Interpreting results requires understanding model predictions and decision-making. Model comparison evaluates multiple configurations to find the best model, while threshold adjustment optimises classification thresholds for metrics, [18].

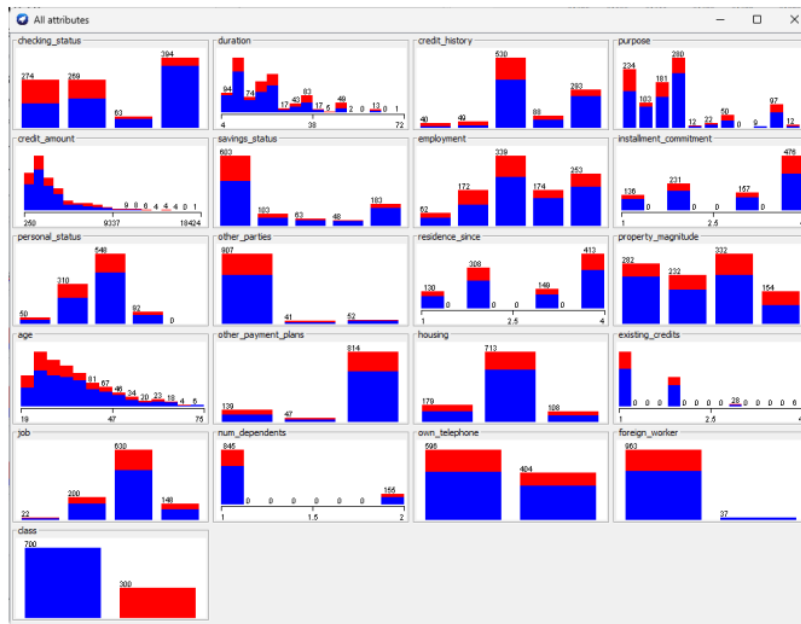


Figure 6. Visualisation of South German dataset

Visualisation facilitates the communication of results by means of charts and graphs, while deployment preparation guarantees that the model is fully prepared for real-world use. Above shows the visualisation of the South German dataset.

The post-processing processes are essential for evaluating the reliability of the model, extracting insights, improving performance, and effectively conveying findings to stakeholders. Ultimately, these procedures enhance transparency and trust in the model. [7].

4. Sample Runs

In order to conduct sample runs, navigate to the "classify" tab in Weka and proceed to test the "credit_risk" dataset. There will be multiple algorithms employed to assess the precision of the dataset.

4.1 Random Forest

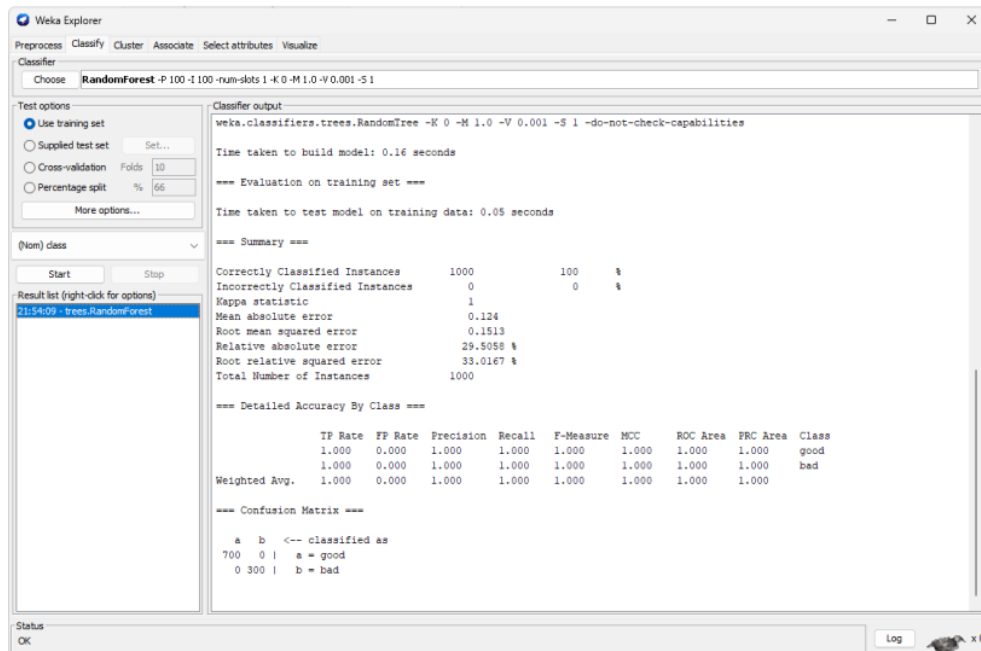


Figure 7. Random Forest training set

In Figure 7, the Random Forest technique was used as the classifier model without any normalisation, with 100 iterations. The training set model construction duration of 0.16 seconds, along with the 100% accuracy in classifying instances, suggests that the dataset is probably not large, and the approach is efficient for the size of the data.

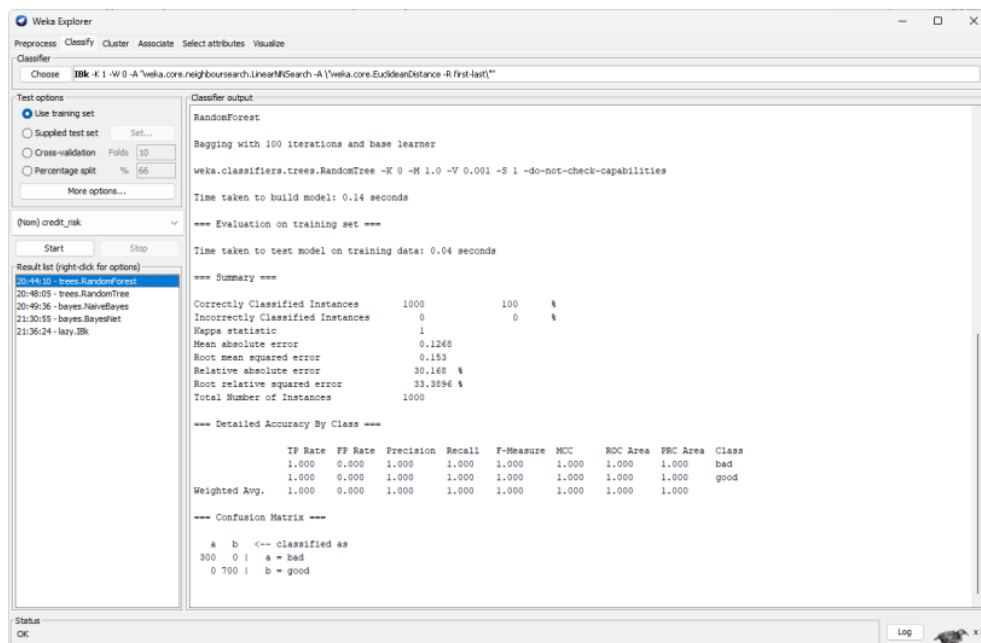


Figure 8. Random Forest normalised training set

In Figure 8, the Random Forest technique was used as the classifier model with normalisation, with 100 iterations. The training set model construction duration of 0.14 seconds, along with the 100% accuracy in classifying instances.

4.2 Random Trees

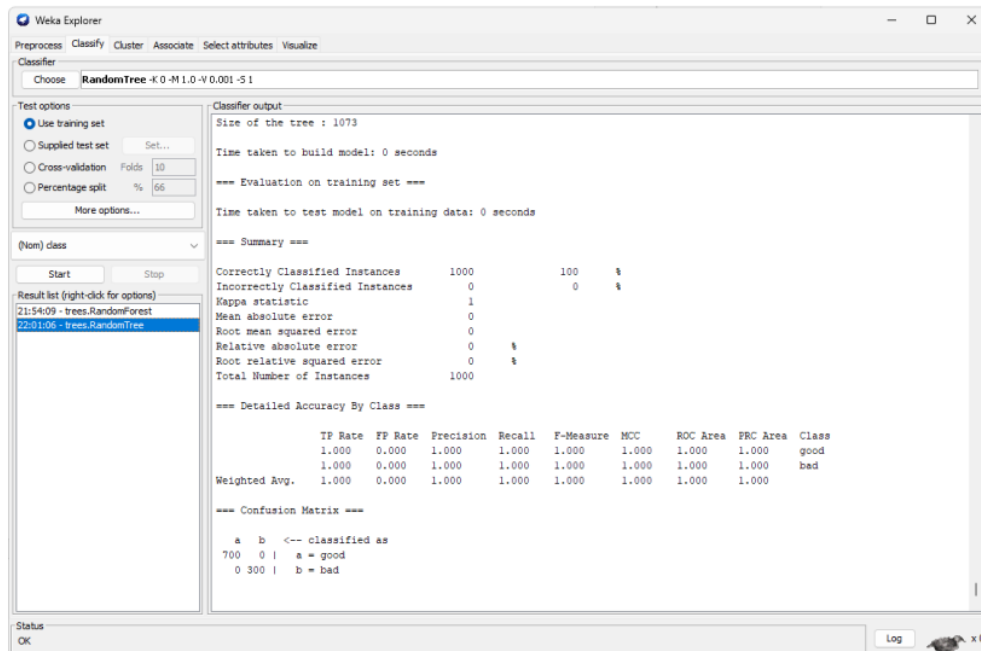


Figure 9. Random Trees training set

In Figure 9, the Random Trees technique was used as the classifier model, with a 1073 tree size. The training set model construction duration of 0 seconds, along with the 100% accuracy in classifying instances, suggests that the dataset is probably not large, and the approach is efficient for the size of the data.

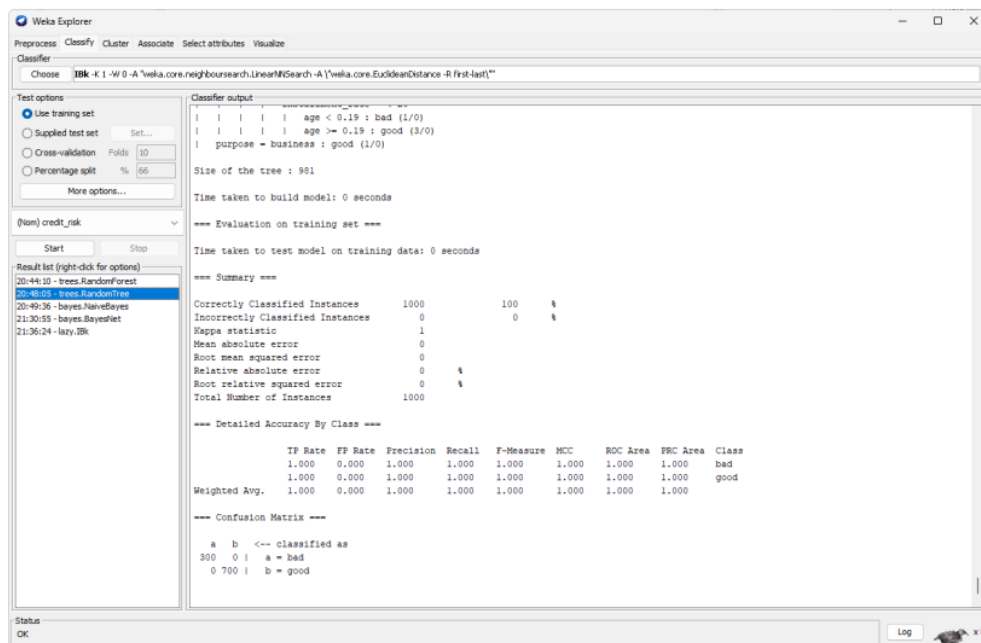


Figure 10. Random Trees normalised training set

In Figure 10, the Random Trees technique was used as the classifier model with normalisation, with 981 tree size. The training set model construction duration of 0 seconds, along with the 100% accuracy in classifying instances.

4.3 Naïve Bayes

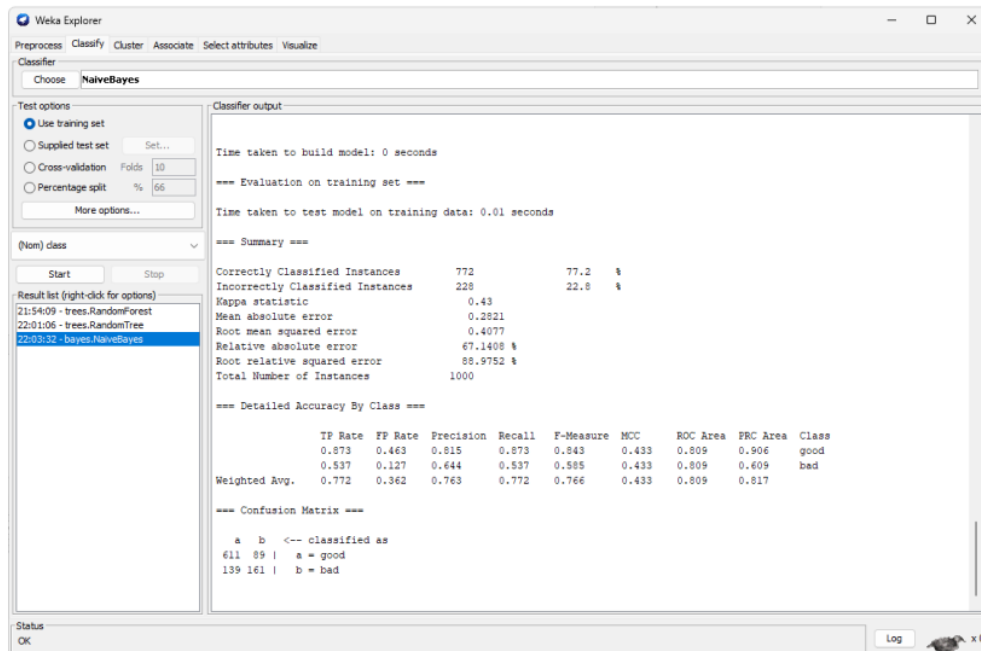


Figure 11. Naive Bayes training set

In Figure 11, the Naïve Bayes technique was used as the classifier model. The training set model construction duration of 0 seconds, along with the 77.2% accuracy in classifying instances and 22.8% incorrectly classified.

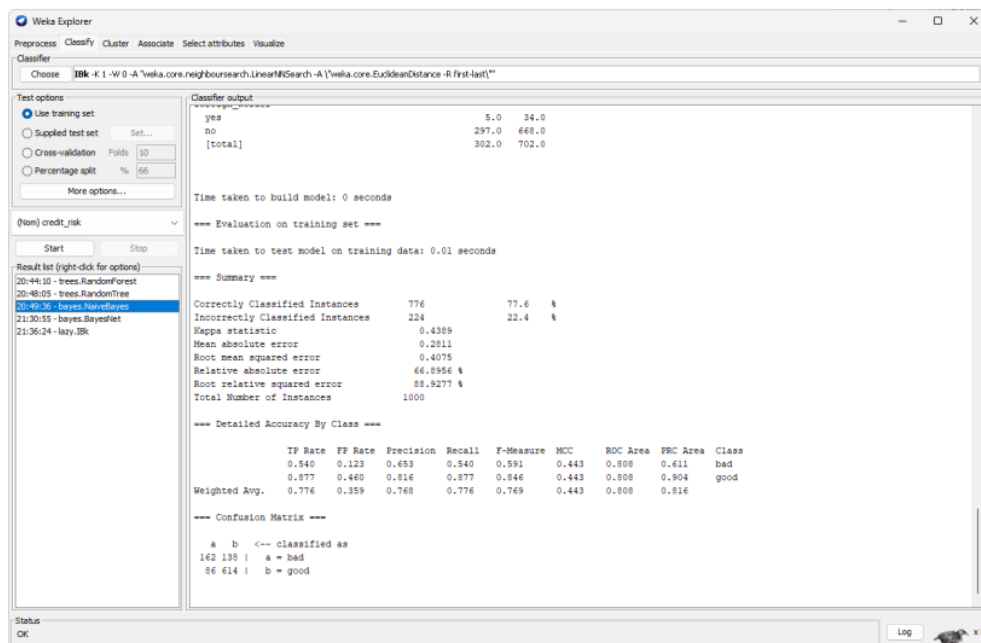


Figure 12. Naive Bayes normalised training set

In Figure 12, the Naïve Bayes technique was used as the classifier model, with normalisation. The training set model construction duration of 0 seconds, along with the 77.6% accuracy in classifying instances and 22.4% incorrectly classified.

4.4 Logistic

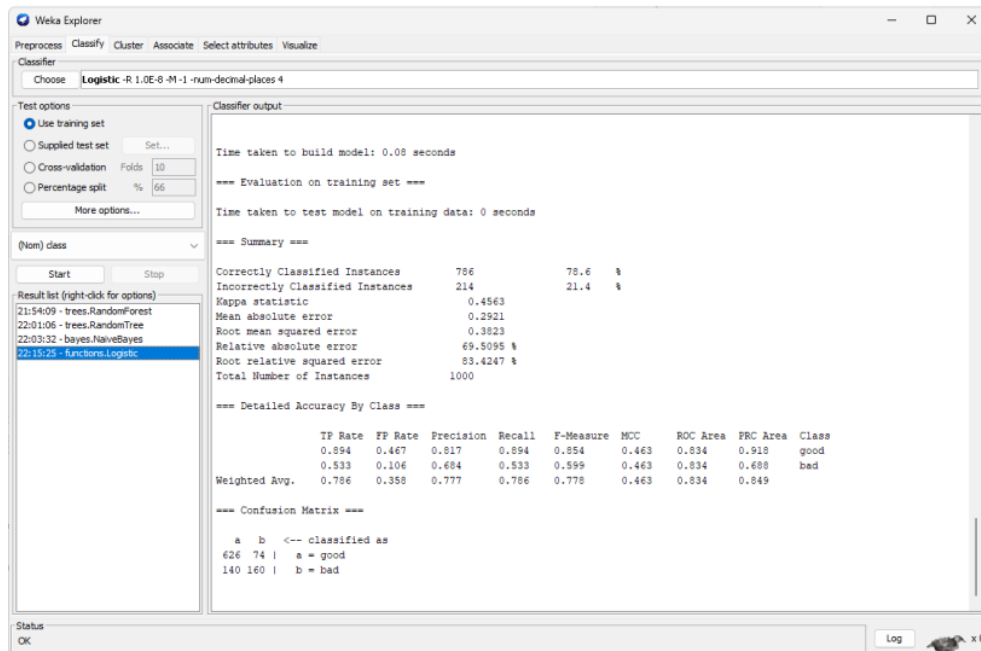


Figure 13. Logistic training set

In Figure 133, the Logistic technique was used as the classifier model. The training set model construction duration of 0.8 seconds, along with the 78.6% accuracy in classifying instances and 21.4% incorrectly classified.

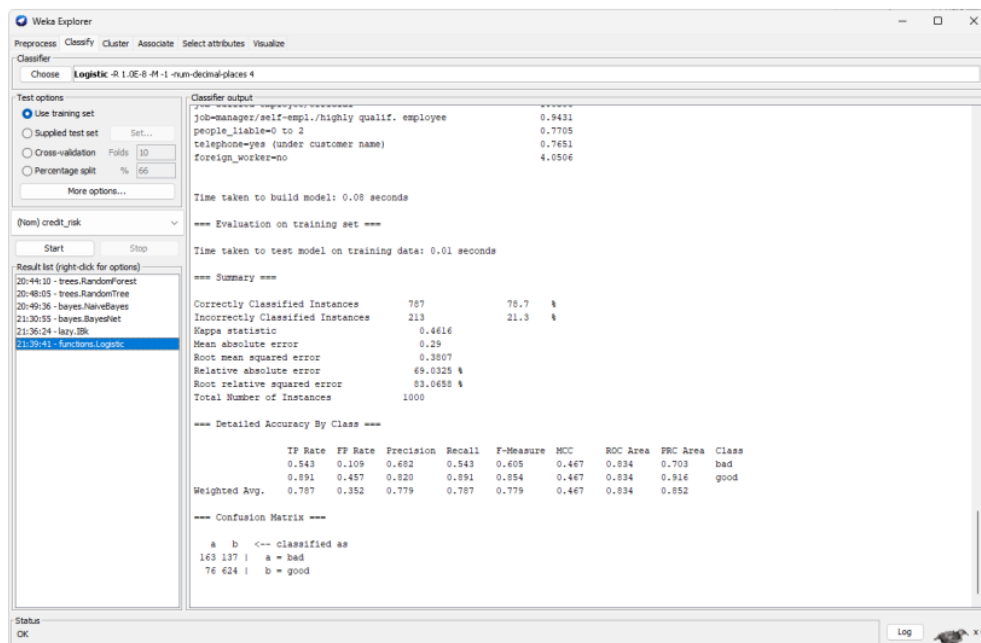


Figure 14. Logistic normalised training set

In Figure 14, the Logistic technique was used as the classifier model, with normalisation. The training set model construction duration of 0.8 seconds, along with the 78.7% accuracy in classifying instances and 21.3% incorrectly classified.

4.5 Bayes Net

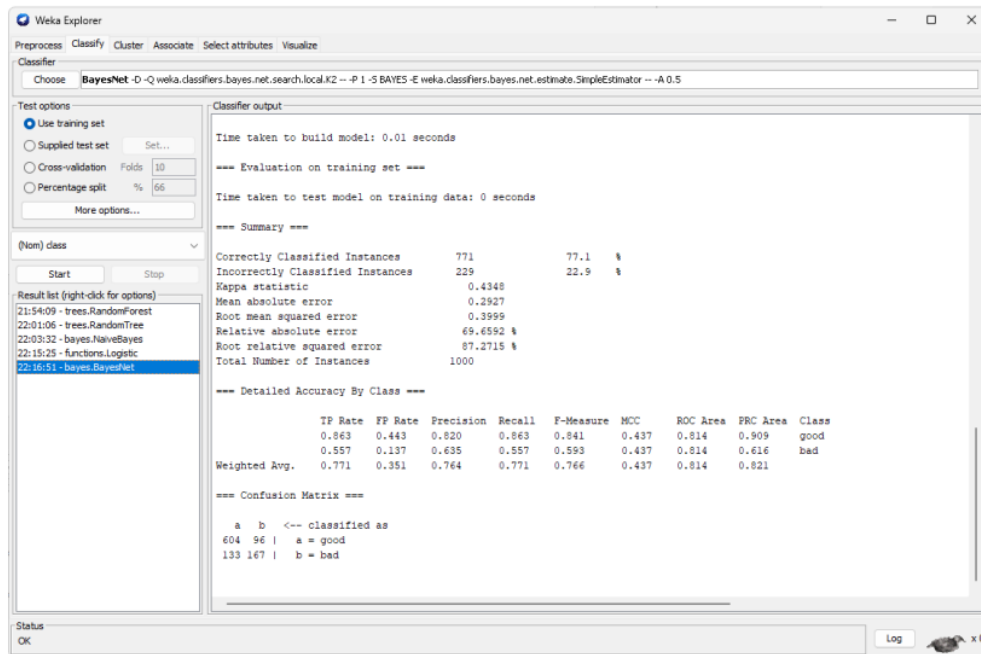


Figure 15. Bayes Net training set

In Figure 15, the Bayes Net was used as the classifier model. The training set model construction duration of 0.1 seconds, along with the 77.1% accuracy in classifying instances and 22.9% incorrectly classified.

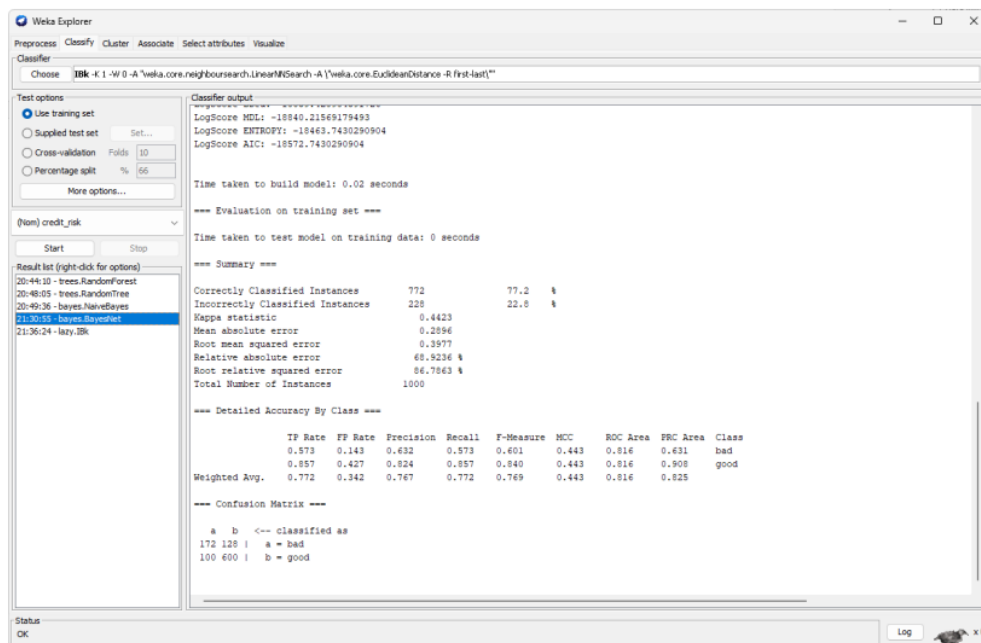


Figure 16. Bayes Net normalised training set

In Figure 16, the Bayes Net technique was used as the classifier model, with normalisation. The training set model construction duration of 0.2 seconds, along with the 77.2% accuracy in classifying instances and 22.8% incorrectly classified.

4.6 KNN

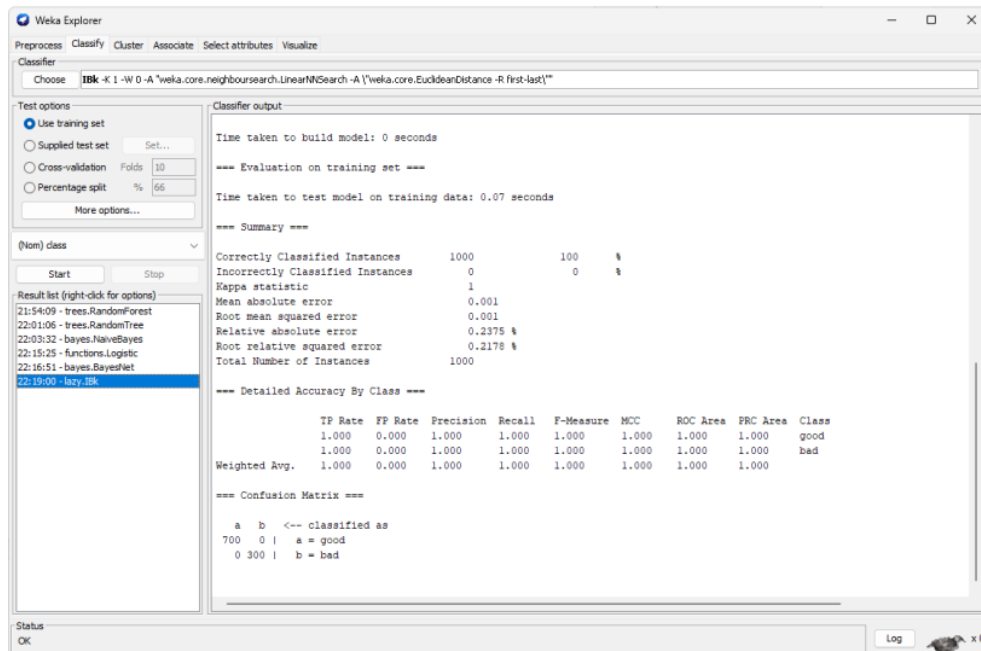


Figure 17. KNN (IBK) training set

In Figure 17, the KNN technique was used as the classifier model. The training set model construction duration of 0 seconds, along with the 100% accuracy in classifying instances

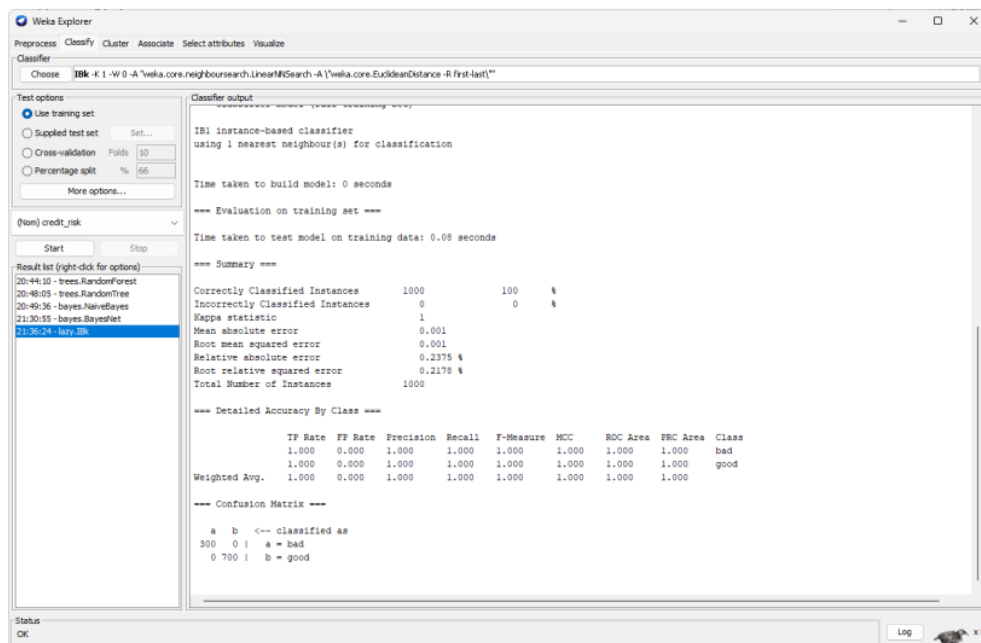


Figure 18. KNN (IBK) normalised training set

In Figure 18, the KNN technique was used as the classifier model, with normalisation. The training set model construction duration of 0 seconds, along with the 100% accuracy in classifying instances.

5. Results

Following multiple iterations of the sample runs, utilising the normalised data would yield the most accurate and conclusive results for the final analysis. These are shown below:

5.1 Random Forest

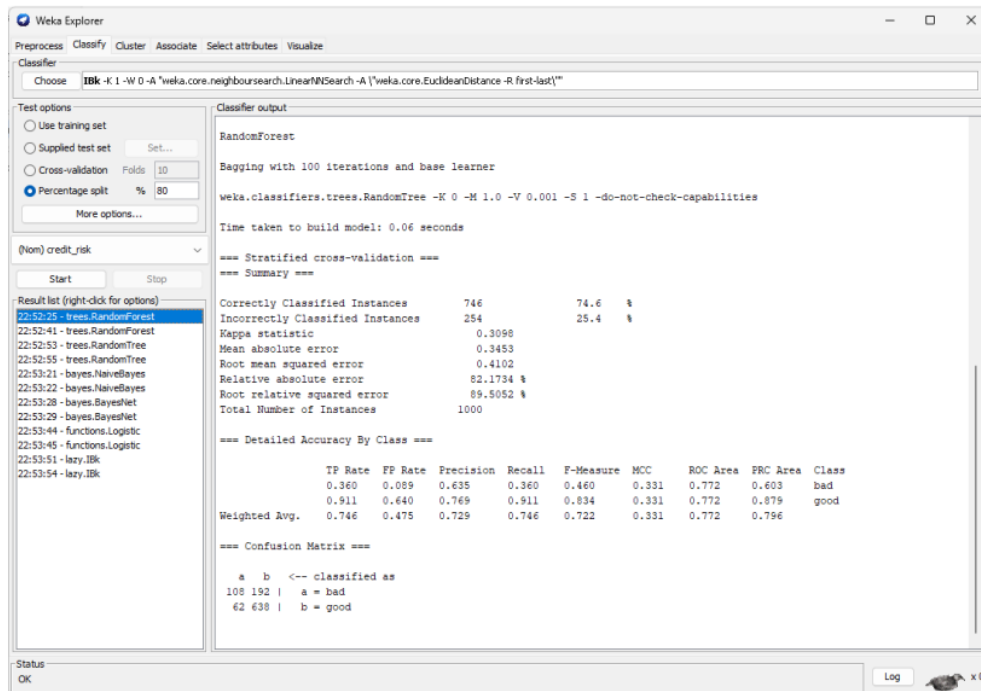


Figure 19. Random Forest 10-fold cross validation

Selection was revealed by evaluation on the Random Forest 10-fold validation, with 1000 instances (74.6%) correctly classified.

The rather low value of 0.3098 obtained from the Kappa statistic, which measures inter-rater agreement, raises the possibility of model fit to the data deficiencies.

Precision, recall, F-measure, and MCC were among the metrics for both "good" and "bad" classes found by detailed accuracy by class analysis.

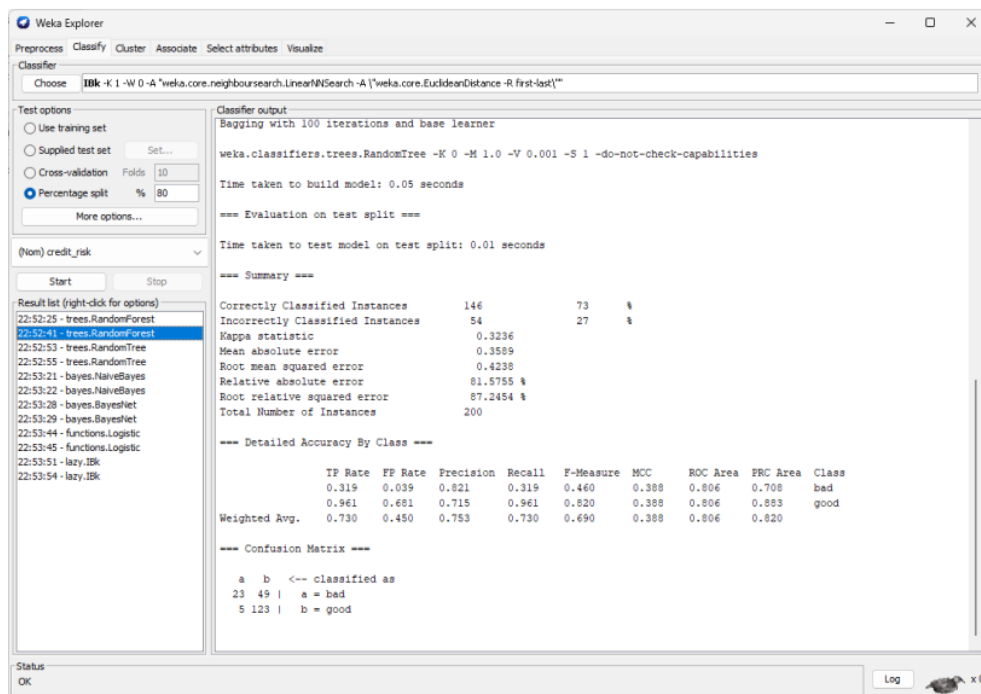


Figure 20. Random Forest 80/20% percentage split

A high performance is shown by the test split evaluation, which shows 73% of cases (200) correctly classified and a Kappa statistic of 0.3236, which by some standards indicates fair agreement.

More investigation explores the specific accuracy by class, emphasising for every class the MCC, F-measure, precision, and recall.

Analysing the confusion matrix clarifies the inclinations of the model. While misclassifying a noteworthy 27% of instances, it correctly identifies 23/5 instances of the "bad" class and 49/123 instances of the "good" class as true positives.

5.2 Random Trees

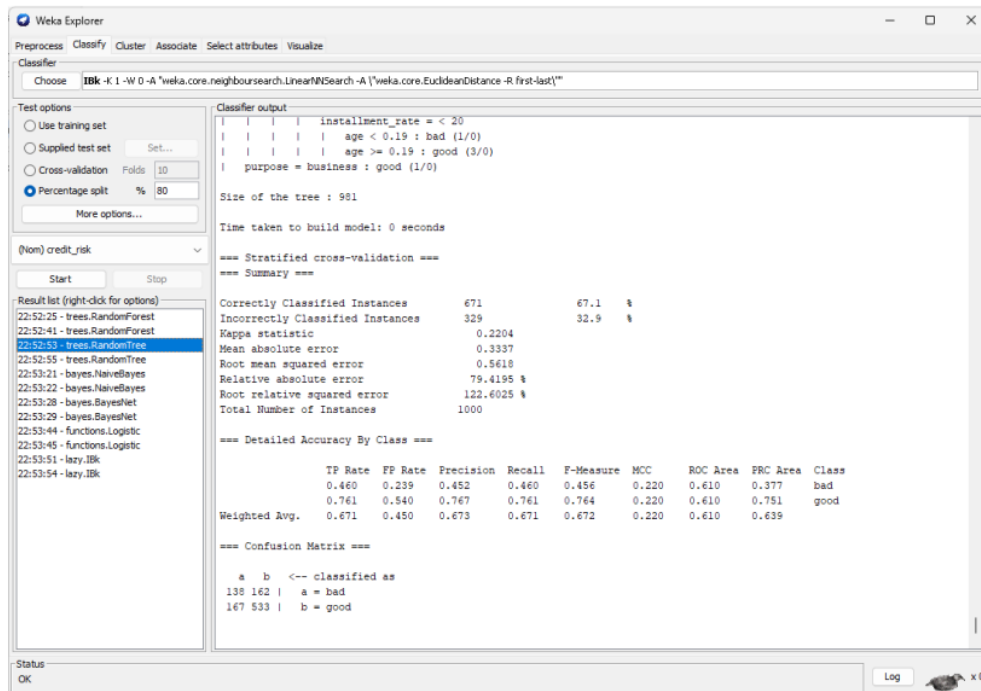


Figure 21. Random Trees 10-fold cross validation

Evaluation on the Random Trees 10-fold validation revealed selection; 1000 cases (67.1%) were correctly classified.

The possibility of model fit to the data shortcomings is decreased by the rather low value of 0.2204 from the Kappa statistic, which gauges inter-rater agreement.

Detailed accuracy by class analysis revealed metrics for both "good" and "bad" classes including precision, recall, F-measure, and MCC.

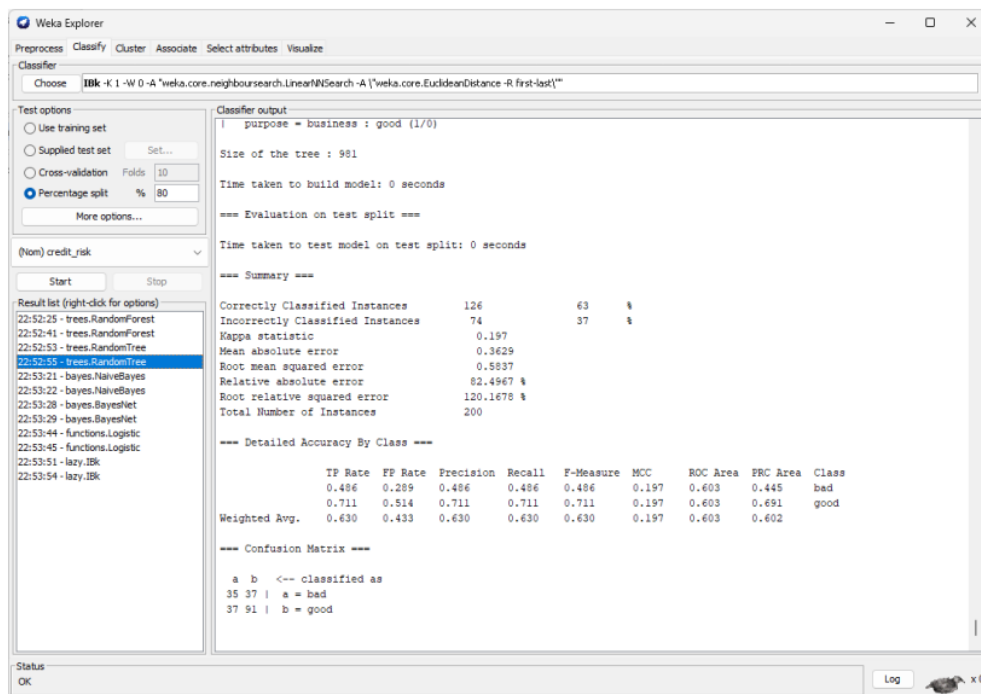


Figure 22. Random Tree 80/20% percentage split

With 63% of cases (200) correctly classified and a Kappa statistic of 0.197, which by some criteria indicates poor agreement, the test split evaluation performs moderately.

Further study looks into the particular accuracy by class, highlighting the MCC, F-measure, precision, and recall for each class.

Clarifying the model's inclinations is the analysis of the confusion matrix. Although it misclassifies a moderate 37% of cases, it correctly classifies 35/37 cases of the "bad" class and 37/91 cases of the "good" class as real positives.

5.3 Naïve Bayes

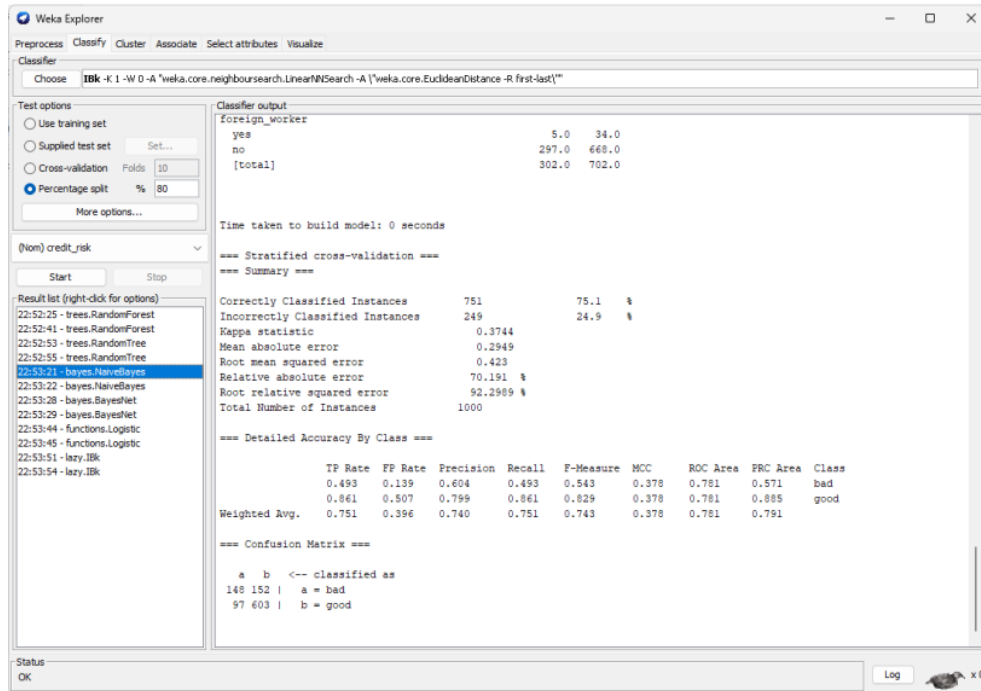


Figure 23. Naive Bayes 10-fold cross validation

Tenfold validation of Naïve Bayes revealed selection; 75.1 percent of 1000 cases were correctly classified.

The rather low-mid value of 0.3744 obtained from the Kappa statistic, which measures inter-rater agreement, raises the likelihood of model fit to the data deficiencies.

Precision, recall, F-measure, and MCC are among the metrics for both "good" and "bad" classes that were revealed by detailed accuracy by class analysis.

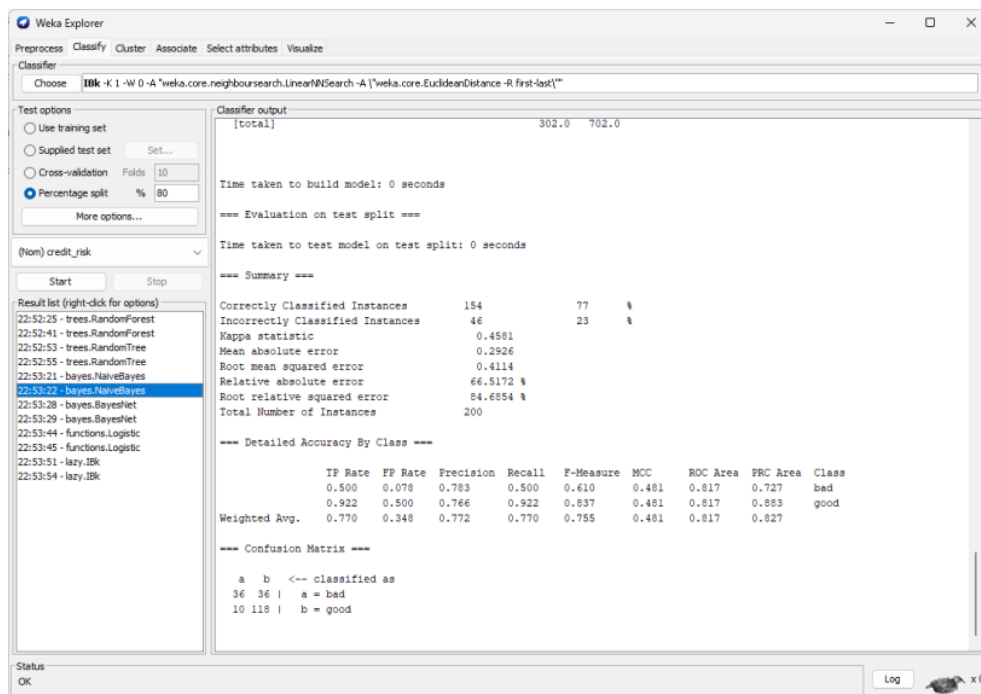


Figure 24. Naïve Bayes 80/20% percentage split

The test split evaluation does rather well with 77% of cases (200) correctly classified and a Kappa statistic of 0.4581, which by some criteria indicates good agreement.

Further investigation examines the specific accuracy by class, emphasising the MCC, F-measure, precision, and recall for every class.

Analysing the confusion matrix helps to clarify the inclinations of the paradigm. It correctly classifies 36/10 cases of the "bad" class and 10/118 cases of the "good" class as actual positives, albeit misclassifying a small 23% of cases.

5.4 Logistic

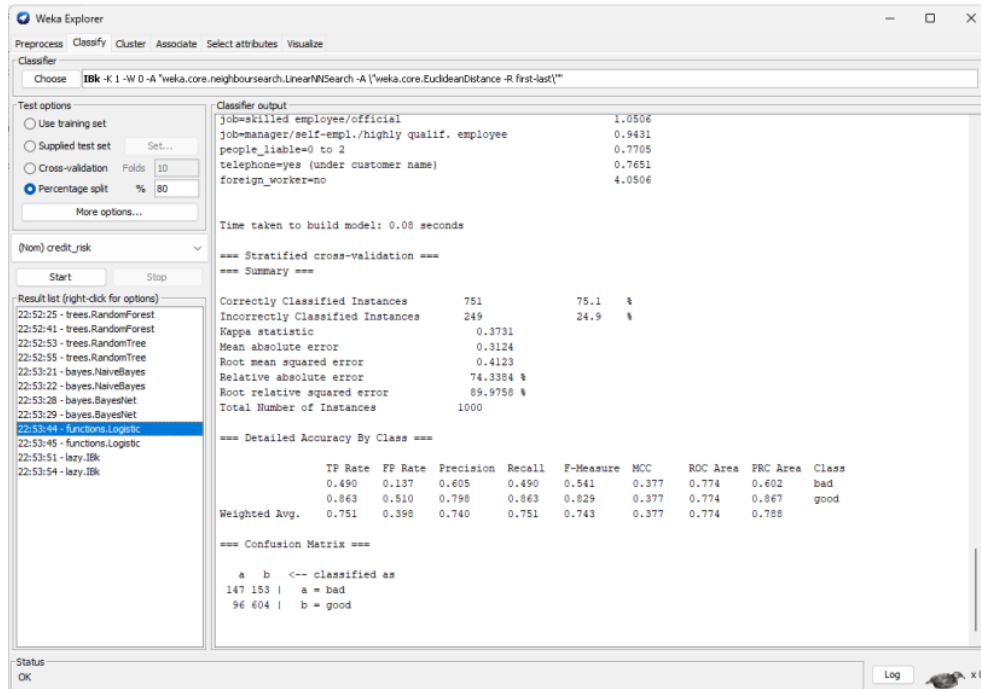


Figure 25. Logistic 10-fold cross validation

A tenfold validation of Logistic revealed selection; 75.1 percent of 1000 cases were correctly classified.

The Kappa statistic's (inter-rater agreement) rather low-mid value of 0.3731 increases the possibility that the model could fit the data shortcomings.

Detailed accuracy by class analysis revealed metrics for both "good" and "bad" classes including precision, recall, F-measure, and MCC.

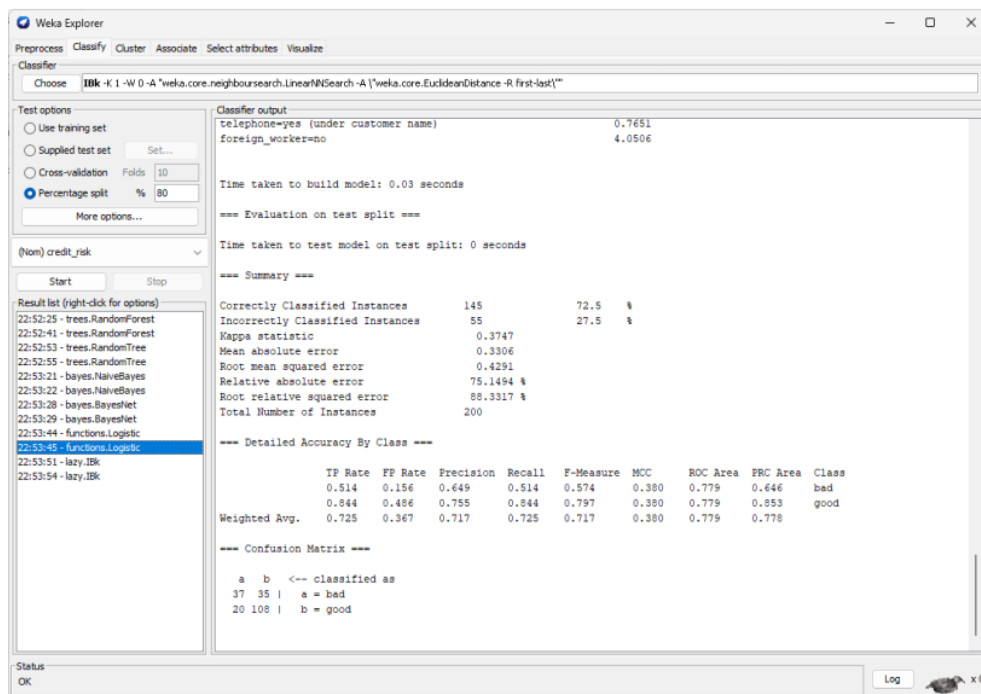


Figure 26. Logistic 80/20% percentage split

With 72.5% of cases (200) correctly classified and a Kappa statistic of 0.3747, which by some standards indicates moderate agreement, the test split evaluation does quite well.

The inclinations of the paradigm become clearer when the confusion matrix is analysed. Though misclassifying just 27.5% of cases, it correctly classifies 37/20 cases of the "bad" class and 35/108 cases of the "good" class as actual positives.

5.5 Bayes Net

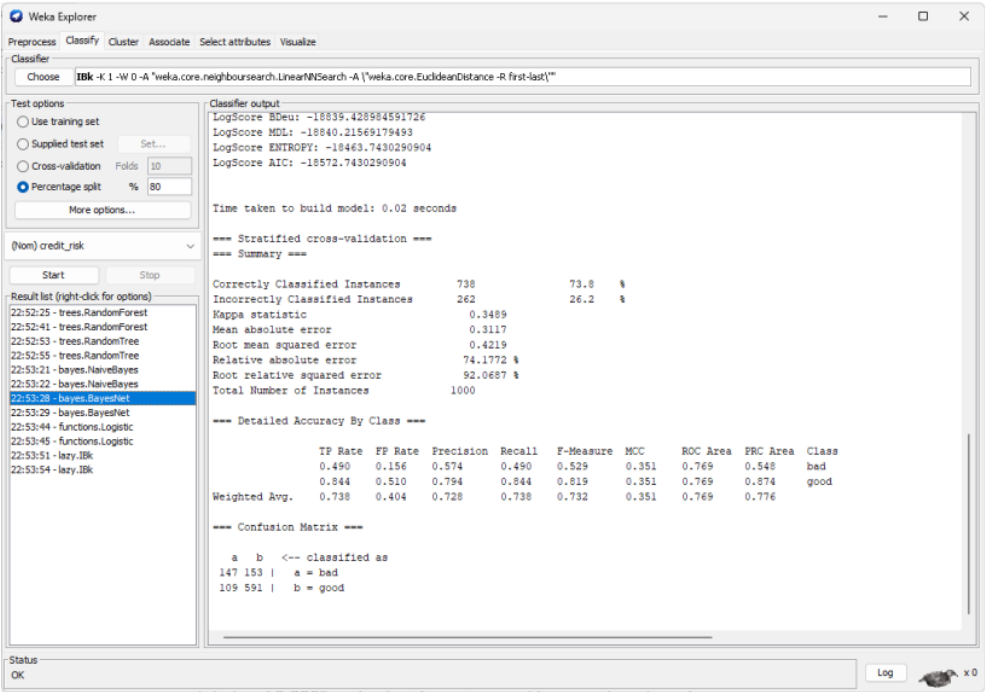


Figure 27. Bayes Net 10-fold cross validation

Selection was shown by tenfold validation of Bayes Net; 73.8 percent of 1000 cases were correctly classified.

The rather low-mid value of the Kappa statistic (inter-rater agreement) of 0.3489 raises the likelihood that the model could account for the limitations of the data.

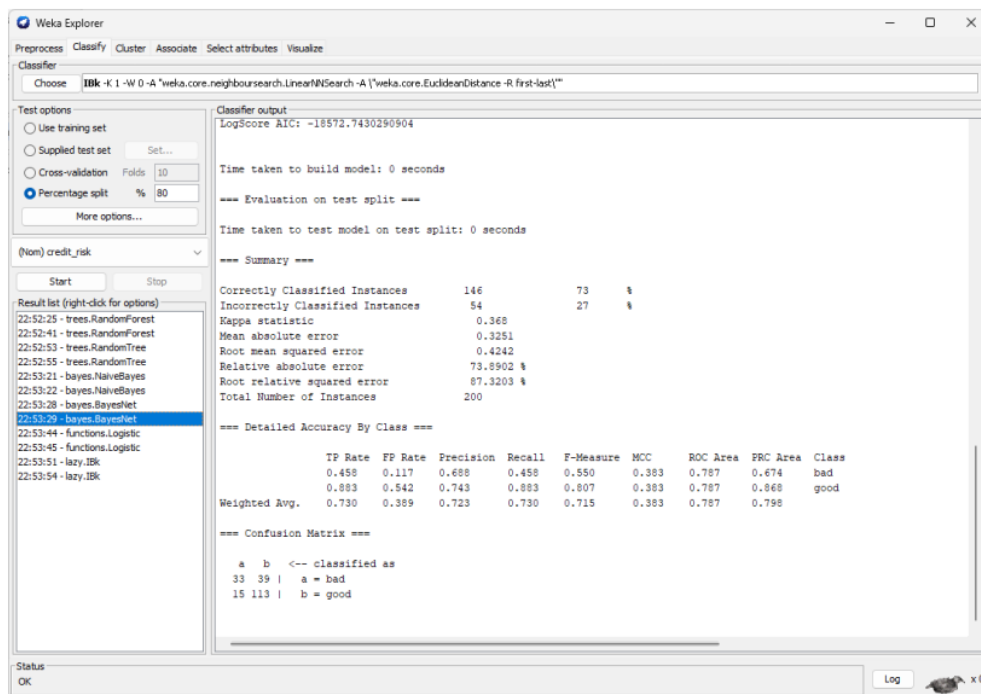


Figure 28. Bayes Net 80/20% percentage split

The test split evaluation performs rather well with 73% of cases (200) correctly classified and a Kappa statistic of 0.368, which by some criteria indicates moderate agreement.

Analysis of the confusion matrix makes the paradigm's inclinations more apparent. Although it misclassifies just 27% of cases, it correctly classifies 33/15 cases of the "bad" class and 39/113 cases of the "good" class as real positives.

5.6 KNN

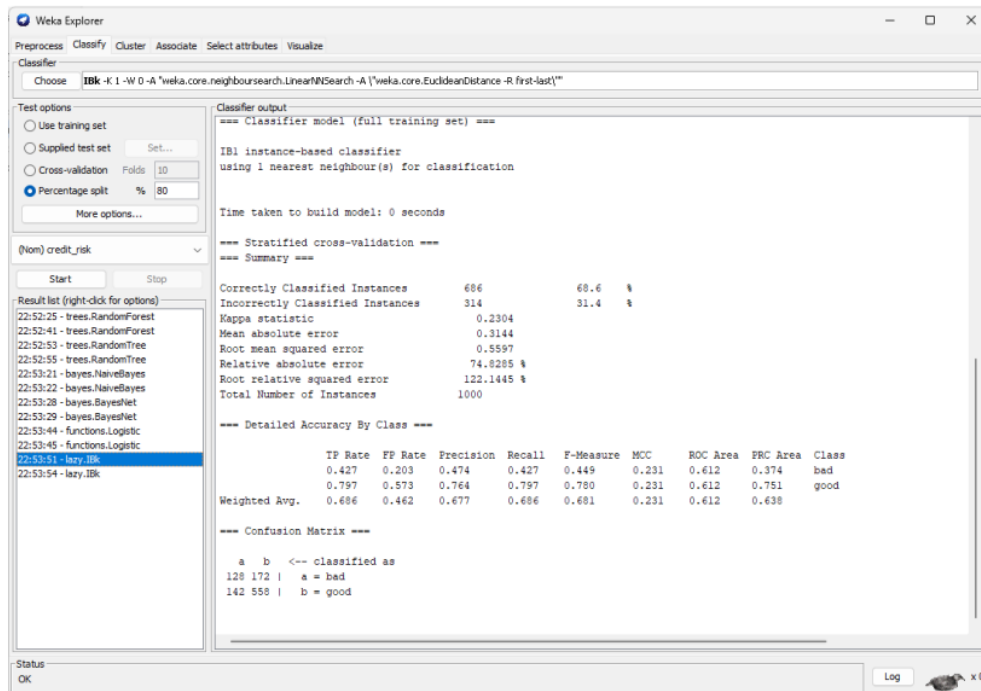


Figure 29. KNN (IBk) 10-fold cross validation

A tenfold validation of Bayes Net demonstrated selection; 68.6% of 1000 cases were correctly classified.

The probability that the model could explain the constraints of the data is increased by the rather low value of the Kappa statistic (inter-rater agreement) of 0.2304.

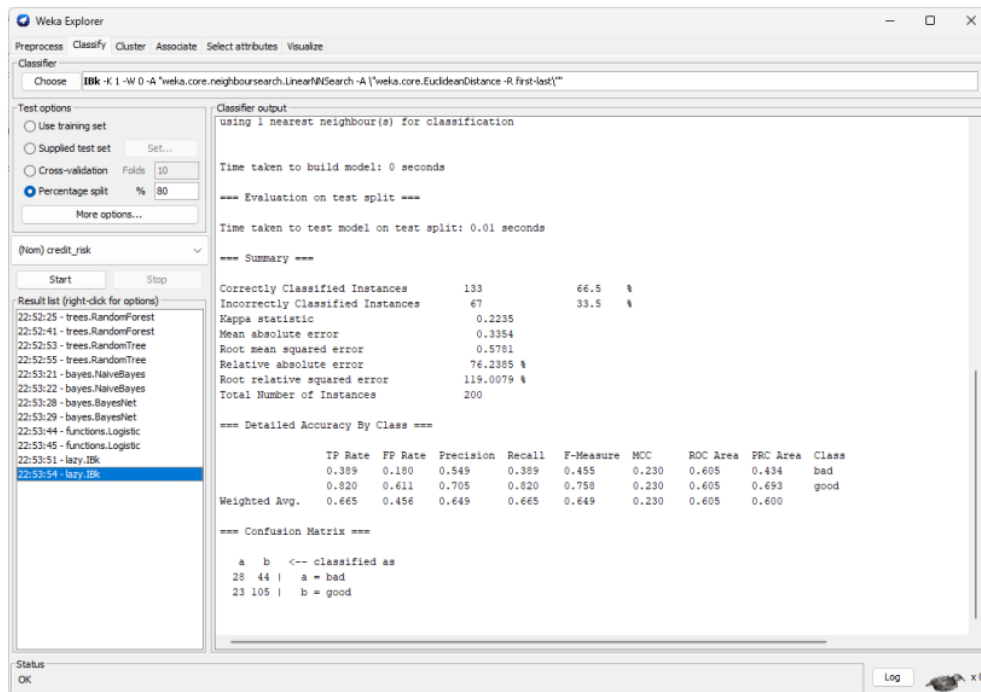


Figure 30. KNN (IBk) 80/20% percentage split

At 66.5% of cases (200) correctly classified and a Kappa statistic of 0.2235, which by some standards indicates moderate agreement, the test split evaluation performs averagely.

The inclinations of the paradigm are made clearer by examination of the confusion matrix. It correctly classifies 28/23 cases of the "bad" class and 44/105 cases of the "good" class as actual positives, albeit misclassifying mid-high 33.5% of cases.

5.7 Classification Results

Classification	Training set (normalised)	10-fold validation	Percentage split
Random Forest	100%	74.6%	73%
Random Trees	100%	67.1%	63%
Naïve Bayes	77.6%	75.1%	77%
Logistic	78.7%	75.1%	72.5%
Bayes Net	77.2%	73.8%	73%
KNN(IBk)	100%	68.6%	66.5%

Figure 31. Classification results

Here are the findings from the analysis conducted on the South German Credit Dataset. The recommended classification method is Naïve Bayes, although Random Forest, Logistic, and Bayes Net are also viable options to consider, [19].

The rationale behind these algorithms is based on research, which indicates that they are extensively employed in classification-oriented data analysis. Since the data set is for classification, this justifies that the most suitable algorithms to use would be those based on these classification, [20], [21], [22], [23].

6. Business Recommendations

After analysing the South German Credit Dataset classifier results, several business recommendations can be made. Random Forest, Random Trees, Naive Bayes, Bayes Net, Logistic, and KNN (IBk) k-nearest neighbours have an accuracy of 70% and are better at identifying "good" credit risks.

Improve agreement metrics, especially Kappa, on the test set for Random Forest. This is crucial when normalisation methods are not used. This suggests overfitting or tailoring the model to the training data. Additionally, there is a noticeable bias towards the majority class "good", which requires strategies to improve the model's ability to categorise minority class "bad" cases.

To improve the model's ability to accurately categorise credit risks as "bad," the dataset must be expanded by adding data points for "bad" credit instances and integrating more relevant features through feature engineering.

When calculating the costs of misclassifying "good" and "bad" classes, other factors must be considered. Understanding these costs helps prioritise model improvements based on business case. Additionally, combining a Decision Tree with Random Forest or KNN can help explain forecasts if understanding them is crucial.

Supplemental verifications for high-risk scenarios, credit insurance, and flexible credit limits may reduce false positives. Using predictive analytics to customise loan offers and create loyalty programmes for good credit customers improves customer management. Retraining the model with new data and implementing a performance monitoring system are necessary to adapt to changing client behaviours and economic conditions.

Ensuring credit decision criteria are open and actively reducing model biases will help comply with regulatory and ethical standards. These comprehensive strategies may improve credit risk assessment, reduce default risks, and boost customer satisfaction and loyalty.

7. Reflection

7.1 Does it work well?

Credit risk assessment models on the South German Credit Dataset perform well, with accuracy rates around 75%. Random Forest, Random Tress, Naïve Bayes, Logistic Bayes Net, and KNN(IBk) are accurate and precise, with a focus on "good" credit risks. The recall and precision metrics for "good" credit classifications show that the models are effective at identifying clients who are likely to repay their debts. However, identifying "high-risk" credit profiles is less accurate, suggesting improvement. The models are good, but they could reduce credit risk more.

7.2 Advantages and Disadvantages

These models accurately predict good credit, which boosts loan approval confidence. The uniformity of each classifiers performance shows the methodology's strength and reliability, ensuring reliable results. Our consistent service performance allows us to tailor credit offerings and implement client loyalty programmes using predictive analytics.

However, significant drawbacks must be considered. The high rate of misclassification for "bad" credit risks poses a business risk and may increase defaults. Advanced models like Random Forest require constant optimisation and retraining with new data to work. The models may also have inherent biases that, if not monitored and changed, could unfairly affect certain client groups.

7.3 Different Approaches (if any)

These issues can be addressed using various methods. Adding "adverse" credit situations or artificial data can help balance the dataset. Feature engineering can also improve model correctness by adding relevant characteristics. Ensemble methods that combine model predictions should improve accuracy and resilience, especially for "poor" credit. Advanced algorithms like gradient boosting machines (GBM) and deep learning can improve performance. Combining statistical methods with machine learning models may improve prediction accuracy by combining their strengths.

Regular audits to identify and address biases can ensure the model is fair and ethical. When dealing with imbalanced datasets, F1-score, Matthews correlation coefficient (MCC), and precision-recall curves can be used to evaluate model performance in addition to accuracy.

7.4 Conclusion

To conclude, the South German Credit Dataset models perform well, especially in identifying "good" credit risks. There is room for improving the classification of "bad" credit risks to reduce defaults. Credit risk evaluation must be improved by improving

data accuracy, investigating complex algorithms, and monitoring biases. These methods can help create more robust models that improve accuracy, fairness, and regulatory compliance. This will improve business decisions and customer satisfaction

8. Bibliography

- [1] K. P. Soman, S. Diwakar and V. Ajay, DATA MINING: THEORY AND PRACTICE, PHI Learning Pvt. Ltd, 2006.
- [2] S. Sumathi and S. N. Sivanandam, "Data Warehousing and Data Mining," in *Introduction to Data Mining and its Applications*, Springer Link, 2006, pp. 1-20.
- [3] B. Mento and B. Rapple, Data Mining and Data Warehousing, Association of Research Libraries, 2003.
- [4] W.-T. Wu, L. Yuan-Jie, A.-Z. Feng, L. Li, T. Huang and X. An-Ding, "Data mining in clinical big data: the frequently used databases, steps, and methodological models," *Military Medical Research*, vol. 8, no. 44, 11 August 2021.
- [5] S. A. Alasadi and W. S. Bhaya, "Review of Data Processing Rechniques in Data Mining," *Journal of Enginerring and Aplied Sciences*, vol. 12, no. 16, pp. 4102-4107, 2017.
- [6] M. J. Zaki and M. Wagner, "Data Mining and Analysis," in *Data Mining and Analysis: Fundamental Concepts and Algorithms*, Cambridge University Press, 2014, pp. 1-30.
- [7] U. M. Fayyad, "Visualizing Data Mining Models," *Information Visualization in Data Mining and Knowledge Discovery*, vol. 1, pp. 212-220, 2002.
- [8] H. Zhang, Data Mining for Business Applications, Springer Science & Business Media, 2008.
- [9] D. L. Olson, "Data mining in business services," *Service Business*, vol. 1, pp. 181-193, 2006.
- [10] G. S. Linoff and M. J. A. Berry, Data Mining Techniques, John Wiley & Sons, 2011.
- [11] K. I. Moin and Q. B. Ahmed, "Use of Data Mining in Banking," *International Journal of Engineering Research and Applications (IJERA)*, vol. 2, no. 2, pp. 738-742, 2012.
- [12] N. Jothi, "Data Mining in Healthcare – A Review," *Procedia Computer Science*, vol. 72, pp. 306-313, 2015.
- [13] L. Sadath, "Data Mining in E-Commerce: A CRM Platform," *International Journal of Computer Applications*, vol. 68, no. 24, 2013.
- [14] A. Saleem, K. H. Asif, A. Ali, S. M. Awan and M. A. Alghamdi, "Pre-processing Methods of Data Mining," in *IEEE International Conference on Utility and Cloud Computing (UCC)*, 2014.
- [15] E. G. Kulkarni and R. B. Kulkarni, "WEKA Powerful Tool in Data Mining," *International Journal of Computer Applications*, pp. 10-15, 2016.

- [16] C. Faloutsos, F. Korn, A. Labrinidis, Y. Kotidis, A. Kaplunovich and D. Perkovi'c, "Quantifiable Data Mining Using Principal Component Analysis," in *Technical Research Report*, Institute for Systems Research and Department of Computer Science.
- [17] S. Singhal and M. Jena, "A Study on WEKA Tool for Data Preprocessing,," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 2, no. 6, pp. 2278-3075, 2013.
- [18] I. Bruha , "Pre- and Post-processing in Machine Learning and Data Mining," *Machine Learning and Its Applications*, pp. 258-266, 2001.
- [19] A. K. Pandey and D. S. Rajpoot, "A comparative study of classification techniques by utilizing WEKA," in *International Conference on Signal Processing and Communication (ICSC)*, 2017.
- [20] Y. Zhongguo, L. Hongqi, S. Ali and A. Yile, "Choosing Classification Algorithms and Its Optimum Parameters," *Journal of Computers*, vol. 28, no. 5, pp. 26-38, 2017.
- [21] S. Ali, "On learning algorithm selection for classification," *Applied Soft Computing*, vol. 6, no. 2, pp. 119-138, 2006.
- [22] V. F. Hutter, *Automated Machine Learning*, Springer, 2019.
- [23] R. Dash, "Selection Of The Best Classifier From Different Datasets Using WEKA," *International Journal of Engineering Research and Technology (IJERT)*, vol. 2, no. 3, 2013.
- [24] ChatGPT, <https://chat.openai.com/>
- [25] Quillbot, <https://quillbot.com/>

9. Appendix

9.1 Full list of Attributes

Attribute	Description
Credit Risk	Indicates whether the credit contract was met (good) or not (bad) (binary).
Foreign Worker	Specifies if the debtor is a foreign worker (binary).
Telephone	Indicates if a telephone landline is registered in the debtor's name (binary; data from the 1970s).
People Liable	Number of individuals financially dependent on the debtor (binary).
Job	Quality level of the debtor's job (ordinal).
Credit Number	Number of credits the debtor has or had at this bank, including the current one (ordinal; discretized quantitative).
Housing	Type of housing where the debtor resides (categorical).
Other Instalment Plans	Presence of instalment plans from providers other than the credit-giving bank (categorical).
Age	Debtor's age in years (quantitative).
Property	Most valuable property owned by the debtor; the highest code is used. Code 2 applies if higher codes don't, but there's a car or other relevant property (ordinal).
Present Residence	Duration (in years) the debtor has lived in their current residence (ordinal; discretized quantitative).
Other Debtors	Indicates if there is another debtor or a guarantor for the credit (categorical).
Sex	Combined information on sex and marital status; sex cannot be determined as male singles and female non-singles are coded the same (2); female widows are not clearly classified.
Instalment Rate	Credit instalments as a percentage of the debtor's disposable income (ordinal; discretized quantitative).
Employment Duration	Length of employment with the current employer (ordinal; discretized quantitative).
Savings	Amount of savings the debtor has (categorical).
Purpose	Reason for which the credit is needed (categorical).
Credit History	Record of compliance with previous or current credit contracts (categorical).
Duration	Length of the credit in months (quantitative).
Status	Status of the debtor's checking account with the bank (categorical).

Figure 32. Full list of Attributes

DWDM Assignment 2 (Nathan Kyle Hudson)

GRADEMARK REPORT

FINAL GRADE

GENERAL COMMENTS

63/100

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

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

PAGE 35

PAGE 36

PAGE 37

PAGE 38

PAGE 39

PAGE 40