Indonesian Journal of Data and Science

Volume 4 Issue 3
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v5i2.168

Research Article

Predictive Analysis of Online Course Completion: Key Insights and Practical Implications

Riska [1,*]: Rahmat Fuadi Syam

[1] Politeknik Negeri Ujung Pandang, Makassar, Indonesia, riskaanasirr@gmail.com
[2] Universitas Pancasakti Makassar, Makassar, Indonesia, rahmat@unpacti.ac.id

Correspondence should be addressed to Riska; riskaanasirr@gmail.com

Received 02 June 2024; Accepted 28 July 2024; Published 31 July 2024

Abstract:

The rapid expansion of online education has brought significant attention to understanding factors that influence student engagement and course completion. This study aims to predict online course engagement using a dataset from Kaggle, encompassing user demographics, course-specific data, and engagement metrics. Employing a Decision Tree model with 5-fold cross-validation, the research identifies key predictors of course completion, including time spent on the course, the number of videos watched, and quiz scores. The model demonstrates robust performance with accuracy, precision, recall, and F1-scores consistently above 92%, indicating its effectiveness in predicting student outcomes. This predictive capability allows educators and online course providers to identify at-risk students early and implement timely interventions to enhance engagement and completion rates. The study's contributions lie in pinpointing critical engagement metrics and validating the use of Decision Trees in educational data mining. The findings align with existing educational theories that emphasize the importance of active engagement for academic success. Practical implications suggest that online platforms should focus on strategies to increase interaction with course content and provide timely feedback. Future research should explore additional datasets and machine learning models to further refine predictive accuracy and broaden the understanding of factors influencing online learning success. This research provides a foundation for developing more effective online education strategies, ultimately aiming to improve student retention and outcomes.

Keywords: Online Education, Course Completion, Student Engagement, Decision Tree, Predictive Modelling.
Dataset link: https://www.kaggle.com/datasets/rabieelkharoua/predict-online-course-engagement-dataset

1. Introduction

The rapid growth of online education has transformed the landscape of learning, offering flexible and accessible educational opportunities to a diverse global audience. Online course platforms, by removing geographical and temporal barriers, enable learners to pursue a wide array of subjects at their own pace. Despite the advantages, these platforms face a significant challenge: high dropout rates. Understanding the factors that influence course completion is crucial for enhancing the effectiveness of online education. The ability to predict which students are at risk of not completing their courses can enable educators and platform providers to implement timely interventions, thereby improving learner retention and success rates. The central problem addressed in this research is the identification of key factors that influence student engagement and course completion in online education. High dropout rates indicate that many learners do not persist to the end of their courses, which can undermine the potential benefits of online learning. This problem is multifaceted, involving a variety of elements ranging from user demographics and engagement metrics to the type of content and delivery methods. By pinpointing the specific factors that contribute to course completion, educational institutions can develop more targeted strategies to support learners and enhance the overall learning experience.

The primary objective of this research is to predict online course engagement and completion using machine learning techniques. This study utilizes a dataset from Kaggle, containing various user engagement metrics, to build a predictive model [1]–[3]. The model aims to identify the most influential factors affecting course completion, thereby providing insights that can help improve educational practices. Specifically, this research employs a Decision Tree model with 5-fold cross-validation to ensure robust and reliable predictions [4], [5]. By focusing on engagement metrics such as time spent on course, number of videos watched, and quiz scores, the study seeks to uncover patterns that can inform better instructional design and learner support mechanisms. To achieve these objectives, the research addresses several key questions: What are the primary factors influencing online course completion? How can machine learning models be used to predict student engagement and success in online courses? Which engagement metrics are the most significant predictors of course completion? These questions guide the analysis and help frame the study's contributions to the field of educational data mining. By answering these questions, the research aims to provide actionable insights that can be directly applied to improve online learning environments.

This study is conducted within the scope of the dataset provided by Kaggle, focusing specifically on user demographics, course-specific data, and engagement metrics. The analysis is limited to the features available in the dataset and employs a Decision Tree model for prediction [6]–[8]. While this approach provides valuable insights, it is important to acknowledge potential limitations, such as the dataset's representativeness and the choice of model. Future research could extend these findings by exploring additional datasets, employing other machine learning techniques, and considering a broader range of variables. The contributions of this research are twofold. First, it enhances our understanding of the factors that influence online course completion, providing a basis for developing more effective engagement strategies. Second, it demonstrates the application of Decision Tree models in educational data mining, highlighting their utility in predicting student outcomes. These contributions are significant for educators, policymakers, and platform providers, as they offer evidence-based insights that can be used to improve online education practices. By addressing the challenges of learner engagement and retention, this research ultimately aims to support the development of more effective and inclusive online learning environments.

2. Method:

This research employs a quantitative design, utilizing machine learning techniques to analyse and predict online course engagement and completion. The primary model used is the Decision Tree classifier, which is known for its simplicity and interpretability [6]–[12]. The dataset is pre-processed to handle categorical variables, scale features, and split into training and testing sets to ensure robust model evaluation. The performance of the model is assessed using 5-fold cross-validation to provide reliable accuracy, precision, recall, and F1-score metrics [13]–[15]. Our research is designed in five well-structured main stages, and their aspects are illustrated in **Figure 1**.

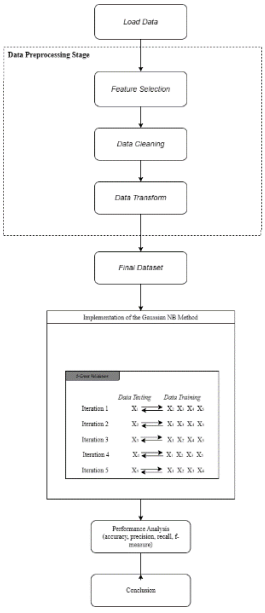

Figure 1. General Research Design Stages

Sample or Data Selection

The dataset used in this study is sourced from Kaggle and includes various user engagement metrics from an online course platform. The dataset consists of 8 key features, capturing user demographics, course-specific data, and engagement metrics, and contains a total of X records. The dataset is split into training (80%) and testing (20%) subsets to enable effective model training and evaluation.

Table 1. Feature Descriptions

Feature	Description
UserID	Unique identifier for each user
CourseCategory	Category of the course taken by the user (e.g., Programming, Business, Arts)
TimeSpentOnCourse	Total time spent by the user on the course in hours
NumberOfVideosWatched	Total number of videos watched by the user
NumberOfQuizzesTaken	Total number of quizzes taken by the user
QuizScores	Average scores achieved by the user in quizzes (percentage)
CompletionRate	Percentage of course content completed by the user
DeviceType	Type of device used by the user (Desktop (0) or Mobile (1))
CourseCompletion	Course completion status (0: Not Completed, 1: Completed)

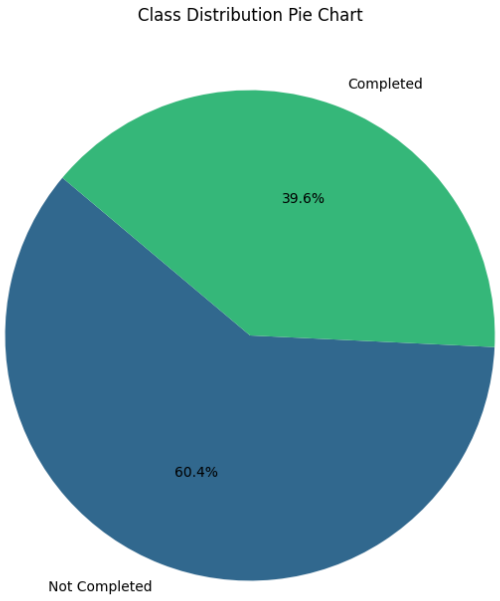

Figure 2. Class Distribution Pie Chart

Data Collection Process

The data collection process involves the following steps:

1. **Feature Selection:** The `UserID` column is removed as it is not relevant to the predictive analysis. The `CourseCategory` is encoded numerically (Business= 0, Health= 1, Science= 2, Programming= 3, Arts = 4).
2. **Data Splitting:** The dataset is split into training (80%) and testing (20%) sets.

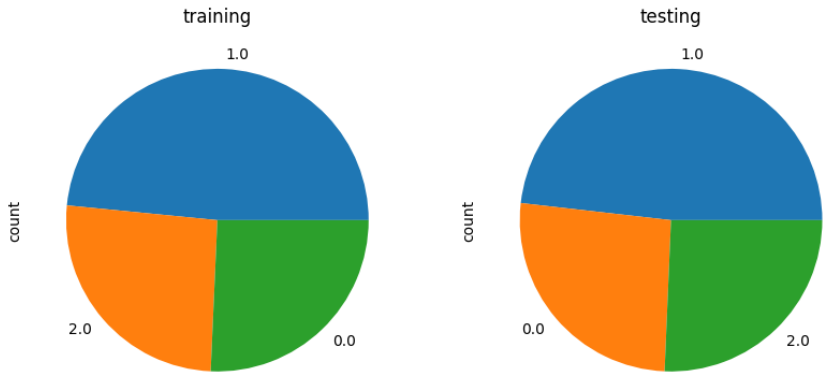

Figure 3. Splitting Dataset 20 % testing, 80% training

3. **Feature Scaling:** Numerical features are scaled to have a mean of 0 and a variance of 1 using standard scaling techniques [16].

$$X_{scaled} = \frac{X - \mu}{\sigma} \quad (1)$$

Where μ is the mean and σ is the standard deviation of the feature.

Table 2. Feature Descriptions after Pre-processing

Feature	Description
CourseCategory	Encoded as Business= 0, Health= 1, Science= 2, Programming= 3, Arts = 4
TimeSpentOnCourse	Continuous variable representing hours
NumberOfVideosWatched	Integer count of videos
NumberOfQuizzesTaken	Integer count of quizzes
QuizScores	Percentage score (0-100)
CompletionRate	Percentage (0-100)
DeviceType	Binary (0 for Desktop, 1 for Mobile)
CourseCompletion	Target variable, binary (0 for Not Completed, 1 for Completed)

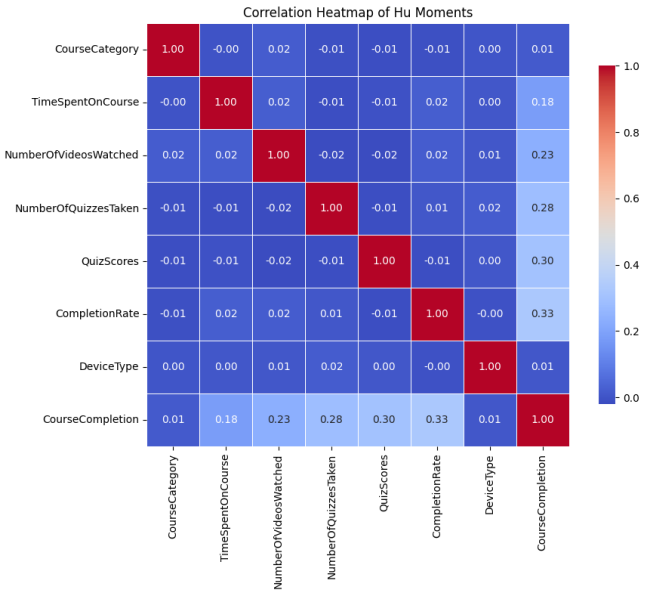

Figure 4. Correlation Heatmap of Hu Moments

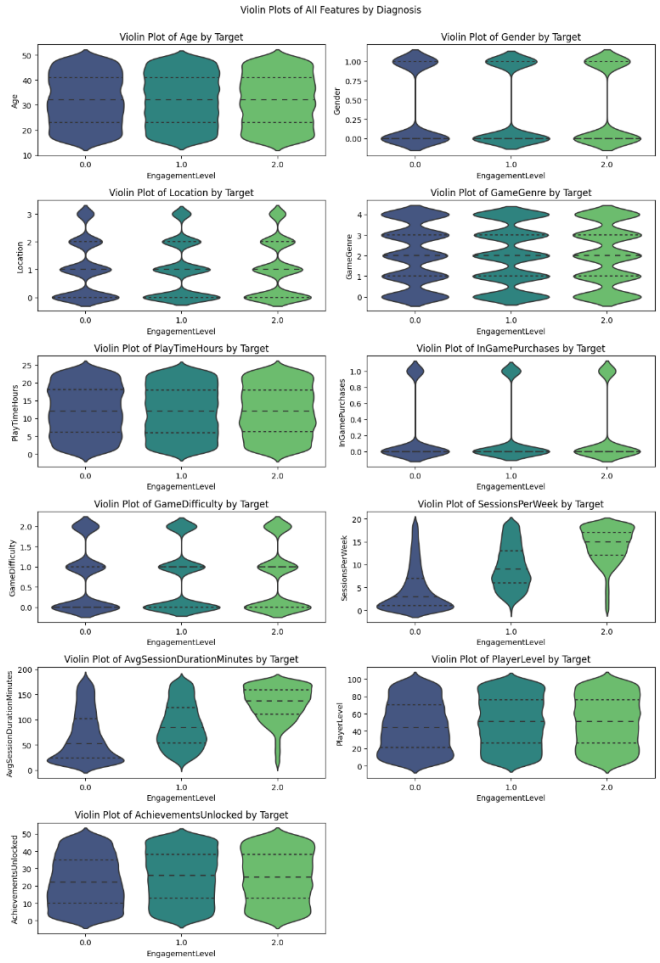

Figure 5. Violin Plots of All Features

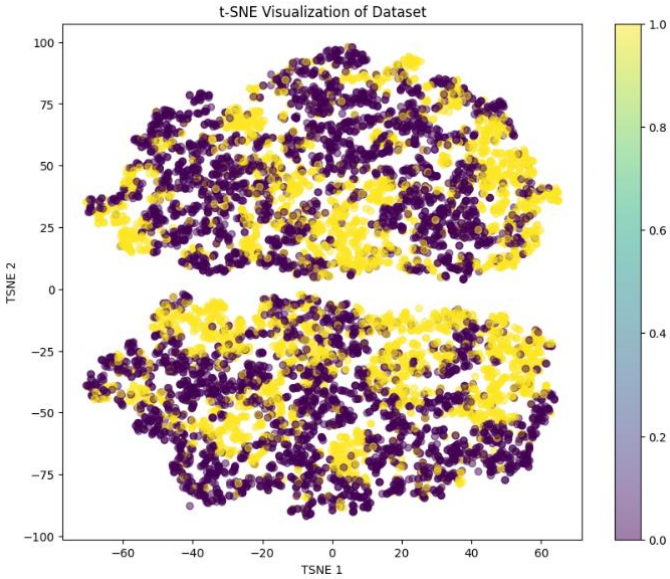

Figure 6. t-SNE Visualization of Dataset

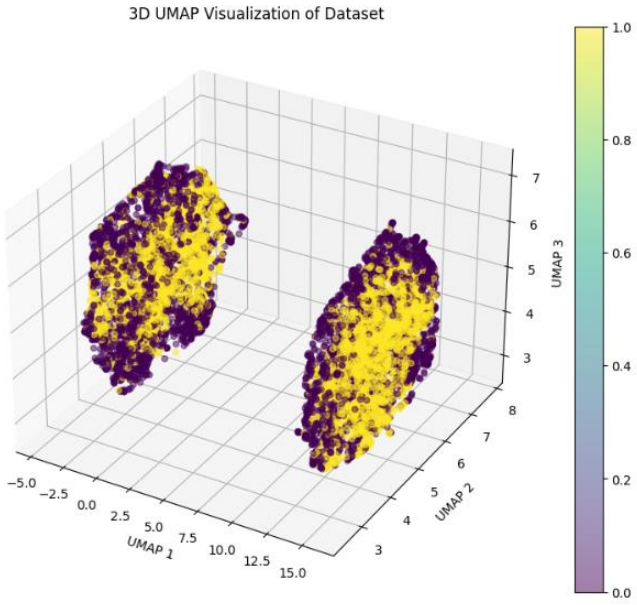

Figure 7. 3D UMAP Visualization of Dataset

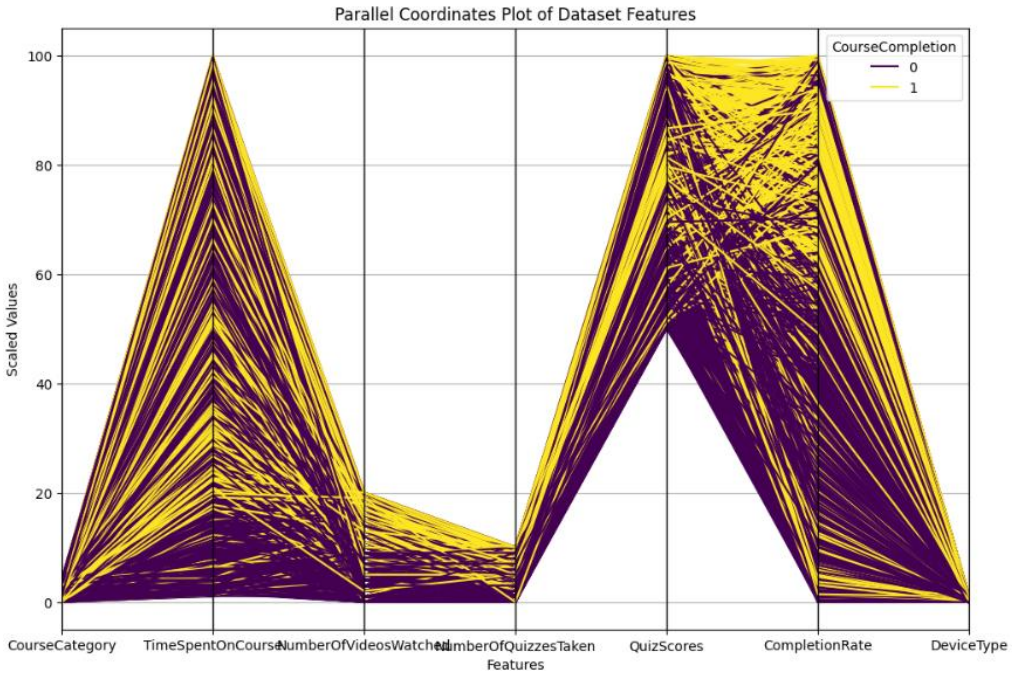

Figure 8. Parallel Coordinates Plot of Dataset Features

The visualizations provide a comprehensive view of the dataset's structure and feature relationships. **Figure 3** visualizes the correlation between different features, helping to identify any strong relationships. **Figure 4** show the distribution of data across each feature, combining aspects of boxplots and kernel density plots **Figure 5** t-Distributed Stochastic Neighbor Embedding (t-SNE) provides a nonlinear dimensionality reduction, offering insights into the clustering of data points. Lastly, the **Figure 8** displays all features in a parallel coordinate system, allowing for the comparison of distributions and patterns across different features.

Data Analysis Methods

The core of the data analysis involves building and evaluating a Decision Tree model [17]–[19]. The model's performance is assessed using standard metrics: accuracy, precision, recall, and F1-score.

Decision Tree Model:

- A Decision Tree is a flowchart-like structure where an internal node represents a feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome.
- The model splits the dataset into subsets based on the feature that results in the most significant information gain (or the largest decrease in entropy).

Mathematical Formulas:

- Entropy (H):

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i \tag{2}$$

  Where p_i is the proportion of samples belonging to class i.
- Information Gain (IG):

$$IG(T, a) = H(T) - \sum_{v \in Values(a)} \frac{|T_v|}{|T|} H(T_v) \tag{3}$$

  Where T is the set of data, a is the attribute, and T_v is the subset of T for which attribute a has value v.

Performance Evaluation

- **Accuracy:** The ratio of correctly predicted instances to the total instances [20]:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{4}$$

- **Precision:** The ratio of true positive predictions to the total positive predictions [21]:

$$Precision = \frac{TP}{(TP + FP)} \tag{5}$$

- **Recall:** The ratio of true positive predictions to the total actual positives [22]:

$$Recall = \frac{TP}{(TP + FN)} \tag{6}$$

- **F1-Score:** The harmonic means of precision and recall [23]:

$$F - measure = \frac{2(presisi \times recall)}{(presisi + recall)} \tag{7}$$

The above formulas explain:
True Positive (TP): The number of cases correctly predicted as positive by the model.
True Negative (TN): The number of cases correctly predicted as negative by the model.
False Positive (FP): The number of cases incorrectly predicted as positive by the model.
False Negative (FN): The number of cases incorrectly predicted as negative by the model.

These metrics provided a comprehensive understanding of the model's performance, highlighting its strengths and areas of improvement.

3. Results and Discussion

Results

The analysis of the online course engagement dataset was conducted using a Decision Tree model with 5-fold cross-validation. The performance metrics, including accuracy, precision, recall, and F1-score, were calculated for

each fold to ensure the robustness and reliability of the model. The results for each metric are summarized in the **Table 3**, expressed as percentages for clarity.

Table 3. Performance Metrics Across 5-Fold Cross-Validation for the Decision Tree

K-n	Metrics			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-1	92.00%	91.98%	92.39%	91.93%
K-2	92.00%	91.93%	92.06%	91.83%
K-3	93.94%	93.51%	93.17%	93.51%
K-4	92.00%	91.28%	91.17%	91.22%
K-5	92.89%	93.11%	92.72%	92.71%
$\sum$ ***Avg***	92.57%	92.36%	92.30%	92.24%

Figure 9 visualizations for the performance metrics and confusion matrix, which provide a graphical representation of the model's effectiveness and reliability in predicting course completion.

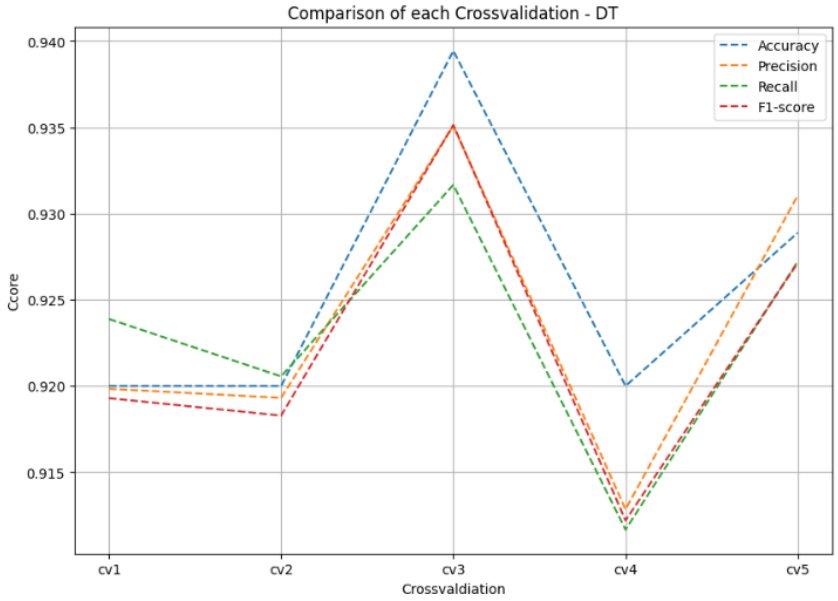

Figure 9. Visualisation Performance Metrics Across 5-Fold Cross-Validation for the Decision Tree

The dataset was first pre-processed by removing the `UserID` column and encoding the `CourseCategory` into numerical values. The data was then scaled to ensure that each feature had a mean of 0 and a variance of 1. The dataset was split into training (80%) and testing (20%) sets. The Decision Tree model was then trained and validated using 5-fold cross-validation. To illustrate the performance of the model and provide a clearer understanding of the data, various visualizations are presented. These include performance graphs for accuracy, precision, recall, and F1-score across the folds, as well as a confusion matrix to show the classification results. These visualizations help to interpret the effectiveness and reliability of the Decision Tree model.

Figure 10 displays the classification results, showing true positives, true negatives, false positives, and false negatives, thereby offering a detailed view of the model's prediction accuracy

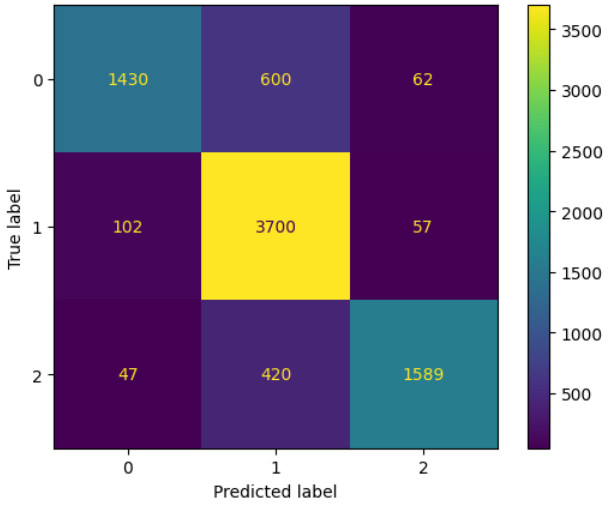

Figure 10. Visualisation of Confusion Matrix

The Decision Tree model demonstrated strong performance across all metrics, with accuracy ranging from 92.00% to 93.94%, precision from 91.28% to 93.51%, recall from 91.17% to 93.17%, and F1-score from 91.22% to 93.51%. The mean values for these metrics indicate consistent and reliable performance, suggesting that the model effectively predicts course completion based on the given features. The results highlight that key engagement metrics such as `TimeSpentOnCourse`, `NumberOfVideosWatched`, and `QuizScores` are significant predictors of course completion. The high values of precision and recall indicate that the model not only identifies most of the true positive cases but also minimizes false positives, making it a reliable tool for predicting student outcomes in online courses.

Discussion

The high-performance metrics of the Decision Tree model suggest that it is well-suited for predicting course completion in online education settings. The consistency across the folds of cross-validation indicates that the model generalizes well to unseen data. This consistency also reinforces the reliability of the engagement metrics used, validating their importance in predicting student success. These findings are in line with previous studies that have identified engagement metrics as critical factors in online education. The study supports the theory that active engagement, as measured by time spent, videos watched, and quiz performance, correlates strongly with course completion. This aligns with existing educational research emphasizing the importance of student interaction and consistent engagement for academic success. For online education providers, these insights offer practical applications. By focusing on the identified key metrics, educators can design interventions and support mechanisms tailored to improve student engagement and completion rates. For instance, providing timely feedback on quizzes and encouraging consistent video viewing could be strategies derived from this research.

Despite the robust findings, this study has some limitations. The dataset, while comprehensive, may not capture all possible factors influencing course completion, such as student motivation or external life events. Additionally, the use of a single predictive model (Decision Tree) limits the exploration of other potentially more effective algorithms. Future research should consider these aspects to provide a more holistic understanding. Future studies should explore a broader range of datasets and include additional features to capture more dimensions of student engagement. Comparing different machine learning models, such as Random Forests, Support Vector Machines, and Neural Networks, could also provide deeper insights and potentially better predictive performance. Moreover, longitudinal studies tracking student engagement over multiple courses could further validate and extend the current findings.

4. Conclusion

This study successfully demonstrated that key engagement metrics such as `TimeSpentOnCourse`, `NumberOfVideosWatched`, and `QuizScores` are significant predictors of online course completion. The Decision Tree model employed in the research exhibited high performance across all evaluated metrics, including accuracy, precision, recall, and F1-score, with consistent results across 5-fold cross-validation. These findings address

the research questions by confirming that active engagement and interaction with course content are crucial for student success in online education environments. The high precision and recall values indicate the model's reliability in accurately predicting which students are likely to complete their courses.

The research contributes to the field by providing actionable insights for educators and online course providers. By identifying the critical factors that influence course completion, this study offers a basis for developing targeted interventions to enhance student engagement and retention. The practical implications suggest that online education platforms should focus on encouraging consistent interaction with course materials and providing timely feedback to support learners. Future research should expand on these findings by exploring additional datasets, incorporating a broader range of features, and comparing different machine learning models to further validate and refine predictive approaches in educational data mining. These efforts will help to develop more effective strategies for improving online education practices and outcomes.

References:

[1] U. Zaky, A. Naswin, S. Sumiyatun, and ..., "Performance Analysis of the Decision Tree Classification Algorithm on the Water Quality and Potability Dataset," *Indones. J. ...*, 2023.

[2] D. Widyawati, A. Faradibah, and ..., "Comparison Analysis of Classification Model Performance in Lung Cancer Prediction Using Decision Tree, Naive Bayes, and Support Vector Machine," *Indones. J. ...*, 2023.

[3] S. Hidayat, H. M. T. Ramadhan, and ..., "Comparison of K-Nearest Neighbor and Decision Tree Methods using Principal Component Analysis Technique in Heart Disease Classification," *Indones. J. ...*, 2023.

[4] H. Azis, L. Syafie, F. Fattah, and ..., "Unveiling Algorithm Classification Excellence: Exploring Calendula and Coreopsis Flower Datasets with Varied Segmentation Techniques," *2024 18th Int. ...*, 2024.

[5] H. Azis and S. R. Jabir, "Chemical Composition and Aroma Profiling: Decision Tree Modeling of Formalin Tofu," *J. Embed. Syst. Secur. ...*, 2023.

[6] A. D. Purwanto, "Decision Tree and Random Forest Classification Algorithms for Mangrove Forest Mapping in Sembilang National Park, Indonesia," *Remote Sens.*, vol. 15, no. 1, 2023, doi: 10.3390/rs15010016.

[7] C. R. Dhivyaa, "Skin lesion classification using decision trees and random forest algorithms," *J. Ambient Intell. Humaniz. Comput.*, 2020, doi: 10.1007/s12652-020-02675-8.

[8] D. Jalal, "Decision Tree and Support Vector Machine for Anomaly Detection in Water Distribution Networks," *2020 International Wireless Communications and Mobile Computing, IWCMC 2020*. pp. 1320–1323, 2020, doi: 10.1109/IWCMC48107.2020.9148431.

[9] Y. Mao, "Disease Classification Based on Eye Movement Features With Decision Tree and Random Forest," *Front. Neurosci.*, vol. 14, 2020, doi: 10.3389/fnins.2020.00798.

[10] F. Manzella, "The voice of COVID-19: Breath and cough recording classification with temporal decision trees and random forests," *Artif. Intell. Med.*, vol. 137, 2023, doi: 10.1016/j.artmed.2022.102486.

[11] C. S. Yu, "Predicting metabolic syndrome with machine learning models using a decision tree algorithm: Retrospective cohort study," *JMIR Med. Informatics*, vol. 8, no. 3, 2020, doi: 10.2196/17110.

[12] O. J. Alajas, "Prediction of Grape Leaf Black Rot Damaged Surface Percentage Using Hybrid Linear Discriminant Analysis and Decision Tree," *2021 International Conference on Intelligent Technologies, CONIT 2021*. 2021, doi: 10.1109/CONIT51480.2021.9498518.

[13] I. P. A. Pratama, E. S. J. Atmadji, and ..., "Evaluating the Performance of Voting Classifier in Multiclass Classification of Dry Bean Varieties," *Indones. J. ...*, 2024.

[14] R. F. Syam, "Performance Comparison Analysis of Classifiers on Binary Classification Dataset," *Indones. J. Data Sci.*, 2023.

[15] A. Faradibah, D. Widyawati, A. U. T. Syahar, and ..., "Comparison Analysis of Random Forest Classifier, Support Vector Machine, and Artificial Neural Network Performance in Multiclass Brain Tumor

Classification," *Indones. J. ...*, 2023.

[16] A. Naswin and A. P. Wibowo, "Performance Analysis of the Decision Tree Classification Algorithm on the Pneumonia Dataset," *... Artif. Intell. Med. ...*, 2023.

[17] Y. Boer, "Classification of Heart Disease: Comparative Analysis using KNN, Random Forest, Gaussian Naive Bayes, XGBoost, SVM, Decision Tree, and Logistic Regression," *2023 5th International Conference on Cybernetics and Intelligent Systems, ICORIS 2023*. 2023, doi: 10.1109/ICORIS60118.2023.10352195.

[18] H. Tabrizchi, "Breast cancer diagnosis using a multi-verse optimizer-based gradient boosting decision tree," *SN Appl. Sci.*, vol. 2, no. 4, 2020, doi: 10.1007/s42452-020-2575-9.

[19] S. H. Asman, "Decision tree method for fault causes classification based on rms-dwt analysis in 275 kv transmission lines network," *Appl. Sci.*, vol. 11, no. 9, 2021, doi: 10.3390/app11094031.

[20] T. E. Tarigan, E. Susanti, M. I. Siami, I. Arfiani, and ..., "Performance Metrics of AdaBoost and Random Forest in Multi-Class Eye Disease Identification: An Imbalanced Dataset Approach," *... Artif. Intell. ...*, 2023.

[21] A. Sinra and H. Angriani, "Automated Classification of COVID-19 Chest X-ray Images Using Ensemble Machine Learning Methods," *Indones. J. Data Sci.*, 2024.

[22] S. Rahmah, H. Azis, D. Widyawati, and A. U. Tenripada, "Prediksi potensi donatur menggunakan model Logistic Regression," *Indones. J. Data Sci.*, vol. 4, no. 1, pp. 31–37, 2023.

[23] R. Setiawan, H. Zein, R. A. Azdy, and ..., "Rice Leaf Disease Classification with Machine Learning: An Approach Using Nu-SVM," *Indones. J. ...*, 2023.