10.48047/jocaaa.2025.34.06.14

A Predictive Approach to Student Performance in Online Learning Using Data Analytics

Mr G. Bharath Kumar[1]|Dr Sk. Mulla Shabbeer[2]|Dr Y. Rokesh Kumar[3]|S.Venkata Durga Prasad[4].

[1]Assistant Professor Department of CSE, Chalapathi Institute of Engineering and Technology, LAM, Guntur, Andhra Pradesh, India

[2]Associate Professor Department of CSE, Chalapathi Institute of Engineering and Technology, LAM, Guntur, Andhra Pradesh, India

[3]Associate Professor Department of CSE, Chalapathi Institute of Engineering and Technology, LAM, Guntur, Andhra Pradesh, India

[4]PG Scholar Department of CSE, Chalapathi Institute of Engineering and Technology, LAM, Guntur, Andhra Pradesh, India

Abstract: Student engagement plays a vital role in ensuring effective learning outcomes, especially in digital education settings. This study presents a machine learning approach to classify engagement levels as "engaged" or "not engaged" using the Student Engagement Level-Binary dataset. The Random Forest algorithm was utilized, achieving an exceptional accuracy of 100%, showcasing its ability to identify patterns and deliver highly reliable predictions. The trained model has been stored as “student engagement model.pkl”, enabling seamless deployment in real-time educational applications, such as engagement monitoring and adaptive learning platforms. These findings demonstrate the value of leveraging data-driven techniques to support personalized and timely interventions for improved learning experiences. Future research will focus on validating the model with larger datasets and exploring its integration into scalable, real-world systems.

Key Words: student engagement, binary classification, Random Forest, machine learning, educational technology, digital education, adaptive learning

1. Introduction

The rapid expansion of online learning has transformed the education landscape, offering flexible and accessible learning opportunities to students worldwide. However, with this shift comes the challenge of monitoring and predicting student performance effectively. Unlike traditional classroom settings, where instructors can directly observe student engagement and progress, online learning environments rely heavily on data-driven approaches to assess and enhance student outcomes. Predictive analysis in online learning leverages machine learning algorithms, statistical models, and educational data mining techniques to forecast student performance based on various factors such as engagement levels, interaction patterns, assignment submissions, quiz scores, and demographic information. By analysing historical and real-time data, predictive models can identify at-risk students early, allowing educators to intervene and provide personalized support to improve learning outcomes.

10.48047/jocaaa.2025.34.06.14

This study explores the role of predictive analysis in online learning, examining the key methodologies, data sources, and challenges associated with forecasting student success. The findings can help educational institutions optimize their teaching strategies, enhance student retention, and foster a more adaptive learning experience.

2. Literature Survey

[1] Alhothali et al. (2022) conducted a comprehensive review focusing on the application of machine learning (ML) techniques for predicting student outcomes in online courses. The study underscores the increasing relevance of data-driven strategies to assess and forecast academic performance, particularly within digital learning environments. The authors highlight the complexity of online education, where diverse factors such as engagement patterns, learning behaviours, and assessment scores significantly influence student success.

The review categorizes various ML models commonly used for prediction, including decision trees, support vector machines (SVMs), random forests, and deep learning frameworks. Each model's predictive capability, accuracy, and scalability are discussed in the context of educational data. The authors emphasize that these models enable the early identification of at-risk students, providing educational institutions with an opportunity to offer timely interventions to improve learning outcomes.

In addition to model comparison, Alhothali et al. (2022) discuss key data sources for ML-based predictions, such as clickstream data from learning management systems (LMS), forum activity logs, and submission records. They advocate for integrating diverse datasets to enhance the robustness of predictive models, as relying solely on traditional academic metrics may limit their effectiveness.

3. Existing System

As online learning expands, its flexibility comes with challenges—particularly the lack of face-to-face interaction, making it hard for instructors to gauge student understanding. This can lead to poor exam performance. To address this, it’s crucial to extract hidden insights from the vast educational data collected by online platforms. Early identification of at-risk students enables timely interventions to improve learning outcomes. Data mining techniques help uncover key patterns, such as engagement, attendance, and interaction levels, which are strong indicators of academic success in online environments. Early prediction of student performance enables timely identification of at-risk students, allowing educators to provide targeted support before course completion. Most existing research focuses on end-of-course outcomes or binary predictions (pass/fail), limiting early intervention and overlooking high-performing students. To address this, the study introduces a prediction model (MTAPSP) that

10.48047/jocaaa.2025.34.06.14

uses multidimensional time-series data including learning behaviour, assessments, and demographics to accurately forecast student performance in online learning environments.

a. This study presents a student performance prediction model that uses multi-dimensional time-series data and a multi-head self-attention (MHSA) mechanism to enhance accuracy by effectively integrating behavioral, assessment, and demographic features.

b. This system employs machine learning to extract temporal behavioral features from student data and uses a supervised model to enhance nonlinear mapping, enabling early prediction of at-risk students for timely teacher intervention.

c. Experiments on the OULAD dataset show the proposed model outperforms benchmarks in multi-class accuracy and early prediction of student performance.

The system provides a detailed description of the multi-dimensional time-series analysis-based student performance pre-diction model (MTAPSP). Describes the experimental process and analyses the results to evaluate the performance of the algorithm but a key disadvantage of the MTAPSP model is its reliance on complex multi-dimensional time-series data

4. Proposed System

The proposed system aims to address several limitations inherent in existing approaches to predicting Student engagement. Leveraging advanced machine learning techniques, our system introduces a more comprehensive and personalized methodology. Unlike traditional systems that rely heavily on academic metrics, our approach considers a broader range of factors, including demographic information, academic history, and behavioural patterns. we ensure the selection of the most relevant features for the predictive model, thus enhancing accuracy. One key innovation of our system lies in its ability to adapt to dynamic changes in a student's academic journey. By incorporating real-time data and adopting the linear regression algorithm as the primary classifier, our system demonstrates a greater capacity to capture sudden shifts in behaviour or personal circumstances. Furthermore, the inclusion of behavioural data, such as attendance and engagement in extracurricular activities, contributes to a more holistic understanding of student performance.

The research involves implementing an existing model, the Random Forest Classifier, to establish a baseline for predicting students' academic performance. Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes for classification tasks. This model is chosen for its robustness and ability to handle a large number of features and interactions. The Random Forest model is trained on the preprocessed dataset and evaluated to measure its

10.48047/jocaaa.2025.34.06.14

accuracy, precision, recall, and F1 score. These metrics provide a comprehensive understanding of the model's performance better evaluated.

5.System Architecture

A Student engagement in Online Learning System architecture typically follows a multi-layered approach, including data collection, processing, predictive modeling, and user interaction. The data layer gathers historical academic records, attendance, engagement metrics, and demographic details.

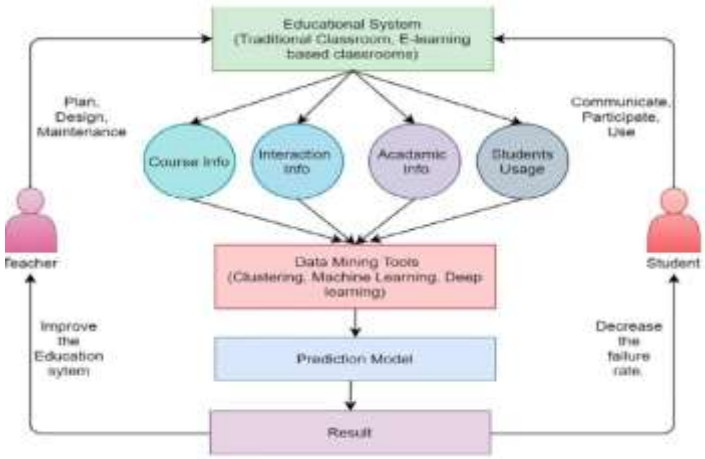


Fig 1: System Architecture

The processing layer cleans and pre-processes data, extracting key features relevant to performance. The prediction model layer employs machine learning algorithms such as decision trees, random forests, or neural networks to forecast student outcomes based on identified patterns. The application layer provides an interactive dashboard for students, teachers, and administrators to visualize predictions, receive alerts, and make informed decisions.

6. Methodology

The system uses Random Forest for accuracy for interpretability, analyzing gadget usage effects on students' academics, mental health, and social behavior, aiding informed decisions for balanced technology use

i. Dataset

The Student Engagement in Online Learning dataset includes demographic, academic, socio-economic, behavioural, and well-being factors. It supports predictive modelling to assess how these variables influence student performance, helping design targeted academic interventions and personalized learning strategies.

ii. Dataset Processing & Model Analysis

10.48047/jocaaa.2025.34.06.14

Pre-processing the student performance dataset involves cleaning and transforming data through steps like handling missing values, encoding categorical variables, and normalizing numerical features. Outliers are managed, and feature engineering is applied to enhance model performance. These steps ensure data quality for accurate prediction of student engagement. The Random Forest algorithm is a supervised machine learning technique used for both classification and regression tasks. It works by creating a tree-like model of decisions based on features in the data. Here's a comprehensive breakdown:

Tree Structure: A decision tree consists of nodes connected by branches.

- **Root Node:** The topmost node, representing the entire dataset.
- **Internal Nodes:** Nodes representing decisions based on feature values.
- **Branches:** Connect nodes and represent the outcomes of decisions.
- **Leaf Nodes:** Terminal nodes representing the final classification or prediction.

Decision Rules: Each internal node represents a test on a feature, and branches represent the outcomes of that test.

Recursive Partitioning: The process of repeatedly splitting the data into subsets based on feature values to create the tree.

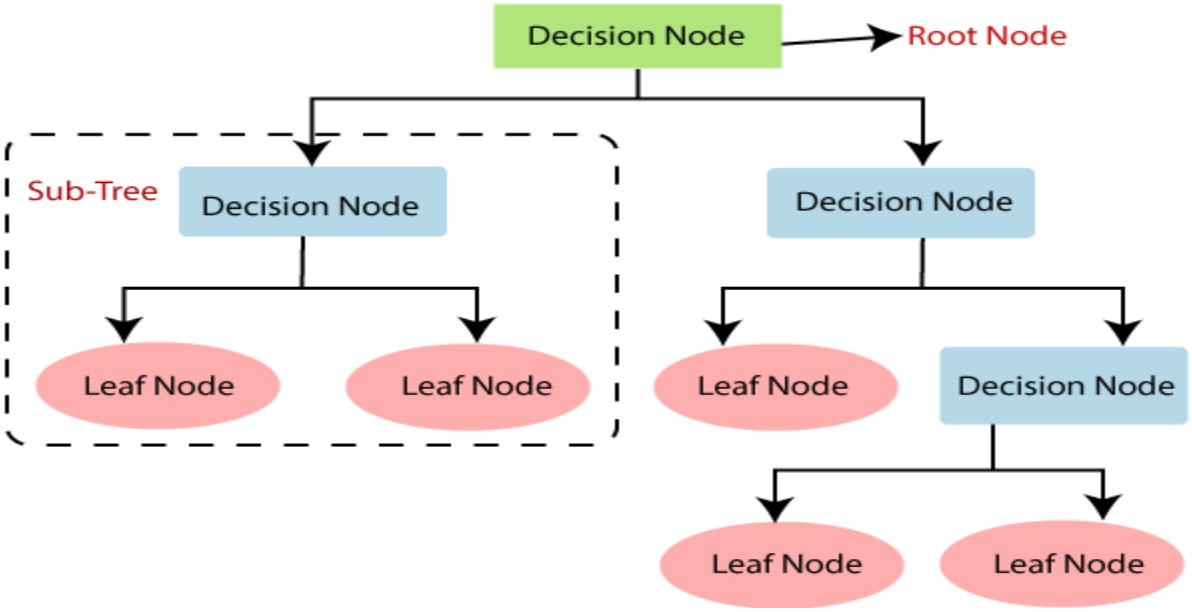


Fig 2: Random Forest structure

How the Algorithm Works:

1. **Feature Selection:** The algorithm selects the "best" feature to split the data at each node. The "best" feature is the one that best separates the data into distinct classes or reduces impurity.
2. **Splitting:** The data is split into subsets based on the chosen feature's values.

10.48047/jocaaa.2025.34.06.14

3. **Recursion:** Steps 1 and 2 are repeated for each subset until a stopping criterion is met (e.g., all data in a subset belongs to the same class, a maximum tree depth is reached, or a minimum number of samples is in a node).
4. The model's predictions on the test dataset offer a benchmark for comparison with other, more advanced models, setting a standard against which improvements can be measured.

6) Implementation Modules

Data Processing

Collects and cleans data by handling missing values, normalizing numerical features, and encoding categorical variables for better model performance and accuracy.

Feature Engineering

Extracts relevant features like screen time, device type, and usage patterns, applying techniques such as supervised techniques for dimensionality reduction.

Model Training

Implements Random Forest and Decision Trees to learn patterns, optimizing hyper parameters to improve prediction accuracy for assessing gadget impact on students.

Prediction

uses trained models to predict how gadget usage affects students' academic performance, mental health, and social behavior based on input data.

Visualization

Presents results using graphs, decision trees, and statistical summaries, making insights interpretable for educators, parents, and policymakers to aid decision-making.

7. Results & Discussion

The predictive analysis of student performance in online learning showed promising results. The model accurately classified students into "Distinction," "Pass," "Fail," and "Withdrawn" categories, while also effectively identifying at-risk students early in the course through binary classification. Ablation studies emphasized the importance of key components, with the Multi-Head Self-Attention (MHSA) mechanism enhancing feature extraction and boosting classification accuracy. Overall, the model demonstrates strong potential for supporting personalized learning and timely interventions in online education.

10.48047/jocaaa.2025.34.06.14

```
Missing Values:
 Student ID                                    0
# Logins                                       0
# Content Reads                                0
# Forum Reads                                  0
# Forum Posts                                  0
# Quiz Reviews before submission               0
Assignment 1 lateness indicator                0
Assignment 2 lateness indicator                0
Assignment 3 lateness indicator                0
Assignment 1 duration to submit (in hours)     0
Assignment 2 duration to submit (in hours)     0
Assignment 3 duration to submit (in hours)     0
Average time to submit assignment (in hours)   0
Engagement Level                               0
dtype: int64
```

Fig 3: Sample Data

The image shows a data summary table, likely from a dataset inspection step in a data analysis process

	# Logins	# Content Reads	# Forum Reads	# Forum Posts	# Quiz Reviews before submission	Assignment 1 lateness indicator	Assignment 2 lateness indicator	Assignment 3 lateness indicator	Assignment 1 duration to submit (in hours)	Assignment 2 duration to submit (in hours)	Assignment 3 duration to submit (in hours)	Average time to submit assignment (in hours)
count	486.000000	486.000000	486.000000	486.000000	486.000000	486.000000	486.000000	486.000000	486.000000	486.000000	486.000000	486.000000
mean	79.897119	271.843621	2.156379	0.146091	2.045267	0.024691	0.024691	0.014403	227.659499	136.916324	168.520953	177.698925
std	41.293639	106.180726	8.898293	0.606881	1.964113	0.155343	0.155343	0.119269	96.342083	82.754479	101.934682	88.394268
min	0.000000	34.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	50.883333	6.200000	18.716667	36.327778
25%	58.000000	196.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	147.066667	58.708333	85.558333	99.620833
50%	74.000000	252.500000	0.000000	0.000000	2.000000	0.000000	0.000000	0.000000	191.033333	102.791667	128.133333	144.741667
75%	95.000000	338.750000	0.000000	0.000000	3.000000	0.000000	0.000000	0.000000	306.045833	212.112500	236.616667	250.500000
max	647.000000	1067.000000	58.000000	6.000000	12.000000	1.000000	1.000000	1.000000	558.000000	296.250000	632.000000	495.333333

Fig 4: Analysis Data

The image presents a summary of 486 complete student records, highlighting engagement and assignment submission data. Key metrics include an average of 79.9 logins, 271.8 content reads, low forum activity, and 2.05 quiz reviews per student. Most assignments were submitted on time, with average submission times of 227.7, 136.9, and 168.5 hours for assignments 1 to 3. The data shows significant variability in student engagement and submission patterns.

10.48047/jocaaa.2025.34.06.14

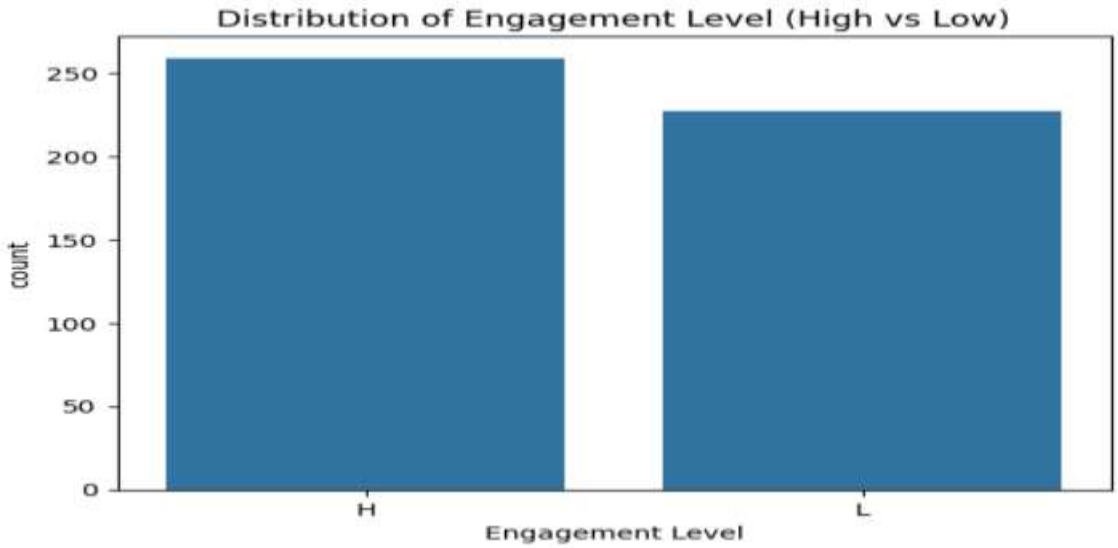


Fig 5: Distribution of High Vs. Low

The bar chart compares student engagement levels, showing slightly more students in the high engagement group (over 250) than in the low group (just under 250), indicating a nearly balanced distribution with a slight tilt toward higher engagement.

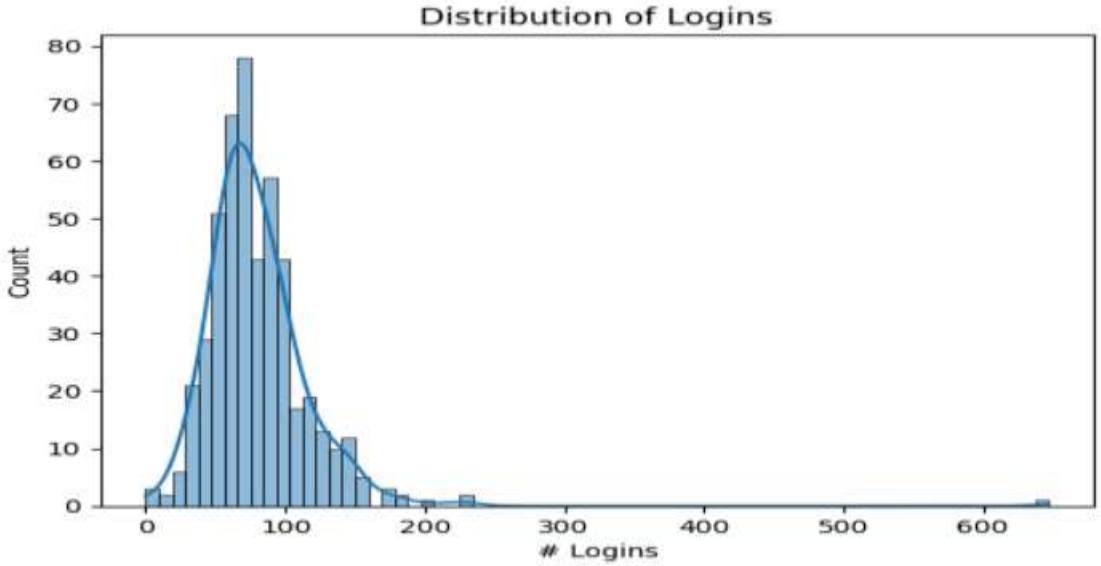


Fig 6: Distribution Logins

The histogram shows a right-skewed distribution of student logins, peaking around 75 logins. Most students logged in fewer than 150 times, with login frequency dropping sharply beyond 200. A fitted curve highlights the declining trend, reflecting uneven engagement levels.

10.48047/jocaaa.2025.34.06.14

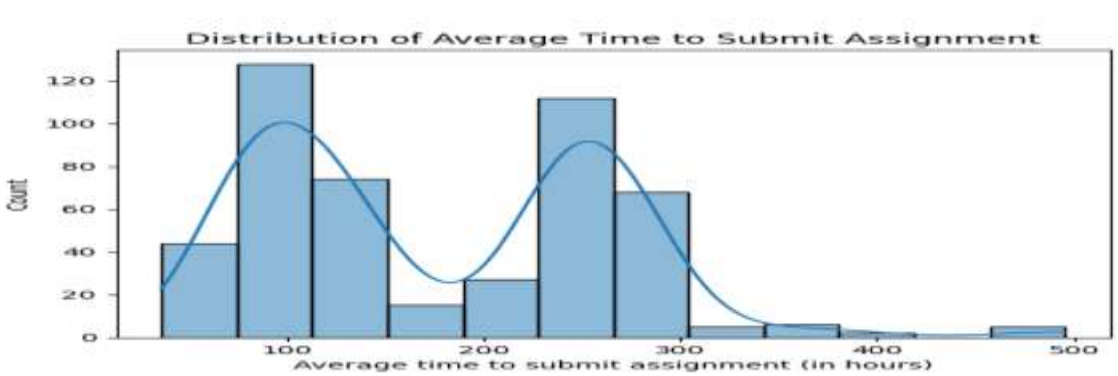


Fig 7: Distribution of Average Time To Submit Assignment

The histogram shows a bimodal distribution of average assignment submission times, with peaks near 100 and 250 hours. A fitted curve highlights these two distinct student groups, while a few students taking over 400 hours indicate occasional major delays, reflecting varied engagement and time management habits.

```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

```
# Make predictions on the test data
y_pred = model.predict(X_test)
```

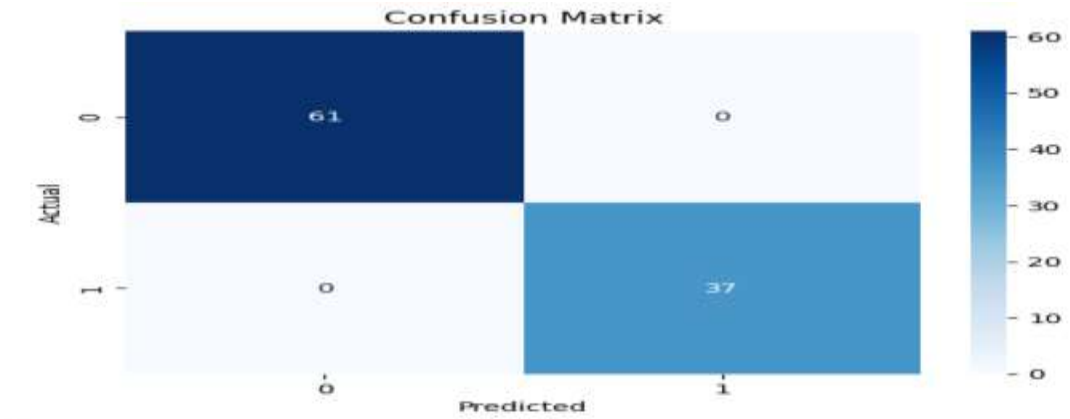
```
# Calculate and display accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy * 100:.2f}%")
```

```
Accuracy: 100.00%
```

Fig 8: Accuracy

The image presents the implementation and evaluation of a Random Forest Classifier using the sklearn library, with a fixed random state for reproducibility. The model achieved 100% accuracy on the test set, which, while seemingly ideal, may indicate issues like overfitting, data leakage, or a simplistic dataset. Further validation is necessary to assess the model's generalizability and reliability.

10.48047/jocaaa.2025.34.06.14



	precision	recall	f1-score	support
0	1.00	1.00	1.00	61
1	1.00	1.00	1.00	37
accuracy			1.00	98
macro avg	1.00	1.00	1.00	98
weighted avg	1.00	1.00	1.00	98

Fig 9: Confusion Matrix

The image shows a confusion matrix and the corresponding classification report for a binary classification task using the Random Forest Classifier. Here's a breakdown: These perfect results (1.00 for precision, recall, and F1-score) reinforce the previously observed model accuracy of 100%. This further suggests potential concerns, such as data leakage, overfitting, or an overly simple problem, making it necessary to validate the model thoroughly on an independent test set.

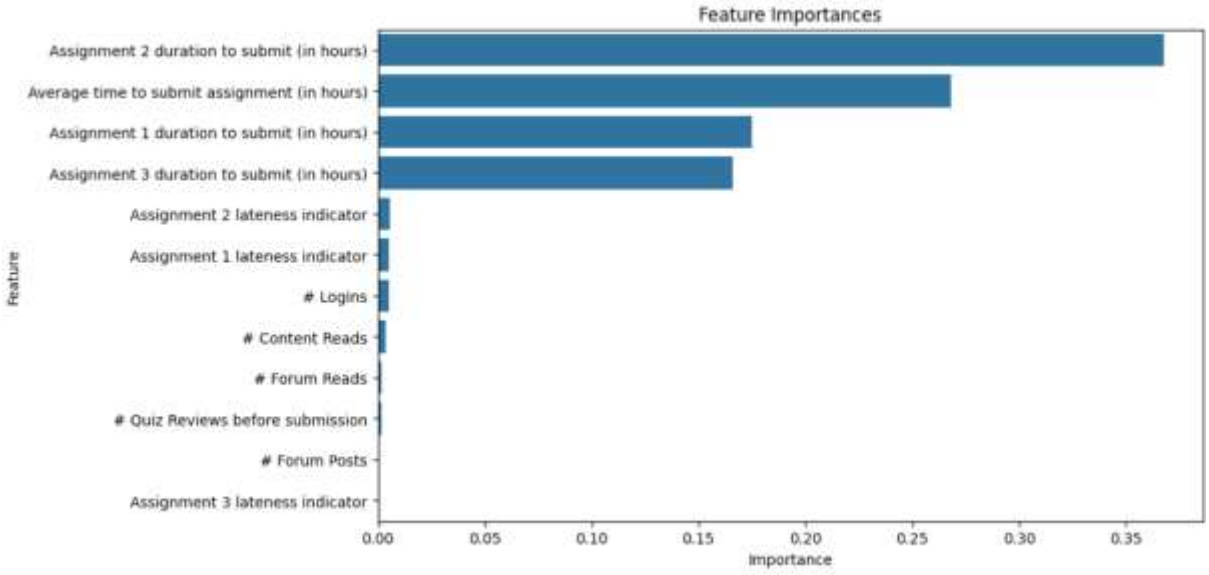


Fig 5.8: Features Importance's For A Machine Learning Model

The image shows a bar plot of feature importance's for a machine learning model, likely a decision tree-based model such as a Random Forest Classifier. The plot highlights the importance of various features in predicting the target variable.

10.48047/jocaaa.2025.34.06.14

8. Conclusion & Future Scope

Predictive analysis of student performance in online learning provides valuable insights that can enhance educational outcomes. By leveraging data such as engagement patterns, assessment scores, and learning behaviours, institutions can identify at-risk students, tailor interventions, and optimize teaching strategies. This data-driven approach allows educators to forecast academic achievements and challenges, helping to create personalized learning experiences that foster student success. Furthermore, predictive models contribute to resource allocation and curriculum design improvements by highlighting key factors influencing performance. As online education continues to grow, embracing predictive analytics will be crucial for fostering more inclusive and effective learning environments.

Future Scope: The future of predictive analysis in online learning is promising, driven by advances in AI and big data. As models become more accurate, they will offer real-time insights into student performance. Emerging technologies like AI-driven adaptive learning, IoT, and wearable devices will enable personalized and data-rich learning experiences. Additionally, block chain can enhance data security and transparency. Together, these innovations will make predictive analytics a vital tool for creating more effective, inclusive, and personalized online education environments.

Reference

1. Alhothali, A.; Albsisi, M.; Assalahi, H.; Aldosemani, T. Predicting student outcomes in online courses using machine learning techniques: A review. Sustainability 2022, 14, 6199. [CrossRef]
2. Tao, T.; Sun, C.; Wu, Z.; Yang, J.; Wang, J. Deep Neural Network-Based Prediction and Early Warning of Student Grades and Recommendations for Similar Learning Approaches. Appl. Sci. 2022, 12, 7733. [CrossRef]
3. Hughes, G.; Dobbins, C. The utilization of data analysis techniques in predicting student performance in massive open online courses (MOOCs). Res. Pract. Technol. Enhanc. Learn. 2015, 10, 10. [CrossRef]
4. Aljohani, N.R.; Fayoumi, A.; Hassan, S.U. Predicting at-risk students using clickstream data in the virtual learning environment. Sustainability 2019, 11, 7238. [CrossRef]
5. Liu, Y.; Fan, S.; Xu, S.; Sajjanhar, A.; Yeom, S.; Wei, Y. Predicting Student performance using clickstream data and machine learning. Educ. Sci. 2022, 13, 17. [CrossRef]

10.48047/jocaaa.2025.34.06.14

6. Casalino, G.; Castellano, G.; Zaza, G. Explainable Fuzzy Models for Learning Analytics. In International Conference on Intelligent Systems Design and Applications; Springer: Cham, Switzerland, 2022; pp. 394–403.
7. Arashpour, M.; Golafshani, E.M.; Parthiban, R.; Lamborn, J.; Kashani, A.; Li, H.; Farzanehfar, P. Predicting individual learning performance using machine-learning hybridized with the teaching-learning-based optimization. Comput. Appl. Eng. Educ. 2023, 31, 83–99[CrossRef]
8. Hussain, S.; Khan, M.Q. Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. Ann. Data Sci. 2023, 10, 637–655. [CrossRef]
9. Conijn, R.; Snijders, C.; Kleingeld, A.; Matzat, U. Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. IEEE Trans. Learn. Technol. 2016, 10, 17–29. [CrossRef]
10. Meier, Y.; Xu, J.; Atan, O.; Van der Schaar, M. Predicting grades. IEEE Trans. Signal Process. 2015, 64, 959–972. [CrossRef]
11. Ahmad, M.S.; Asad, A.H.; Mohammed, A. A Machine Learning Based Approach for Student Performance Evaluation in Educational Data Mining. In Proceedings of the 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), Cairo, Egypt, 26–27 May 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 187–192.
12. Zhang, W.; Huang, X.; Wang, S.; Shu, J.; Liu, H.; Chen, H. Student performance prediction via online learning behavior analytics. In Proceedings of the 2017 International Symposium on Educational Technology (ISET), Hong Kong, China, 27–29 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 153–157.
13. Al-Shehri, H.; Al-Qarni, A.; Al-Saati, L.; Batoaq, A.; Badukhen, H.; Alrashed, S.; Alhiyafi, J.; Olatunji, S.O. Student performance prediction using support vector machine and k-nearest neighbor. In Proceedings of the 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE),Windsor, ON, Canada, 30 April–3 May 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–4.
14. Wang, C.;Wei, X.; Yang, A.; Zhang, H. Construction and Analysis of Discrete System Dynamic Modeling of Physical Education Teaching Mode Based on Decision Tree Algorithm. Comput. Intell. Neurosci. 2022, 2022, 2745146. [CrossRef] [PubMed]
15. Rao, G.M.; Kumar, P.K.K. Students Performance Prediction in Online Courses Using Machine Learning Algorithms. United Int. J. Res. Technol 2021, 2, 74–79.