Unveiling the Power: A Comparative Analysis of Data Mining Tools through Decision Tree Classification on the Bank Marketing Dataset

ELIF AKKAYA[1], SAFIYE TURGAY[2]
[1]Department of Electric and Electronic Engineering,
Sakarya University,
54187, Esentepe Campus Serdivan-Sakarya,
TURKEY

[2]Department of Industrial Engineering,
Sakarya University,
54187, Esentepe Campus Serdivan-Sakarya,
TURKEY

Abstract: - The importance of data mining is growing rapidly, so the comparison of data mining tools has become important. Data mining is the process of extracting valuable data from large data to meet the need to see relationships between data and to make predictions when necessary. This study delves into the dynamic realm of data mining, presenting a comprehensive comparison of prominent data mining tools through the lens of the decision tree algorithm. The research focuses on the application of these tools to the BankMarketing dataset, a rich repository of financial interactions. The objective is to unveil the efficacy and nuances of each tool in the context of predictive modeling, emphasizing key metrics such as accuracy, precision, recall, and F1-score. Through meticulous experimentation and evaluation, this analysis sheds light on the distinct strengths and limitations of each data-mining tool, providing valuable insights for practitioners and researchers in the field. The findings contribute to a deeper understanding of tool selection considerations and pave the way for enhanced decision-making in data mining applications. Classification is a data mining task that learns from a collection of data to accurately predict new cases. The dataset used in this study is the Bank Marketing dataset from the UCI machine-learning repository. The bank marketing dataset contains 45211 instances and 17 features. The bank marketing dataset is related to the direct marketing campaigns (phone calls) of a Portuguese banking institution and the classification objective is to predict whether customers will subscribe to a deposit (variable y) in a period. To make the classification, the machine learning technique can be used. In this study, the Decision Tree classification algorithm is used. Knime, Orange, Tanagra, Rapidminerve, Weka yield mining tools are used to analyze the classification algorithm.

Key-Words: - Data Mining Tools, BankMarketing Dataset, Feature Selection, Performance Evaluation, Decision Trees, Evaluation Metrics.

Received: August 31, 2023. Revised: February 5, 2024. Accepted: March 7, 2024. Published: May 13, 2024.

1 Introduction

In a data mining environment where there is a limitless response in terms of data size and data sources, usefulness of the toolset you select is a primary tool for extracting valuable information. The investigation turns on the examination of the features of the variety of data mining tools by the means of decision tree methodology. And notably, the backdrop for this exploration is the BankMarketing dataset, which is a kind of treasure box of daily financial interactions that gives the right context for predictive analytics in banking.

Data mining is the realization from a complex dataset that is created from patterns, trends, and knowledge. Flooded with many of them as there are, you need to pick the most appropriate one. Decision tree classification, which is very commonly used in data mining, is popular for its interpretability and utility. The capacities of the AI include identifying complex associations in data, which can provide the most appropriate solution for this comparative analysis. The main target of the study is a careful analysis and evaluation of numerous data mining tools, by using the decision tree classifier on the BankMarketing dataset. By using this filter,

therefore, we would like to take a closer look at the special advantages as well as disadvantages of each tool. Particular performance metrics, for instance, the accuracy, precision, recall, and F1-score, will be used to judge the effectiveness of the method. The main goal of this in-depth study is to feed the existing data mining tool choice domain with informed perspectives and practical suggestions, thus helping the professionals and researchers in their decision-making processes.

The remainder of this paper unfolds as follows: Section 2 presents a comprehensive literature review, it proves the study in the research design, and showcases significant findings. In part 3, a detailed study of decision tree algorithm is presented along with any specific customized changes made for this experiment.

Section 4 provides an explanation for the criteria and rationale behind selecting the appropriate data mining tools used in that experimentation methodology. At the same time, it shows the available report metrics and compares the results. The results in Section 1 and the challenges and limits are given in Section 5.

2 Literature Survey

A study of data mining tools used in decision tree classifications is a vital part of BankMarketing dataset optimization. This literature analysis strives to disseminate the various powers that each data mining tool possesses as it is capable of identifying the strengths and weaknesses of each tool in addressing the complexities in this dataset. There have been many studies focusing on the importance of data mining tools in the different domains. [1], carried out the iris data set evaluation with the use of Weka, RapidMiner, ApacheSpark but they reported that Weka provided the best accuracy value with 98%. [2], evaluating iris data with 3513 records using SPSS-Clementine, RapidMiner and Weka AmritaNaika and LilavatiSamantb analysed Indian Liver disease patients data using the Decision tree, K-Nearest Neighbor, Naive Bayes algorithms with the help of WEKA, Rapidminer, Tanagra, Orange and Knime. Some researcher focused on the day to day running into the benefits of data mining tools in business and research, revealing the credibility in identifying trends and patterns, [3], [4], [5], [6], [7], [8], [9].

Decision tree classification has come of age as one of the most popular algorithim which comes with the feature of interpretability and usefulness, [10], [11], [12]. The study does the in-depth analysis of decision tree algorithms capacity to find meaningful patterns from the BankMarketing dataset of [13], [14], [15].

Through reviewing studies, the survey determines the data mining tools that were used in the study, the performance of the tools, interpretability factors, and if the tools were scalable enough. This tells us if these tools may have been used in the past and what was wrong with them and whether the tools are appropriate for use in this study [16], [17], [18], [19], [20].

On the other hand, the literature survey scrutinizes various integrated techniques and techniques applied in decision tree classification modeling for the prediction of banking datasets. The field is portrayed as ever-growing and constantly evolving with the development of new techniques; the unseen and unexpected difficulties encountered are recounted alongside the anticipated advancements in data mining tools capacity of forecasting [21], [22], [23], [24], [25], [26]. In the end, the study will show the best strategy to be applied for data mining BankMarketing data to find the necessary information and eventually find ways of improving the data mining methodologies in bank analytics. Indeed, the previous study gave a basis of the usage of data mining tools and decision tree classification algorithms, yet this paper will elaborate on a single dataset–BankMarketing– and provide in-depth analysis and comparison of some specific tools, [27], [28], [29]. The using of decision tree classifier on financial datasets gives new angle to the refrigerator which is relating to data mining technologies, [30], [31], [32], [33], [34], [35].

In the next few sections, we examine at BankMarketing data set, describe of decent tree categorizing algorithm, and explicitly describe how this comparison is conducted in detail. We target this in the sense that besides current knowledge, it becomes possible to have a practical view of the approach to the toolbar in data mining.

3 Methodology

This part gives a bill of fare on the step-by-step manner used to compare different Data Mining tools by using the Decision Tree Classification Algorithm on the BankMarketing data set. These features include not only demographic but also financial aspects, and a wide range of economic indicators and cover all outcome info of the campaign. It is very integral to have a complete grasp on the dataset before moving on to the next stage of analysis because this leaves no place to misunderstand the data.

The choice of parameters for decision tree classifier, that are, the number of tree splits, measure function and pruning strategy, had to be brought in a balance so as not to jeopardize the model for generalization.

Before launching the trainfulness of model, the dataset undertake elaborated correction of any deviance or missing element. Categorical variables were properly encoded and Features marked with a number were normalized by way of scaling of those features so that the model could learn correctly. Missing values were imputed using various techniques, and the outliers were either replaced and tagged for separate consideration or not. Each solicit was prepared with a set of standard options, specifications, and accessories; special attention was paid to the product compatibility for the Decision Tree Classification algorithm. The description for the tools used is composed of factors like the spread, ease of use, and accuracy, which are mostly accepted and to be relied on by the data mining community. The next step involves the classification decision tree algorithm which is used separately on the training set rather than as a combined approach through all selected data mining tools. To find this out, the models' generalization performance was evaluated by cross-leaving, using a specific routine, [e.g., k-fold cross-leaving]. To gauge the efficacy of each data mining tool, a comprehensive suite of performance metrics was employed. These metrics included accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). The choice of metrics aimed to provide a holistic evaluation of each tool's predictive capabilities across various dimensions.

In the study, the accuracy rates of WEKA, Rapidminer, Tanagra, Orange and Knime applications on the determined data sets were compared. Visualization features, data structures, platforms, transfer options of data mining tools were realized and presented (Figure 1). Classification was performed with the C4.5 algorithm of the decision tree. By using the accuracy parameter for analysis, it is aimed to determine how accurate unanalyzed data sets will give accurate results in applications and to help users choose the right data mining tool for data predictions.

RapidMiner: It is a data mining program developed by YALE University scientists in the USA as a result of programming with Java Programming Language. [www.rapidminer.com.].

Weka: It is a Data Mining program that was initially started as a small project and started to be used by many people all over the world. It is a product developed with Java Programming language, [7].

Knime: Konstanz Information Miner (KNIME) is a software developed by the visual data mining research group of the University of Konstanz on the EclipseRich Client Platform. It offers users a software development kit, [8].

Orange: It is software developed by the artificial intelligence research team of the Department of Computer and Informatics Sciences, University of Ljubljana, Slovenia, [4].

Tanagra: is an open source program that includes supervised learning algorithms, especially focusing on the visual and interactive construction of decision trees, [9]. Classification and clustering are among the most important methodologies used in data mining.

Decision tree: Decision trees, which produce class results by branching using the features in the dataset and the values of the features, are a frequently used method for trained learning, [10].

ID3 algorithm: The ID3 algorithm, an entropy-based algorithm, is the most basic and widely used algorithm of decisiontree. The goal of the ID3 algorithm is to keep the tree depth at a minimum level while creating the tree structure. The attributes whose complexity is determined to be minimum with entropy are added to the tree, and the data belonging to these attributes can be discrete data that can be counted or continuous data that can be measured, [1].

C4.5 algorithm: The algorithm J48 or C4.5 which allows both continuous and discrete features and is almost a similar to the algorithm ID3 and was developed to overcome the issues of ID3 algorithm, [1].

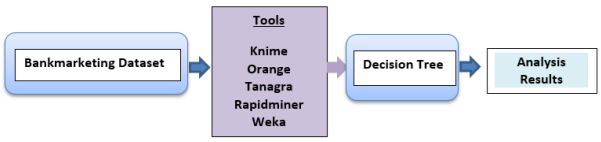

Fig. 1: Suggested study process

The outcomes of every data mining instruments were categorically compared, with the strengths and weaknesses of each belonging to their sphere outlined. Statistical tests like paired t-tests or Wilcoxon signed-rank tests were employed for evaluating the significance of the observed discrepancies in behavior performance. Along with the general performance metrics, the effects of variables selected or the features of the given dataset on the tools were assessed by sensitivity analyses as well. This approach was designed to reveal any

tools-caused sensitiveness or robustness in the situations where different conditions affect them.

This research embodies ethics guidelines in data gathering, interpretation, and analysis. The BankMarketing dataset has the private data amended and dealt with anonymity and the maximum confidentiality. The project is committed to abide by the rules and regulations associated with the protection of data and privacy. As a result the researchers can reproduce the findings and also the transparency in the research is clearly put forward. The following step is the comparison of results which will be then discussed and guided by the context of the tool selection in data mining.

3.1 Decision Tree

The decison tree predictor node generates predictions of each instance in th input data respectively. The predictions are a consequence of the decision rules that have been acquired in training the decision tree. The accuracy of the given output can be determined using several performance measurements, for instance, accuracy, precision, recall, and F1-score. This processing delivers the clarification of model's prediction ability on new data. For certain tools or platforms, the interface may contain features like visualizing decision tree structure or showing decision-making process. This visualization is data that users will understand the model specifically related to how it is predicting and explains the model's behavior. Which iteration will be applied depends on the evaluation results. The next step may involve training of the model, and refining the prediction performance. This repetitive process allows for successive correction of the model performance to perfection for the task implemented.

Undoubtedly, the Decision Tree Predictor node is a crucial part of the application of decision tree regression which makes it possible to translate learned patterns into planes of action that may be used to forecast for new data instances.

Accuracy stands as a key measure in estimating classification models' capability, including decision trees. It means of the correctly predicted instances that of the total instances in the dataset.

$$Accuracy = \frac{Number\ of\ Current\ Predictions}{Total\ Number\ of\ Predictions} \times 100$$

- One needs high precision to be sure that the model makes right predictions for a considerable part of instances. Accuracy is not the only thing because it is not sufficient in imbalance sets.

- Low accuracy score indicates that the model has a difficulty distinguishing between good and bad predictions. Hence trainees could face problems like overfitting, underfitting or the presence of noisy data.

- If the classes that exist in the dataset are imbalanced, accuracy is not the single criterion to study. Being able to suggest other parameters, such as precision, recall, and F1-score, can provide deeper insights, for instance, in case of the program trying to meet the needs of a specific class.

Accuracy needs to be reported on sufficiently, considering also other relevant metrics (if applicable) and keeping the context of specific problem in mind.

3.2 Confusion Matrix

A confusion matrix is an administrative document that many researchers use to denote the performance of the classification model when working on test data with known values. It is specifically good for learning what types of mistakes a model may be making.

- Instruction that asks for positive instances, and are well represented in the model output.

- Negative instances whose predictions as negative are correctly made by the model are all the real cases.

- Instances that are in reality not positive but are, however, incorrectly classified as positive by the model. To err is human, and this is known as a Type I error.

- Case of the positives that are actually good but are labeled as negative incorrectly by the model. And it is also referred to as Type II error.

The confusion matrix is placed in the Table 1 generally to make it easier to view.

Table 1. Confusion Matrix

	Predictive Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

From the confusion matrix, various performance metrics can be derived, including:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$
$$Specificity = \frac{TN}{TN + FP}$$
$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

The confusion matrix provides a clear and concise summary of a model's performance, allowing practitioners to assess both the overall accuracy and the specific types of errors made by the model. The idea is to develop a feature tree; the edges of which represent a tree where each internal node represent one decision based on a particular feature, each branch represents the outcome of the decision, and each leaf node represent the final prediction.

3.3 Decision Tree Classification Algorithm Implementation using KNIME Version 4.5.2

Decision tree classy algorithm was done by KNIME platform which was version 4.5.2 used for this study. KNIME, which is a free of charge data analytics platform that permits by an intuitive graphical interface a smooth and concurrent design of workflows working together with the data. Likewise, Bank Marketing was imported into KNIME which, let me access any necessary dataset or function as well as manipulate this dataset efficiently. Values that were missing were addressed properly through imputation techniques, the effect of which was considered for what remained of the data required for further analysis. A category paired labels were set up and numerical columns were standardized so as to maintain uniformity in the input.

Decision tree learner was used within KNIME to configure tree-based classification algorithm in this specific case. Parameters including tree depth and splitting criteria, and decisions on splitting options were optimized with reference to the analysis results and dataset characteristics. The prepared decision tree algorithm was applied to the training subset of BankMarketing dataset via the Decision Tree Learner node of the pytreebank package. Counterchecking of the model's generalization capabilities, k-fold cross-validation approach (with k = [Specify the number of folds]) was applied. Metrics protoformance adopted examples of accuracy, precision, recall, F1-score, and AUC-ROC were computed using dedicated nodes of the KNIME. However, this common set of steps was followed by the implementation of each tool for the purpose of providing similar experimental design and evaluation measures. Parametric sensitivities were looked at to assess the effect of overall performance of the decision tree algorithm after it was run through KNIME by changing specific parameters. KNIME was used as a whole workflow development environment, covering data preprocessing, decision tree configuration, model training, and verification of success, all with good transparency and reproducibility.

Consequently, after adding metrics and visuals, results were exported for a deeper analyses and data representation.

For implementation purposes, this study is using a version 4.5.2 of KNIME that provides a consistent and user-friendly environment thus accommodating both novice and experienced users to repeat the approach. The final part, thus, presents and discusses the resulting set of the findings, providing more light on the comparison of data mining tools that are the part of decision tree classification on the BankMarketing dataset.

3.4 Mathematical Modeling

Formulating a mathematical model for a comparative study of data mining tools and classifying these into decision tree category comprises identifying all the components that shall be included in the investigation. While the specifics of the model may vary based on the exact approach and algorithms used, here's a conceptual outline:While the specifics of the model may vary based on the exact approach and algorithms used, here's a conceptual outline:

- We have a dataset going by BankMarketing and it is represented as D.
- D is represented as a set of instances (x_i, y_i), wherex_i is a vector of features, andy_i is the corresponding class label.
- D is split into training set D_{train} and testing set D_{test} using a specified ratio.
- Let M_t represent the decision tree model using tool t.
- The model is trained on D_{train}using a decision tree classification algorithm.
- M_t predicts class labels for instances in D_{test}.
  Define metrics P_t(precision), R_t (recall), $F1_t$ (F1-score), and AUC_t (area under the ROC curve) for tool t.

Compare the performance metrics of each tool to determine their effectiveness in classification. Perform statistical tests (e.g., paired t-tests) to assess significant differences.

1. Accuracy (Acc)

$$\text{Acc}_t = \frac{\text{Number of Correct Predictions by } M_t}{\text{Total Number of Predictions by } M_t}$$

2. Precision (P)

$$P_t = \frac{\text{True Positives by } M_t}{\text{True Positives by } M_t + \text{False Positives by } M_t}$$

3. Recall (R)

$$R_t = \frac{\text{True Positives by } M_t}{\text{True Positives by } M_t + \text{False Negatives by } M_t}$$

4. F1 Score

$$F1_t = \frac{2 \times P_t \times R_t}{P_t + R_t}$$

Compute the AUC-ROC for the tool \(t\) based on the true positive rate and false positive rate. Use appropriate statistical tests to compare performance metrics between different tools. The mathematical model provides a systematic framework for comparing the performance of data mining tools through decision tree classification on the BankMarketing dataset. This is the model that helps to determine the degree of efficacy and reveals statistical assessment of each tool, thus covering holistic comparison.

4 Case Study

Summarise the essence of information extraction and Decision tree enabling for a better understanding of BankMarketing dataset Indubitably set objectives, to name a few, a comparison of data mining tools, a decision tree classification, and a crucial metrics assessment. Analysis of data mining tools, decision tree classification, applications considering bank datasets, and their utility. Thorough coverage of BankMarketing dataset features and the target variable (Orchestra). An overview of the data preprocessing steps involved, including, but not limited to, the handling missing values and decoding of the categorical variables.

A (Deep) explanation of the decision tree classification algorithm applied. Talk about changing the parameters of the experiment and tuning, if necessary. Data mining is the term for the process of identifying and using methods and tools that are used for data analysis and modeling. For example, KNIME, Weka, and RapidMiner are used as data mining tools. A discussion on how the identified tools are able to accomplish tasks and meet the unique demands of each user. An extensive description of the experiment set up which includes among others, data partition, decision tree modeling, and evaluation of the performance. Consistency is also important in this case since the paragraph has to start with a reference to the KNIME Version 4.5.2. Presentation of performances of accuracy, precision, recalls, and F1-score for each data mining tool.

Comparative analysis of results, highlighting strengths and weaknesses of each tool. Interpretation of findings, exploring reasons for variations in tool performance. Comparison of decision tree models generated by each tool (Figure 2). Figure 2 and Figure 3 show the screenshots of the Knime toolbox for the statistics and decision tree modules. At the same time, Figure 4 shows the Confusion Matrix result output.

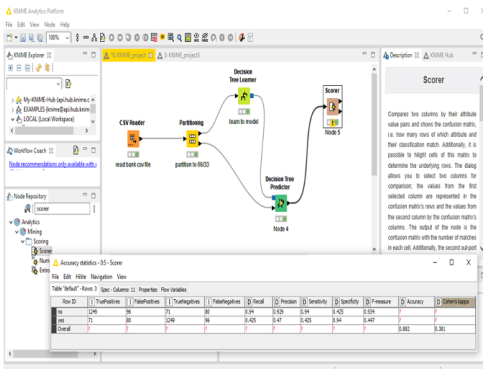

Fig. 2: Knime Accuracy Statistics

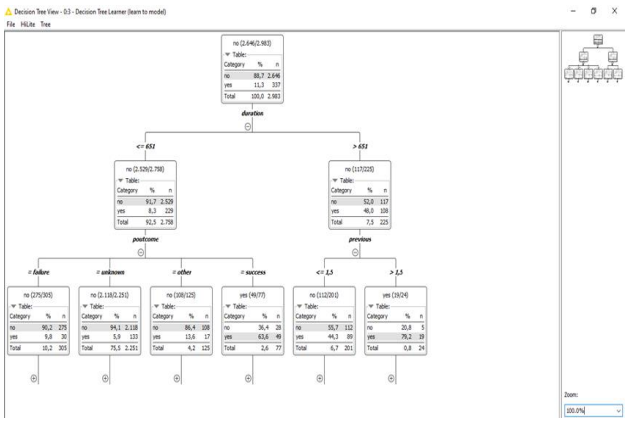

Fig. 3: Knime Decision Tree

Discussion of challenges encountered during the study. Acknowledgment of limitations in the experimental design or dataset. This case study aims to provide a holistic view of the comparative analysis of data mining tools through decision tree classification on the BankMarketing dataset, offering insights into the performance and applicability of each tool in a real-world scenario. Decision trees are valuable tools for exploratory data analysis and building understandable models. However, practitioners often need to consider potential overfitting and explore techniques like pruning or using ensemble methods to enhance their predictive capabilities.

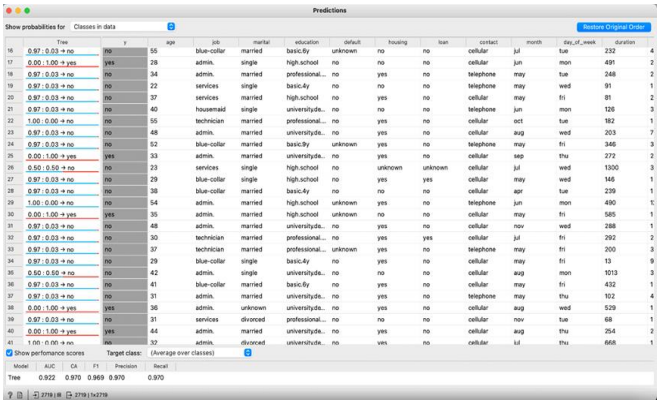

Fig. 4: Knime Performance Scores

Classification is a data mining task that learns from a collection of cases to accurately predict new cases.Many applications such as Weka, Tanagra, RapidMiner, Knime, and Orange use classification with decision trees.

The Bank marketing dataset used in the study, taken from the UCI site, contains 45211 instances and 17 features. The classification was performed on the Bank marketing dataset with the C4.5 algorithm of the decision tree. By using the accuracy parameter for analysis, it is aimed to determine how accurate unanalyzed data sets will give accurate results in applications and to help users choose the right data mining tool for data prediction. Neural Networks gave the best result with 98.66667%. Figure 5 and Figure 6 show the performance scores and confusion matrix screen outputs of Orange Toolbox.

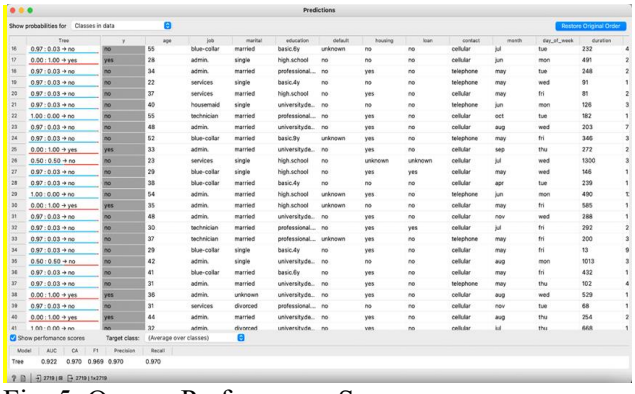

Fig. 5: Orange Performance Scores

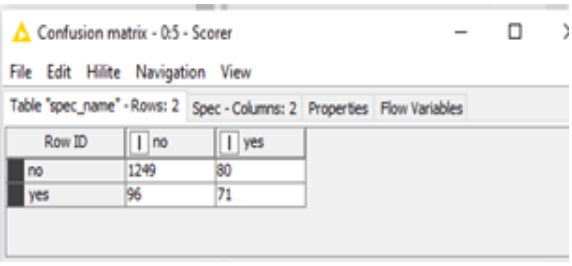

Fig. 6: Orange Confusion Matrix

A brief summary of the key findings from the comparative analysis of data mining tools through decision tree classification on the BankMarketing dataset. Presentation of accuracy, precision, recall, and F1-score for each data mining tool. Comparative analysis of these metrics to highlight the strengths and weaknesses of each tool in predictive modeling. Figure 7 shows the Tanagra classifier performance results.

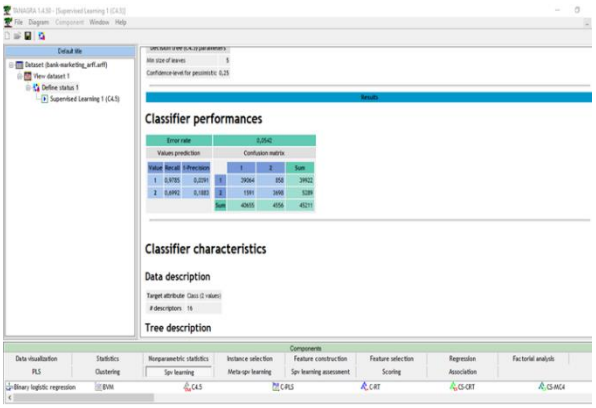

Fig. 7: Tanagra Classifier Performances

Detailed examination of accuracy statistics for each tool, including true positives, true negatives, false positives, and false negatives. Visualization of the confusion matrices to provide a clear understanding of the model's predictive performance. Visualization of key decision nodes and branches to showcase the interpretability and complexity of the resulting models. Exploration of the sensitivity of each data mining tool to variations in parameters or dataset characteristics. Insights into the robustness and adaptability of the models. Figure 8 and Figure 9 show the Rapidminer Tree Description and Rapidminer Decision Tree screen outputs.

Application of statistical tests (e.g., paired t-tests) to assess the significance of observed differences in performance metrics. Identification of tools that significantly outperform others. In-depth discussion and interpretation of the analysis results. Insightful explanations for observed variations in performance and decision tree structures. Figure 10 shows the Weka Classifier Output results.

Application of statistical tests (e.g., paired t-tests) to assess the significance of observed differences in performance metrics. Identification of tools that significantly outperform others. In-depth discussion and interpretation of the analysis results. Insightful explanations for observed variations in performance and decision tree structures. Figure 10 shows the Weka Classifier Output results.

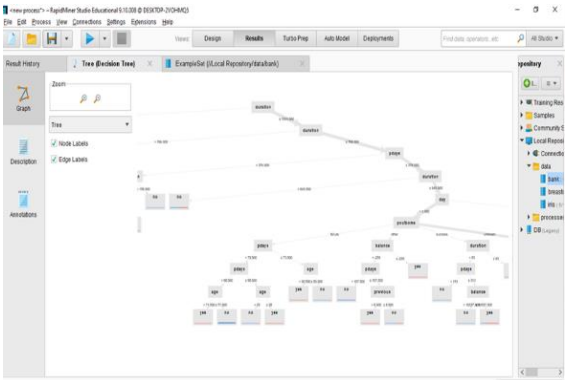

Fig. 8: Rapidminer Decision Tree

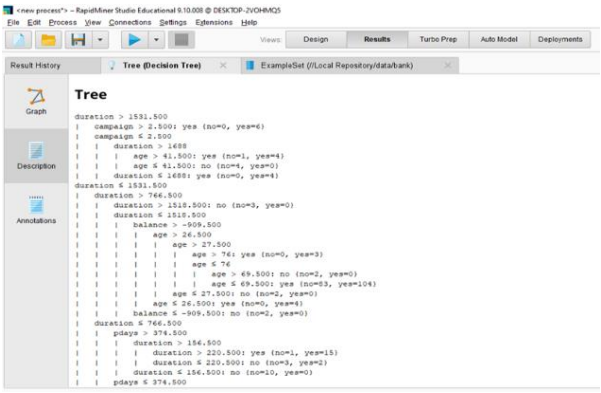

Fig. 9: Rapidminer Tree Description

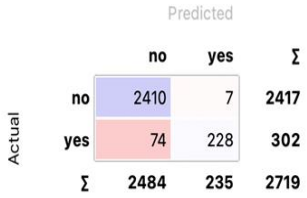

Fig. 10: Weka Classifier Output

Highlighting specific considerations for each data mining tool. Recommendations for tool selection based on the context of the BankMarketing dataset and decision tree classification. Discussion of the practical implications of the analysis results in the context of the banking domain. How the findings can inform decision-making processes and improve predictive modeling in financial institutions. Reflection on challenges encountered during the analysis. Acknowledgment of any limitations in the study and their potential impact on the results C4.5 Decision Tree algorithm was applied to the bank marketing dataset in Knime, Orange, Tanagra, Rapidminer, and Weka tools. When the accuracy values of the data tools were compared, it was observed that the Orange data tool gave the best result. According to the accuracy results, it was concluded that the use of the Orange data tool would be appropriate for datasets similar to the Bank Marketing dataset to be classified with the decision tree algorithm. A study that can help researchers and users to choose the right tool and technique for data analysis and prediction has been created (Table 2).

Table 2. Data tools and accuracy results

TOOL	ACCURACY
KNIME	0.882
ORANGE	0.970
TANAGRA	0.838
RAPIDMINER	0.906
WEKA	0.903

Recommendations for future research based on the insights gained. Suggestions for refining the methodology or exploring additional dimensions in subsequent studies. Putting in weight on the role of research which is central to the appraisal of decision tree tasks tool for data mining. By the mean of the obtained results researchers and those who are in practice get complete knowledge about the performance of various tools of data mining with the selected dataset. As a result, the practice can be improved by practitioners and academicians of the research field pertaining to data mining and predictive modeling.

5 Conclusion

Our research developed around a mission to expose the modus operandi of different data mining tools via a thorough delving into a classifier we encountered in a decision tree on the BankMarketing dataset. The main purpose was to determine how effective each method was in developing predictor models and to disclose the unique advantages and disadvantages of each tool. Let's validate the major discoveries and draw conclusions after our study has been over.

From my comparison analysis, it was established that there were differing metrics of performance across the data mining tools statistics that included accuracy, precision, recall, and F1-score. Unequal tree breeding has decision trees with their own specific traits and as a result, they affect model understandability and complexity. The features of the selected tools were also elaborated, having reminded us that the context of decision tree classification must be taken into consideration. The outcomes of the analysis are of pertinent significance to the decision-making processes in the banking domain given that informed predictive modeling is input for these tasks. The advantages

that seem superior will be encouraged for application in certain types of situations or scenarios.

Practitioners are advised to choose data mining tools by striking a balance between the kind of task involved each time and the tool most appropriate for executing it. Through details regarding tool-oriented side notes such as those offered in the analysis, stakeholders can be enabled to make informed choices.

The research paper does not downplay the fact that the research had certain limitations and encountered some challenges during the analysis, thus pointing to the need to consider them in the conclusion of the findings. This shows new research paths that further investigation of methods improvement, development other algorithms, numerous datasets, and what-all may be done. One facility of this study is that it paves the way for further improvement in the development of predictive models.

This analysis therefore provides further insight into the respective mechanisms and choices using a decision tree classifier. Learning these lessons opens to the end users a valuable view that they can use in the process of choice of the tools readily available. By the final part of the comparative analysis, it becomes evident that the instrumentality of data in classification is profound. In a certain sense, this study is not only a theoretical but also a practical gain as it helps us to understand the situation better and gives practical recommendations to people who are trying to operate in the predictive modeling environment that surrounds them.

In the quest to unveil the power within the realm of data mining, this study serves as a beacon, illuminating the pathways toward informed decision-making and enhanced predictive modeling capabilities. The journey continues, beckoning future researchers to build upon these insights and propel the field toward new horizons.

References:

[1] Dušanka, D., Darko S., Srdjan, S., Marko, A., Teodora, L., "A Comparison of Contemporary Data Mining Tools", XVII International Scientific Conference on Industrial Systems (IS'17), Novi Sad, Serbia.

[2] Moghimipour, I., Ebrahimpour, M., "Comparing Decision Tree Method Over Three Data Mining Software," *Int. J. Stat. Probab.*, vol. 3, no. 3, pp. 147–156, 2014, doi: 10.5539/ijsp.v3n3p147.

[3] Naik A., Samant, L., "Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime," P*rocedia Comput. Sci.*, vol. 85, pp. 662–668, Jan. 2016, doi: 10.1016/J.PROCS.2016.05.251.

[4] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H., "The WEKA data mining software," ACM SIGKDD *Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009, doi: 10.1145/1656274.1656278.

[5] Berthold M. R., "KNIME: The konstanz information miner," *4th Int. Ind. Simul. Conf. 2006*, ISC 2006, vol. 11, no. 1, pp. 58–61, 2006, doi: 10.1145/1656274.1656280.

[6] Afifi, M. A., Ghazal, T. M., Afifi, M. A. M. , Kalra, D., "Data Mining and Exploration: A Comparison Study among Data Mining Techniques on Iris Data Set Linux Desktop View project E-GOVERNANCE View project Data Mining and Exploration: A Comparison Study among Data Mining Techniques on Iris Data Set," *Talent Dev. Excell.*, vol. 12, no. 1, pp. 3854-3861, 2020.

[7] Duan, J., Wang, G., Hu, X., Xia, D., Wu, D., Mining Multigranularity Decision Rules of Concept Cognition for Knowledge Graphs Based On Three-Way Decision, *Information Processing & Management*, Vol. 60, Issue 4, July 2023, 103365.

[8] Yiğit, S., Turgay, S., Cebeci, Ç., Kara, E.S., Time-Stratified Analysis of Electricity Consumption: A Regression and Neural Network Approach in the Context of Turkey", *WSEAS Transactions on Power Systems*, vol. 19, pp. 96-104, 2024, doi:10.37394/232016.2024.19.12.

[9] Kayali, S., Turgay, S., Predictive Analytics for Stock and Demand Balance Using Deep Q-Learning Algorithm. *Data and Knowledge Engineering*, (2023) Vol. 1: 1-10, doi: 10.23977/datake.2023.010101.

[10] Towell, G. G., Shavlik, J. W., Noordeweir, M. O., "Refinement of Approximate Domain Theories by Knowledge-Based Neural Networks," *Proc. Eighth Natl. Conf. Artif. Intell.*, pp. 861–866, 1990, [Online]. https://www.aaai.org/Library/AAAI/1990/aaai90-129.php (Accessed Date: May 2, 2024).

[11] Borges, L. C., Marques, V. M., Bernardino, J., "Comparison of data mining techniques and tools for data classification," *ACM Int. Conf. Proceeding,* Ser., no. October 2014, pp. 113–116, 2013, doi: 10.1145/2494444.2494451.

[12] Charbuty, B., Abdulazeez, A., "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[13] Jin, C., Li, F., Ma, S., Wang, Y., Sampling Scheme-Based Classification Rule Mining Method Using Decision Tree İn Big Data Environment, *Knowledge-Based Systems,* Vol. 244, 23 May 2022, 108522.

[14] Shoo, T.R., Patra, Vipsita, S., Decision Tree Classifier Based on Topological Characteristics of Subgraph for The Mining of Protein Complexes from Large Scale PPI Networks, *Computational Biology and Chemistry*, Vol. 106, October 2023, 107935

[15] Munoz-Rodriguez, J.M.P., Alonso, Pessoa, T., Martin-Lucas, J., Identity Profile Of Young People Experiencing A Sense Of Risk On The Internet: A Data Mining Application Of Decision Tree With Chaid *Algorithm, Computers & Education*, Vol. 197, May 2023, 104743.

[16] Reddy, R., Girija, S.P., Venkatramulu, S., Dorthi, K., Rao, V.C.S.V., A Gradient Boosted Decision Tree with Binary Spotted Hyena Optimizer for Cardiovascular Disease Detection and Classification, *Healthcare Analytics*, Vol. 3, November 2023, 100173

[17] Rahman, R.M., Hasan, F.R., Using And Comparing Different Decision Tree Classification Techniques for Mining Icddr,B Hospital Surveillance Data, *Expert Systems with Applications*, Vol. 38, Issue 9, September 2011, pp.11421-11436

[18] Naik, A., Samant, L., Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime, *Procedia Computer Science*, Vol. 85, 2016, pp.662-668.

[19] Macuacua, J.C., Centeno, J.A.S., Amisse, C., Data Mining Approach for Dry Bean Seeds Classification, *Smart Agricultural Technology*, Vol. 5, October 2023, 100240.

[20] Jurczuk, K., Czajkowski, M., Kretowski, M., Adaptive in-memory representation of decision trees for GPU-accelerated evolutionary induction, *Future Generation Computer Systems*, Vol. 153, April 2024, pp.419-430.

[21] Koulinas, G., Paraschos, P., Koulouriotis, D., A Decision Trees-based knowledge mining approach for controlling a complex production system, *Procedia Manufacturing*, Vol. 51, 2020, pp.1439-1445.

[22] Manzella, F., Pagliarini, G., Sciavicco, G., Stan, I.E., The voice of COVID-19: Breath and cough recording classification with temporal decision trees and random forests, *Artificial Intelligence in Medicine*, Vol. 137, March 2023, 102486.

[23] Ramakrishnan, J., Liu, T., Zhang, F., Seshadri, K., Yu, R., Gou, Z., A decision tree-based modeling approach for evaluating the green performance of airport buildings, *Environmental Impact Assessment Review*, Vol. 100, May 2023, 107070.

[24] Ghiasi, M.M., Zendehboudi, S., Application of decision tree-based ensemble learning in the classification of breast cancer, *Computers in Biology and Medicine*, Vol. 128, January 2021, 104089.

[25] Ghane, M., Ang, M.C., Nilashi, M., Sorooshian, S., Enhanced decision tree induction using evolutionary techniques for Parkinson's disease classification, *Biocybernetics and Biomedical Engineering*, Vol. 42, Issue 3, July–September 2022, pp.902-920.

[26] Mariano, A.M., Ferreira, A.M.L., Santos, M.R., Castilho, M. L., Bastos, A.C.F.L.C., Decision trees for predicting dropout in Engineering Course students in Brazil, *Procedia Computer Science*, Volume 214, 2022, pp.1113-1120.

[27] Hamdi, M., Hilali-Jaghdam, I., Elnaim, B.E., Elhag, A.A., Forecasting and classification of new cases of COVID 19 before vaccination using decision trees and Gaussian mixture model, *Alexandria Engineering Journal*, Vol. 62, January 2023, pp.327-333.

[28] Martinez-Rojas, A., Jimenez-Ramirez, A., Enriquez, J.G., Reijers, H.A., A screenshot-based task mining framework for disclosing the drivers behind variable human actions, *Information Systems*, Vol. 121, March 2024, 102340.

[29] Fa, H., Shuai, B., Yang, Z., Niu, Y., Huang, W., Mining the accident causes of railway dangerous goods transportation: A Logistics-DT-TFP based approach, *Accident Analysis & Prevention*, Vol. 195, February 2024, 107421.

[30] Naik, D.A., Burunda, C.J., Seea, S.D., A Feasible Dashboard to predict Patent Mining Using Classification Algorithms, *Procedia Computer Science*, Vol. 167, 2020, pp.2011-2021.

[31] Varra, M.O., Husakova, L., Patocka, J., Ghidini, S., Zanard,, E., Classification of Transformed Anchovy Products based on the

Use of Element Patterns and Decision Trees to Assess Traceability and Country of Origin Labelling, *Food Chemistry*, Vol. 360, 30 October 2021, 129790

[32] Ganti, P.K., Naik, H., Barada, M.K., Environmental impact Analysis and Enhancement of Factors Affecting the Photovoltaic (PV) Energy Utilization in Mining Industry by Sparrow Search Optimization Based Gradient Boosting Decision Tree Approach, *Energy*, Vol. 244, Part A, 1 April 2022, 122561

[33] Rutkowski, L., Jaworski, M., PiPietruczuk, L., Duda, P., The CART Decision Tree for Mining Data Streams, Information Sciences, Volume 266, 10 May 2014, pp.1-15.

[34] Quash, Y., Kross, A., Jaeger, J.A., Assessing the impact of Gold Mining on Forest Cover in the Surinamese Amazon from 1997 to 2019: A Semi-Automated Satellite-Based Approach, *Ecological Informatics*, Vol. 80, May 2024, 102442.

[35] Dash, C.S:K., Behera, A.K., Dehuri, S., Ghosh, A., An Outliers Detection and Elimination Framework in Classification Task of Data Mining, *Decision Analytics Journal*, Vol. 6, March 2023, 100164

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

- E.Akkaya, S.Turgay – investigation,
- E.Akkaya- validation and
- S.Turgay writing & editing.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare.