A Journal of Culture, English Language, Teaching & Literature
ISSN 1414-3320 (Print), ISSN 2502-4914 (Online)
Vol. 23 No.2; December 2023

A Comparative Analysis of Decision Tree Classification Algorithms for Blended Learning Analytics in WEKA

[1]Marivic Mitschek and [2]Rosanna Esquivel

[1,2]Angeles University Foundation, Graduate School, and De La Salle University-Dasmariñas, Information Technology Department, Dasmariñas, Philippines

[1]mrmitschek@dlsud.edu.ph, [2]rosanna.esquivel@gmail.com

A Comparative Analysis of Decision Tree Classification Algorithms for Blended Learning Analytics in WEKA

[1]Marivic Mitschek and [2]Rosanna Esquivel

[1]mrmitschek@dlsud.edu.ph, [2]rosanna.esquivel@gmail.com

Graduate School of Angeles University Foundation and Information Technology Department of De La Salle University-Dasmariñas, Philippines

Abstract: This study explores the application of decision tree classification algorithms for analyzing student performance data within a blended learning environment. The analysis, conducted using WEKA 3.8.6, focused on four attributes believed to influence student performance: course type, course level outcome (CLO), topic learning outcome (TLO), and level of assessment. A comparative analysis of J48, Random Forest, and SimpleCart algorithms revealed valuable insights. J48 demonstrated efficiency in model building, while Random Forest offered a balance between interpretability and accuracy. SimpleCart achieved the highest classification accuracy but could be less interpretable. The selection of the optimal algorithm depends on the analytical goals. J48 is suitable for rapid exploration, while SimpleCart prioritizes accuracy. Random Forest offers a compromise for scenarios where both understanding and accuracy are important. This study provides a foundation for understanding student performance through decision trees and highlights opportunities for further exploration using additional attributes, rule-based learners, and other machine-learning algorithms. By leveraging these techniques, educators within blended learning environments can gain a deeper understanding of student performance and tailor their practices to optimize learning outcomes.

Key words: blended learning, learning analytics, decision tree algorithms, J48, random forest, simplecart

Abstrak: *Penelitian ini mengeksplorasi penerapan algoritma klasifikasi pohon keputusan untuk menganalisis data kinerja siswa dalam lingkungan pembelajaran campuran. Analisis yang dilakukan menggunakan WEKA 3.8.6 berfokus pada empat atribut yang diyakini memengaruhi kinerja siswa: jenis kursus, hasil tingkat kursus (CLO), hasil pembelajaran topik (TLO), dan tingkat penilaian. Analisis komparatif algoritma J48, Random Forest, dan SimpleCart mengungkapkan wawasan yang berharga. J48 menunjukkan efisiensi dalam pembuatan model, sementara Random Forest menawarkan*

keseimbangan antara interpretabilitas dan akurasi. SimpleCart mencapai akurasi klasifikasi tertinggi tetapi kurang dapat diinterpretasikan. Pemilihan algoritma yang optimal tergantung pada tujuan analisis. J48 cocok untuk eksplorasi cepat, sedangkan SimpleCart mengutamakan akurasi. Random Forest menawarkan kompromi untuk skenario yang mengutamakan pemahaman dan akurasi. Studi ini memberikan landasan untuk memahami kinerja siswa melalui pohon keputusan dan menyoroti peluang untuk eksplorasi lebih lanjut menggunakan atribut tambahan, pembelajar berbasis aturan, dan algoritma pembelajaran mesin lainnya. Dengan memanfaatkan teknik-teknik ini, para pendidik dalam lingkungan pembelajaran campuran dapat memperoleh pemahaman yang lebih mendalam tentang kinerja siswa dan menyesuaikan praktik mereka untuk mengoptimalkan hasil pembelajaran.

Kata kunci: *blended learning, learning analytic, algoritma pohon keputusan, J48, random forest, simplecart*

INTRODUCTION

Blended learning offers a powerful solution for enriching the English language learning experience of overseas students. By strategically integrating online components with traditional face-to-face instruction (Graham et al., 2013), this approach bridges geographical divides. Overseas students benefit from asynchronous access to learning materials and activities hosted on a learning management system (LMS), overcoming time zone differences and fostering self-directed learning. Blended learning also facilitates personalized instruction through targeted online exercises that address individual needs. Face-to-face sessions then provide opportunities for instructors to offer personalized feedback and guidance. Furthermore, online forums and discussion boards within the LMS create a virtual space for overseas students to connect and collaborate with peers, fostering a sense of community and enhancing crucial intercultural communication skills. The incorporation of multimedia resources like interactive exercises, podcasts, and online games caters to the digital learning preferences of overseas students, promoting engagement and interactivity in the process.

Effectively gauging student learning outcomes in these hybrid environments remains a challenge. Learning analytics (LA) emerges as a powerful tool to address this gap (Ferguson, 2012). LA encompasses the measurement, collection, analysis, and reporting of data on learners' interactions within a learning management system (LMS) (Siemens & Long, 2011). Educators can identify areas of difficulty, gain important insights about students' development, and customize the learning process by utilizing LA approaches (Lang et al., 2017). An especially valuable LA method uses classification algorithms. Based on a variety of variables taken from LMS data, including student learning habits, performance indicators, and dropout risk, these algorithms classify students (Sadiq et al., 2014). This paper explores the use of decision tree classification algorithms for blended learning data analysis, with a focus on the Waikato Environment for Knowledge Analysis (WEKA) platform.

WEKA is an open-source software suite written in Java, designed for data mining tasks and knowledge discovery (Hall et al., 2009). For data preparation, classification, regression, clustering, association rule mining, and visualization, it provides a range of machine learning tools and

techniques. Notably, WEKA provides a user-friendly interface for implementing decision tree algorithms. These algorithms are tree-like structures where each internal node represents a test on a single attribute (e.g., quiz score), and each branch represents the outcome of that test. Leaves of the tree represent the predicted class labels (e.g., high performer, at-risk student).

The power of WEKA in this context lies in its ability to simplify the application of decision tree algorithms for educators with limited technical expertise. WEKA offers a graphical user interface (GUI) that eliminates the need for writing complex code. Users can import data directly from their LMS (assuming proper data export formats) and select the desired decision tree algorithm from a menu. WEKA then handles the heavy lifting – building the tree model, performing the classification, and presenting the results in a visual format. Additionally, WEKA provides interpretable outputs, allowing educators to understand the decision-making logic behind the classifications. By analyzing these decision trees, educators can gain insights into the factors that influence student performance within the blended learning environment. This knowledge empowers them to personalize instruction, identify students at risk, and ultimately optimize the learning experience for overseas students.

DECISION TREE CLASSIFICATION: A FOUNDATIONAL APPROACH

One particularly effective learning analytics (LA) technique involves classification algorithms. These algorithms function as supervised learning models that categorize students based on various factors extracted from LMS data. These factors can include learning styles, performance metrics, or even the risk of dropping out, as identified through sentiment analysis or login frequency (Sadiq et al., 2014). A popular LMS platform like Moodle provides a treasure trove of data suitable for such analysis. This data can encompass quiz scores, forum participation, assignment submissions, and even time spent on specific learning modules. By leveraging WEKA, educators can utilize decision trees to analyze this rich LMS data and gain valuable insights into student learning patterns within the blended learning environment.

Decision trees are supervised learning algorithms that employ a tree-like model for data classification. Each node in the tree represents a decision point, where a specific attribute of the learning data (e.g., time spent on online modules, number of forum posts) is evaluated using a splitting criterion. The branches emanating from each node represent the possible outcomes of this decision. By recursively following branches based on attribute values, the algorithm traverses the tree until it reaches a final leaf node, which assigns a predicted class label (e.g., "high risk," "low engagement"). Extracting data from a learning management system (LMS) like Moodle allows for an in-depth analysis of student performance within a blended English learning environment. WEKA, a software suite for data mining tasks, facilitates the application of decision tree algorithms for this purpose. These algorithms build tree-like models where internal nodes represent tests on LMS data attributes (e.g., quiz scores, forum participation) and branches represent the outcome of those tests. Leaf nodes denote predicted class labels like "course success" or "course risk." For instance, a decision tree might analyze performance on a placement test, followed by forum participation for students scoring below average. High participation could indicate an "engaged student" branch leading to "course success" if video lecture quizzes and adaptive learning software usage are also high. Conversely, a "disengaged student" branch with low participation and minimal software usage might predict "course risk." By analyzing these decision paths, educators gain insights into factors influencing student success. This knowledge empowers them to identify at-risk students, optimize technology usage within the blended

environment, and tailor instruction based on individual learning styles and needs revealed by the decision tree analysis.

The figure presents an example of a decision tree framework for analyzing and addressing late-night craving scenarios faced by students.

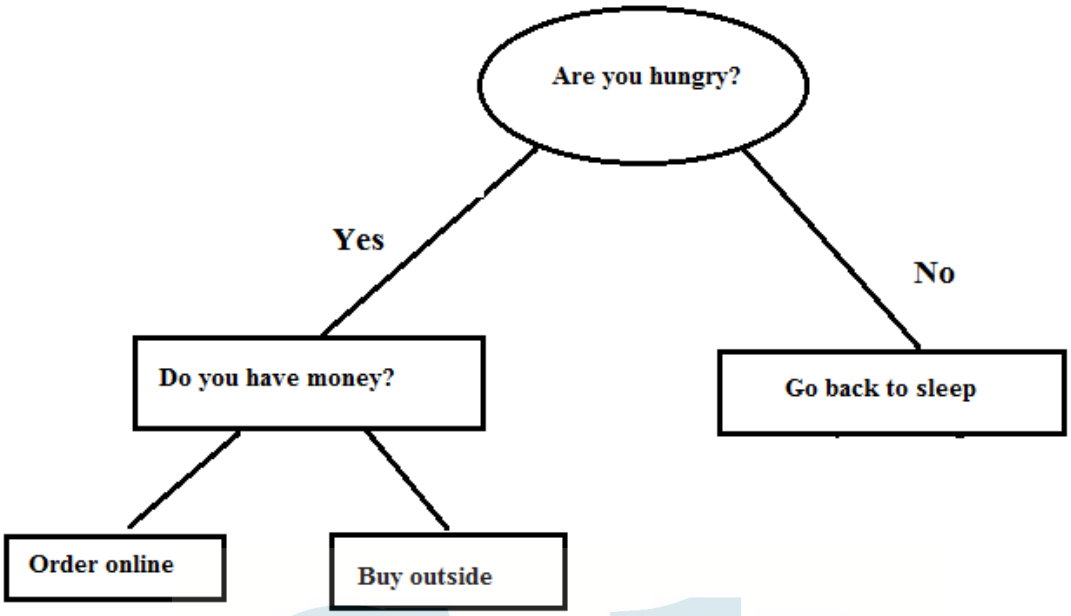

Figure 1:
Decision Tree Classification for late night cravings

As exemplified, the decision tree classification starts with the question, "Are you hungry?" and gives two possible answers, which are either "Yes" or "No". In the sample of "Yes" as an answer, two possible answers are also given, ie. either "order online" or "buy outside". If, however, the chosen answer is "No", then the follow-up action is "Go back to sleep".

WEKA: A USER-FRIENDLY PLATFORM FOR DECISION TREE ANALYSIS

The Waikato Environment for Knowledge Analysis (WEKA) stands as a prominent open-source software suite (Hall et al., 2009) widely adopted for its extensive collection of machine-learning tools (Holmes et al., 1994). Among its rich functionalities, WEKA offers a diverse array of decision tree algorithms, making it a valuable platform for analyzing blended learning data (Romero & Ventura, 2013). This variety is crucial for the decision tree classification approach. Educators can leverage WEKA to experiment with different algorithms like J48 (optimized for accuracy) or SimpleCart (known for interpretability). This flexibility allows them to select the algorithm that best suits their analysis goals, prioritizing interpretability for understanding student behavior or accuracy for pinpointing at-risk students. Furthermore, WEKA facilitates the exploration of multiple algorithms on the same data set, enabling comparison of resulting decision trees and potentially uncovering alternative insights into student performance within the blended learning environment. Ultimately, WEKA's diverse decision tree toolkit empowers educators to tailor the classification approach to their specific context and optimize the extraction of knowledge from LMS data.

A. Key Advantages of WEKA for Educators:

There are three key advantages of WEKA for educators: (1) accessibility, (2) decision tree algorithm variety, and (3) streamlined work analysis. The following are the details of each.

1. Accessibility: WEKA's intuitive graphical user interface (GUI) presents a significant advantage, particularly for educators with limited programming expertise (Bouckaert et al., 2010). This user-friendly interface empowers them to readily utilize the power of decision tree algorithms without delving into complex coding requirements.

2. Decision Tree Algorithm Variety: WEKA incorporates a vast repertoire of decision tree algorithms, including J48 (Quinlan, 1993), Random Forest (Breiman, 2001), and SimpleCart (Karatas & Verhoef, 2018). impacting online teaching success through their varying strengths. J48 prioritizes speed and interpretability, offering a quick understanding of student success factors but potentially sacrificing accuracy. Random Forest balances interpretability and accuracy by combining multiple decision trees, making it suitable for educators seeking both insights and reliable predictions. SimpleCart prioritizes accuracy with complex decision trees, delivering highly precise results but potentially hindering interpretability. This trade-off empowers educators to choose the algorithm that aligns with their goals. J48 suits initial exploration, Random Forest offers balance, and SimpleCart excels in pinpoint accuracy but demands more interpretation effort. This extensive selection allows educators to select the most suitable algorithm based on the specific characteristics of their blended learning data and the desired analytical goals.

3. Streamlined Analysis Workflow: WEKA streamlines the entire data analysis workflow. It facilitates data pre-processing tasks like cleaning, normalization, and transformation (Witten et al., 2016). Additionally, it offers functionalities for model building, evaluation, and visualization (Hall et al., 2009). This comprehensive suite of tools enables educators to efficiently conduct their analysis within a single platform. WEKA's interface simplifies decision tree analysis for online learning platforms. Educators can extract student data (quiz scores, forum activity, logins) from their LMS and import it into WEKA. Algorithms like J48 build decision trees to identify at-risk students based on these factors. By uncovering patterns (e.g., low quiz scores and forum participation predict struggle), educators gain actionable insights. This allows for targeted interventions like personalized feedback or additional materials, ultimately improving the online learning experience.

B. Impact on Blended Learning:

By leveraging WEKA's user-friendly interface and diverse decision tree algorithms, educators within blended learning environments can gain valuable insights into student learning patterns. These insights can be used to:

1. Identify At-Risk Students: Decision tree models can assist in identifying students at risk of dropping out or performing poorly by analyzing factors like participation levels, quiz scores, and time spent on learning materials (Sadiq et al., 2014).

2. Personalize Learning Experiences: Educators can tailor learning experiences by catering to individual student needs based on the classifications generated by decision tree models. This can involve providing additional support for struggling students or offering

advanced challenges for high performers (Ferguson, 2012). WEKA's decision tree outputs can further guide differentiated instruction. Struggling students flagged by the model can receive targeted interventions like one-on-one tutoring or scaffolded activities. Conversely, high performers can be offered enrichment activities like independent research or project-based learning, extending their knowledge beyond the core curriculum.

3. Improve Blended Learning Design: By analyzing student performance data through decision trees, educators can identify areas where the blended learning design might be hindering progress. This knowledge can inform modifications to the curriculum, delivery methods, or online resources to optimize the learning environment (Lang et al., 2017). WEKA's data-driven approach can inform not only differentiated instruction but also broader improvements to the online learning environment. Educators can leverage insights to modify the curriculum, delivery methods, or online resources (Lang et al., 2017). This continuous optimization cycle is exemplified by Filipino classes in the Philippines. Blended learning, which combines traditional classroom instruction with interactive online activities, has revitalized these classes. Platforms like Quizizz allow for gamified learning experiences, while educational simulations on websites can provide immersive cultural exploration. These online applications, coupled with in-person discussions and activities, create a more engaging and effective learning experience for Filipino students.

COMPARATIVE ANALYSIS OF DECISION TREE ALGORITHMS IN WEKA

WEKA offers a diverse range of decision tree algorithms, each with unique characteristics and suitability for specific scenarios. Here's a comparative analysis of some commonly employed algorithms:

1. CART (Classification And Regression Trees): CART forms the foundation for many decision tree algorithms. It is renowned for its simplicity, interpretability, and ability to handle both categorical and numerical data. CART employs the Gini impurity measure to select the optimal splitting attribute at each node, maximizing information gain and classification purity (Breiman et al., 1984). Gini impurity reflects how mixed up the data is at a particular point in the tree. Lower Gini signifies a clearer separation (like apples vs oranges), making future classifications more accurate. This essentially helps CART ask the most informative questions to build a strong decision tree.

2. J48: J48 is an implementation of the C4.5 decision tree algorithm, known for its efficiency and accuracy in handling large datasets. It utilizes a gain ratio criterion for attribute selection, which incorporates a penalty for datasets with a high number of branches, promoting a balance between information gain and tree complexity (Breiman et al., 1984). Imagine sorting words by grammatical function (noun, verb, etc.). J48 doesn't just pick the easiest feature, like capitalization (which might create many subcategories). Instead, it considers how many questions it'll take overall. It might ask "Does it end in -ing?" to efficiently group verbs, even if it requires a few more steps than a simpler question. This keeps the organization clear and efficient, making J48 a great tool for handling complex English data.

https://doi.org/10.24167/celt.v23i2; ISSN: 1412-3320 (print); ISSN: 2502-4914 (online); Accredited; DOAJ

3. Random Forest: Random Forest is an ensemble learning technique that combines multiple decision trees generated using random subsets of features and data points. This ensemble approach often leads to improved performance and robustness compared to single decision trees, especially for complex classification problems. Each tree analyzes random subsets of data and features (e.g., quiz scores, participation) to predict student performance. This ensemble approach improves accuracy, especially for complex tasks like predicting essay writing struggles. By identifying at-risk students beforehand, educators can provide targeted interventions like workshops or personalized feedback.

SELECTION CRITERIA FOR DECISION TREE ALGORITHMS

The optimal decision tree algorithm for a specific blended learning analysis hinges on several factors, including:

1. Classification Goal: If high classification accuracy is paramount, J48 might be a preferred choice.
2. Data Complexity: For datasets with a high number of features, algorithms like CART or Random Forest might be suitable due to their ability to handle complex data structures effectively.
3. Model Interpretability: If understanding the rationale behind the classification is crucial, PART (Partitioning Around Medians) could be preferred due to its rule-based nature.

WEKA's decision tree algorithms unlock valuable insights from blended learning data. Educators can identify students at risk by analyzing quiz scores, forum activity, and logins. This allows for early intervention and targeted support.

Decision trees can also help optimize learning activities by revealing which ones correlate with strong outcomes. Difficulty levels can be tailored based on student performance, and personalized learning paths can be suggested using data and learning style preferences. WEKA empowers educators to translate blended learning data into actionable improvements for all students. This empowers them to design more targeted instruction, personalize learning experiences, and ultimately, enhance student outcomes.

METHOD

This manuscript shares the result of a study, which investigated the efficacy of various decision tree classification algorithms for analyzing blended learning data in WEKA 3.8.6. The primary objective is to identify patterns associated with student performance within the Learning Management System (LMS) of DLSUD, specifically focusing on courses offered by the College of Science and Computer Studies (CSCS).

A. Data and Preprocessing: The data for this study was extracted from the LMS of DLSUD for courses offered by the CSCS. The analysis focused on four key attributes believed to influence student performance:

B. Course Type: This attribute categorizes the course as either theoretical (lecture-based) or practical (hands-on activities).

C. Course Level Outcome (CLO): This attribute represents the overarching learning objectives established for the course.

D. Topic Learning Outcome (TLO): This attribute specifies the learning objectives associated with specific topics within the course.

E. Level of Assessment: This attribute categorizes the assessment type used to evaluate student learning (e.g., quizzes, assignments, final exams). The data was preprocessed to make sure decision tree algorithms could use it before analysis. This might involve handling missing values, converting categorical attributes into numerical representations suitable for WEKA, and potentially scaling numerical attributes if they exhibit significant variations in range.

F. Classification Algorithms and Evaluation: This study employed three prominent decision tree classification algorithms (J48, Random Forest, and PART) within WEKA to conduct a comparative analysis and identify the most efficient model for analyzing blended learning data.

1. J48: This algorithm, an implementation of C4.5, is known for its efficiency and accuracy in handling large datasets. It utilizes a gain ratio criterion to select the optimal splitting attribute at each node, balancing information gain with tree complexity.

2. Random Forest: This ensemble learning technique combines multiple decision trees generated using random subsets of features and data points. This approach often leads to improved performance and robustness compared to single decision trees.

3. Simple Cart: This variant of the CART algorithm focuses on simplicity and interpretability. It employs the Gini impurity measure for attribute selection, aiming to maximize information gain and classification purity.

To assess the performance of these algorithms, a 10-fold cross-validation technique was employed. This technique randomly partitions the data into 10 folds. This process is repeated ten times, ensuring a robust evaluation across the entire dataset. This ensures the chosen algorithm generalizes well on unseen data, leading to reliable insights for educators. These insights can be used to identify at-risk students, optimize learning activities, and personalize learning paths, ultimately improving student outcomes.

G. Performance Metrics: The effectiveness of each decision tree algorithm will be evaluated using relevant performance metrics. Here are some commonly used metrics: Accuracy: The proportion of correctly classified instances.

1. Precision: The ratio of true positives (correctly classified positive instances) to the total number of predicted positive instances.

2. Recall: The ratio of true positives to the total number of actual positive instances.

By comparing these metrics across the three algorithms, we can identify the most suitable approach for uncovering student performance patterns within the blended learning environment of DLSUD's CSCS courses.

RESULTS AND DISCUSSION

Figure 2 shows the result of the 10-fold cross-validation using the J48 classification model. The model achieved an accuracy of 80.4951%, meaning it correctly classified slightly over 80% of the data instances.

```
Number of Leaves  :     16

Size of the tree :      19


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        5105               80.4951 %
Incorrectly Classified Instances      1237               19.5049 %
Kappa statistic                          0.4719
Mean absolute error                      0.2963
Root mean squared error                  0.381
Relative absolute error                 76.9526 %
Root relative squared error             86.8366 %
Total Number of Instances             6342

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.890    0.438    0.853      0.890    0.871      0.474    0.751     0.855     Major
               0.562    0.110    0.643      0.562    0.600      0.474    0.751     0.603     GenEd
Weighted Avg.  0.805    0.352    0.798      0.805    0.801      0.474    0.751     0.789

=== Confusion Matrix ===

    a    b   <-- classified as
 4177  515 |    a = Major
  722  928 |    b = GenEd
```

Figure 2:
Model for J48 Algorithm

The model showed good performance in both precision (79.8%) and recall (80.5%), indicating it can avoid misclassifications and identify most of the relevant cases.

Figure 3 illustrates the outcome of a 10-fold cross-validation using the Random Forest classification model. The model achieved an accuracy of 81.6146%, indicating correct classification for slightly over 80% of the data instances.

```
RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.87 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        5176               81.6146 %
Incorrectly Classified Instances      1166               18.3854 %
Kappa statistic                          0.4955
Mean absolute error                      0.2889
Root mean squared error                  0.3736
Relative absolute error                 75.0322 %
Root relative squared error             85.1449 %
Total Number of Instances             6342

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.905    0.436    0.855      0.905    0.879      0.499    0.802     0.904     Major
               0.564    0.095    0.676      0.564    0.615      0.499    0.802     0.662     GenEd
Weighted Avg.  0.816    0.347    0.808      0.816    0.810      0.499    0.802     0.841

=== Confusion Matrix ===

    a    b   <-- classified as
 4245  447 |    a = Major
  719  931 |    b = GenEd
```

Figure 3:
Model for RandomForest Algorithm

This model demonstrated strong performance in both precision (80.8%) and recall (81.6%), suggesting its ability to minimize misclassifications and effectively identify the most relevant cases.

Figure 4 shows the visualization presents the results of employing a 10-fold cross-validation technique with the SimpleCart classification model. The model attained an accuracy score of 81.7408%, signifying accurate classification for slightly more than 80% of the data instances.

```
Number of Leaf Nodes: 65

Size of the Tree: 129

Time taken to build model: 0.81 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        5184               81.7408 %
Incorrectly Classified Instances      1158               18.2592 %
Kappa statistic                          0.4977
Mean absolute error                      0.2888
Root mean squared error                  0.3735
Relative absolute error                 75.025  %
Root relative squared error             85.1215 %
Total Number of Instances             6342

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.907    0.437    0.855      0.907    0.880      0.501    0.802     0.904     Major
               0.563    0.093    0.680      0.563    0.616      0.501    0.802     0.657     GenEd
Weighted Avg.  0.817    0.348    0.810      0.817    0.811      0.501    0.802     0.840

=== Confusion Matrix ===

    a    b   <-- classified as
 4255  437 |    a = Major
  721  929 |    b = GenEd
```

Figure 4:
Model for SimpleCart Algorithm

https://doi.org/10.24167/celt.v23i2; ISSN: 1412-3320 (print); ISSN: 2502-4914 (online); Accredited; DOAJ

The model in Figure 3 exhibited robust performance in terms of precision (81.0%) and recall (81.7%), indicating its proficiency in reducing misclassifications and accurately identifying most relevant cases.

The three models, ie. J48, Random Forest, and SimpleCart can be compared and classified as shown in Table 1 below:

Table 1:
Comparison of the Three Classification Models

Model	J48		RandomForest		SimpleCart	
Time taken to build the model	0.01 seconds		0.87 seconds		1.68 seconds	
Size of the Tree	19		100		129	
Correctly classified instances	5105	80.50%	5176	81.61%	5184	81.74%
Incorrectly classified instances	1237	19.50%	1166	18.39%	1158	18.26%
Kappa Statistics	0.4719		0.4955		0.4977	

The table informs that a granular examination of the decision tree algorithms employed for student performance analysis in WEKA unveils intriguing performance variations. J48 outperforms in terms of training efficiency; it takes only 0.01 seconds to generate the decision tree model. This underscores its suitability for scenarios demanding rapid model generation. SimpleCart follows at 1.68 seconds, while Random Forest exhibits a longer training duration of 0.87 seconds. These disparities in training times can be attributed to the inherent algorithmic complexities. J48's focus on parsimony translates to faster model construction, whereas Random Forest's ensemble nature, involving the generation of multiple trees, necessitates a longer training period.

Model complexity, as measured by the number of nodes in the decision tree, offers another perspective. Among the models, J48 has the fewest number of nodes (19), which may suggest that its prediction requires less processing. On the other hand, SimpleCart's model, which has 129 nodes, is the most complex and may affect interpretability due to its more complex approach to decision-making. Combining straightforwardness and potential precision, Random Forest's 100-node tree finds the middle ground.

In terms of classification performance, SimpleCart takes the lead with an accuracy of 81.74%. This suggests its effectiveness in accurately distinguishing between different student performance categories within the dataset. Random Forest follows closely with an accuracy of 81.61%, demonstrating competitive performance. At 80.50%, J48 has the least amount of precision, which may be an upside for its simplicity and efficiency as a model.

The Kappa statistic delves deeper than basic accuracy by incorporating the notion of agreement beyond chance. All three models exhibit moderate agreement levels (above 0.4) between their predictions and the actual classifications. SimpleCart maintains its dominance with the highest Kappa statistic (0.4977), followed by Random Forest (0.4955) and J48 (0.4719). This

indicates that SimpleCart's predictions have a stronger concordance with the actual student performance outcomes when accounting for chance agreement.

Particular analysis priority will determine which method is best. If interpretability and rapid model generation are paramount, J48 might be a compelling choice due to its concise model and training speed. However, if maximizing classification accuracy is the primary goal, SimpleCart appears to be the most effective in this scenario. Random Forest presents a compromise between interpretability and accuracy but with a trade-off in training time.

It's crucial to acknowledge that this comparison is specific to the provided dataset, and the performance of these algorithms can fluctuate based on the data characteristics. Additionally, other factors beyond those presented in the table might influence the decision. For instance, if understanding the reasoning behind the model's predictions is critical, a rule-based learner like PART could be explored as an alternative to these decision tree algorithms. By incorporating further analysis techniques and exploring additional algorithms, a more comprehensive understanding of student performance within the blended learning environment can be achieved.

SUMMARY AND CONCLUSION

This study looked at the use of decision tree classification methods to analyze student performance data in the College of Science and Computer Studies (CSCS) blended learning environment at DLSUD. Using WEKA 3.8.6, the study concentrated on four main factors that were thought to affect student performance: course type, topic learning outcome (TLO), course level outcome (CLO), and degree of assessment (Sadiq et al., 2014).

The comparative analysis of J48, Random Forest, and SimpleCart algorithms revealed valuable insights. J48 demonstrated exceptional efficiency in model building, making it ideal for rapid analysis (Quinlan, 1993). However, its efficiency resulted in a less complex model with potentially lower accuracy. Conversely, SimpleCart constructed the most intricate model, achieving the highest classification accuracy but potentially sacrificing interpretability (Karatas & Verhoef, 2018). Random Forest offered a balanced approach, striking a middle ground between interpretability, accuracy, and training time (Breiman, 2001).

The selection of the optimal algorithm hinges on the specific analytical goals within the CSCS blended learning environment. J48 is well-suited for rapid exploration and initial model development due to its speed (Quinlan, 1993). If maximizing classification accuracy for identifying student performance patterns is the primary focus, then SimpleCart appears to be the most effective choice (Karatas & Verhoef, 2018). Random Forest provides a compromise for scenarios where both interpretability and accuracy are important (Breiman, 2001).

While this study provides a foundation for understanding student performance through decision tree algorithms, it represents a starting point. The performance of these algorithms can vary depending on the data, and other factors beyond the scope of this study could be explored for a more comprehensive analysis. Additionally, investigating rule-based learners like PART could offer valuable insights into the decision-making processes within the models (Witten et al., 2016). By incorporating these considerations and expanding the analysis, educators within the DLSUD CSCS can gain a deeper understanding of student performance within the blended

learning environment. This understanding can empower them to tailor pedagogical practices, optimize learning materials, and ultimately enhance student learning outcomes.

RECOMMENDATIONS

In the initial stages of analyzing student performance data, algorithm selection is paramount. For exploratory data analysis (EDA) focused on rapidly generating initial models, J48 reigns supreme due to its exceptional computational efficiency during training (Quinlan, 1993). This prowess makes it ideal for swiftly constructing foundational models and uncovering preliminary patterns within the student performance data.

However, if maximizing classification accuracy is the primary objective, then the SimpleCart algorithm emerges as the most effective choice (Karatas & Verhoef, 2018). SimpleCart meticulously analyzes the data to identify student performance patterns with the highest degree of precision. It's important to note, however, that this pursuit of accuracy can potentially lead to a decrease in model interpretability, making it more challenging to understand the internal logic behind the model's predictions.

Fortunately, a middle ground exists. RandomForest presents a valuable compromise for those who require both interpretability and strong classification accuracy (Breiman, 2001). By leveraging an ensemble of decision trees, it achieves a balance, providing educators with clear insights into the model's reasoning alongside robust performance on the classification task.

As the analysis progresses, data expansion becomes an intriguing next step. Consider incorporating additional attributes beyond course type, course learning objectives (CLOs), assessment levels, and student demographics (Sadiq et al., 2014). Perhaps including factors like learning styles or student engagement metrics could offer a more comprehensive understanding of the variables influencing student performance.

Furthermore, venturing beyond decision trees and exploring the performance of other machine learning algorithms opens new avenues for exploration. Techniques like Support Vector Machines (SVMs) (Cortes & Vapnik, 1995) or Neural Networks (Schmidhuber, 2015) might offer different perspectives on student performance patterns. Additionally, delving into rule-based learning with algorithms like PART could be insightful. While potentially less accurate than decision trees, these algorithms offer a more interpretable view of the decision-making process within the model (Witten et al., 2016).

The pursuit of knowledge in this domain is an ongoing endeavor. Future research directions include exploring the impact of these additional attributes on student performance (Sadiq et al., 2014). Comparing the performance of other machine learning algorithms for comparative analysis would also be a valuable step (Breiman, 2001; Cortes & Vapnik, 1995; Schmidhuber, 2015). Ultimately, the goal is to leverage the insights gained from this study to develop effective early intervention strategies and provide crucial support for struggling students.

REFERENCES

Breiman, L. (2001). Random forests. Machine learning, 45(3), 5–32.

Cortes, C., & Vapnik, V. N. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Ferguson, R. (2012). Learning analytics: Driving innovation in education. EDUCAUSE Review, 47(7), 8–10.

Graham, C. R., Felder, R. M., & Marshak, P. (2013). Blended learning: Research perspectives (Vol. 2). Sterling, VA: Stylus Publishing, LLC.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter, 11(1), 10–18.

Holmes, G., Donkin, A., & Witten, I. H. (1994). WEKA: A knowledge-based environment for machine learning. Proceedings of the 1994 IEEE conference on artificial intelligence applications in manufacturing (pp. 1268–1275). IEEE.

Karatas, İ., & Verhoef, N. (2018). SimpleCart: A new tree induction algorithm. Knowledge and Information Systems, 54(2), 437–466.

Lang, C., Siemens, G., Walker, J., & Koper, R. (2017). Learning analytics in higher education. New Directions for Institutional Research, 2017(177), 107–126.

Quinlan, J. R. (1993). C4.5: Programs for machine learning. Morgan Kaufmann Publishers.

Sadiq, S., Shahi, M., & Tian, Z. (2014). Learning analytics: A comprehensive survey of the literature. Education and Information Technologies, 19(4), 751–783.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85–117.

Siemens, G., & Long, P. (2011). Penetrating the noise: A conversation about learning analytics. EDUCAUSE Review, 46(5), 3–10.

What is Weka ("Waikato environment for knowledge analysis"). IGI Global. (n.d.). https://www.igi-global.com/dictionary/a-triple-bottom-line-approach-based-clustering-study-for-the-sustainable-development-goals-of-the-european-countries/109964

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data mining: Practical machine learning tools and techniques (Vol. 3). Morgan Kaufmann.