Early Detection of At-Risk Students Using Machine Learning: An Empirical Study on the OULAD with a Focus on Ethical Feature Selection

Dr. Mahmoud Yousef AlFaress[1] *0009-0001-8949-024X*, *Prof. Dr. Midhunchakkaravarthy Janarthanan*[2] *0000-0002-0107-885X*, *Prof. Dr. Chandra Kumar Dixit* [3]

[1,2] Lincoln University College – Malaysia; [3] Institute of Engineering and Technology, DSMNRU, Lucknow UP India

pdf.yousef@lincoln.edu.my; midhun@lincoln.edu.my; ckdixit@dsmnru.ac.in

---

Abstract: Timely and accurate prediction of at-risk students remains a central challenge in learning analytics. While machine learning (ML) provides strong predictive capability, the selection of input features is a high-stakes decision with direct implications for accuracy, fairness, and actionability. Building on our prior work that synthesized ethical trade-offs in feature selection across educational datasets [1, 2], this study implements an empirical predictive pipeline using the Open University Learning Analytics Dataset (OULAD). We engineered features from student demographics, virtual learning environment (VLE) interactions, and assessment records, and evaluated five ML models. The Gradient Boosting classifier achieved the best performance (Accuracy: 91.02%, F1-Score: 91.18%, Precision: 94.72%). These results indicate that behavioral engagement data are strong predictors of student success, consistent with our earlier cross-dataset findings [2]. However, the model's high precision warrants caution: it may be driven by features correlated with socioeconomic status or behavioral patterns that do not directly reflect academic ability. This paper contributes a validated technical framework for at-risk prediction and situates the results within the ethical context established in our prior research, underscoring the need for bias auditing prior to any deployment.

Keywords: Learning Analytics, Educational Data Mining, Machine Learning, Student At-Risk Prediction, OULAD, Feature Engineering, Algorithmic Bias, Ethical AI.

---

1. Introduction

Artificial Intelligence (AI) has the potential to transform education by offering personalized support to students. A key application is the early identification of those who might fail or drop out, allowing teachers to help them on time [3]. However, using AI in education comes with significant challenges. As we showed in our earlier work, one of the biggest challenges is choosing what data to use in the AI models [1], [2]. This decision is not just technical; it has major ethical implications.

The data we choose can introduce bias, violate privacy, or lead to unfair outcomes for certain groups of students [2], [4]. Our analysis of several educational datasets revealed a clear pattern: using detailed data on student behavior (like clicks in a virtual learning environment) leads to accurate predictions but raises privacy concerns. Using demographic data (like socioeconomic status) can make models unfair by repeating existing societal inequalities [2].

This study moves from analysis to action. We use the Open University Learning Analytics Dataset (OULAD) to ask a practical question: **Can we build a highly accurate model to predict at-risk students using OULAD, while carefully considering the ethical risks we know exist?**
We built a complete machine learning pipeline. Our results confirm that data on student engagement is a powerful predictor, which matches other research [5] and our own previous findings [2]. However, building an accurate model is only the first step. This paper discusses these results through an ethical lens, making this study the technical foundation for the next phase of our research: a full investigation into the model's fairness [1].

2. Literature Review & Theoretical Framework

The field of using data to understand education has grown rapidly. Early work used simple statistics on grades, but now complex algorithms analyze how students interact with online systems [6]. Research consistently shows that data from Learning Management Systems (LMS) can predict student success [5], [7].

Our previous research provides the foundation for this study. First, **AlFaress et al. (2024)** [1] provided a clear method for preparing educational data and creating meaningful features for machine learning. Second, **AlFaress et al. (2025)** [2] analyzed the ethics of choosing data features, showing that one must balance prediction accuracy with fairness, privacy, and explainability. That paper warned that while OULAD's data is very predictive, it might unfairly disadvantage students who learn differently or have limited technology access [2].

This paper builds on that foundation. We use the data engineering methods from [1] and follow the ethical guidelines from [2]. For this study, we made a conscious choice: we used the data available in OULAD to make the model as accurate as possible, but we will not use features like imd_band (a measure of socioeconomic status) to target interventions, due to the high risk of bias [2], [4]. These features are included only so we can later analyze the model for bias. This makes the current study a key technical step in a larger, ethically-guided research plan.

3. Methodology

3.1. Data Source: The OULAD Dataset

This study utilizes the Open University Learning Analytics Dataset (OULAD) [8], a comprehensive public dataset containing data from 32,593 students. We integrated multiple tables, the scale of which is summarized below:

Table 1: Description of Raw OULAD Data Files

File Name	Records	Description
studentInfo	32,593	Demographic info & student outcomes.
studentVle	10,655,280	Student interactions with VLE materials.

studentAssessment	173,912	Student submissions & scores.
assessments	206	Assessment details & weightings.
courses	22	Course presentation information.

3.2. Data Preprocessing and Feature Engineering

Following our previous work [1], we transformed the raw data into features that represent student behavior and performance.

- **Engagement Features (From how students use the online system):**
  - total_clicks: Total number of clicks per student.
  - unique_vle_materials: Number of different learning materials accessed.
  - avg_clicks_per_day: Average clicks per day during the course.
- **Performance Features (From assessments):**
  - average_assessment_score: The student's average score on assignments.
  - assessment_score_trend: Whether their scores were improving or getting worse.
  - num_late_submissions: How many assignments were submitted late.
- **Demographic Data:** Information like gender, region, and imd_band (a measure of socioeconomic deprivation) was converted into a numerical format. Based on our ethical framework [2], these features were included for analysis but were not used as the primary basis for predicting risk, to avoid building bias into the model.

The final dataset ready for modeling contained **32,593 students and 44 features**.

3.3. Target Variable Definition

We defined our goal based on the student's final result:

- **At-Risk (1):** Fail or Withdrawn
- **Not At-Risk (0):** Pass or Distinction

3.4. Machine Learning Models and Evaluation

The data was split: 70% for training the models and 30% for testing them. We made sure both sets had a similar mix of at-risk and not-at-risk students. We trained five different algorithms:

1. **Logistic Regression:** A simple, interpretable baseline model.
2. **Decision Tree:** A model that makes decisions based on rules.
3. **Random Forest:** A powerful model that combines many decision trees.
4. **Gradient Boosting:** Another powerful model that builds trees sequentially to correct errors.
5. **Support Vector Machine (SVM):** A model effective for complex data.

We evaluated the models based on **Accuracy, Precision, Recall, and F1-Score**. We also created confusion matrices and ROC curves to understand their performance in detail.

4. Results and Analysis

The performance of all five models is summarized in Table 2. The Gradient Boosting model was the best.

Table 2: Model Performance on the Test Set

Model	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	**0.9102**	**0.9472**	0.8789	**0.9118**
Logistic Regression	0.9039	0.9326	0.8816	0.9064
Random Forest	0.9009	0.9369	0.8710	0.9027
Decision Tree	0.8803	0.8876	**0.8855**	0.8865
Support Vector Machine (SVM)	0.8150	0.8603	0.7755	0.8157

The Analysis:

- The **Gradient Boosting Machine** was the most accurate (91.02%) and had the best F1-Score (91.18%). Most importantly, its **precision was 94.72%**. This means that when it predicts a student is at-risk, it is correct 95 times out of 100. This is crucial for avoiding wasted resources on false alarms.
- The strong performance of the Random Forest and Gradient Boosting models shows that the patterns in the data are complex and non-linear.
- These results prove that the features we created from engagement and performance data are highly predictive.

4.1. Visual Model Evaluation

The confusion matrix and ROC curve for the top-performing Gradient Boosting model provide deeper insight into its performance.

Figure 1 shows the confusion matrix for the Gradient Boosting model. The large numbers on the diagonal (from top-left to bottom-right) show the correct predictions. The very small number in the top-right corner (False Positives) visually confirms the model's high precision

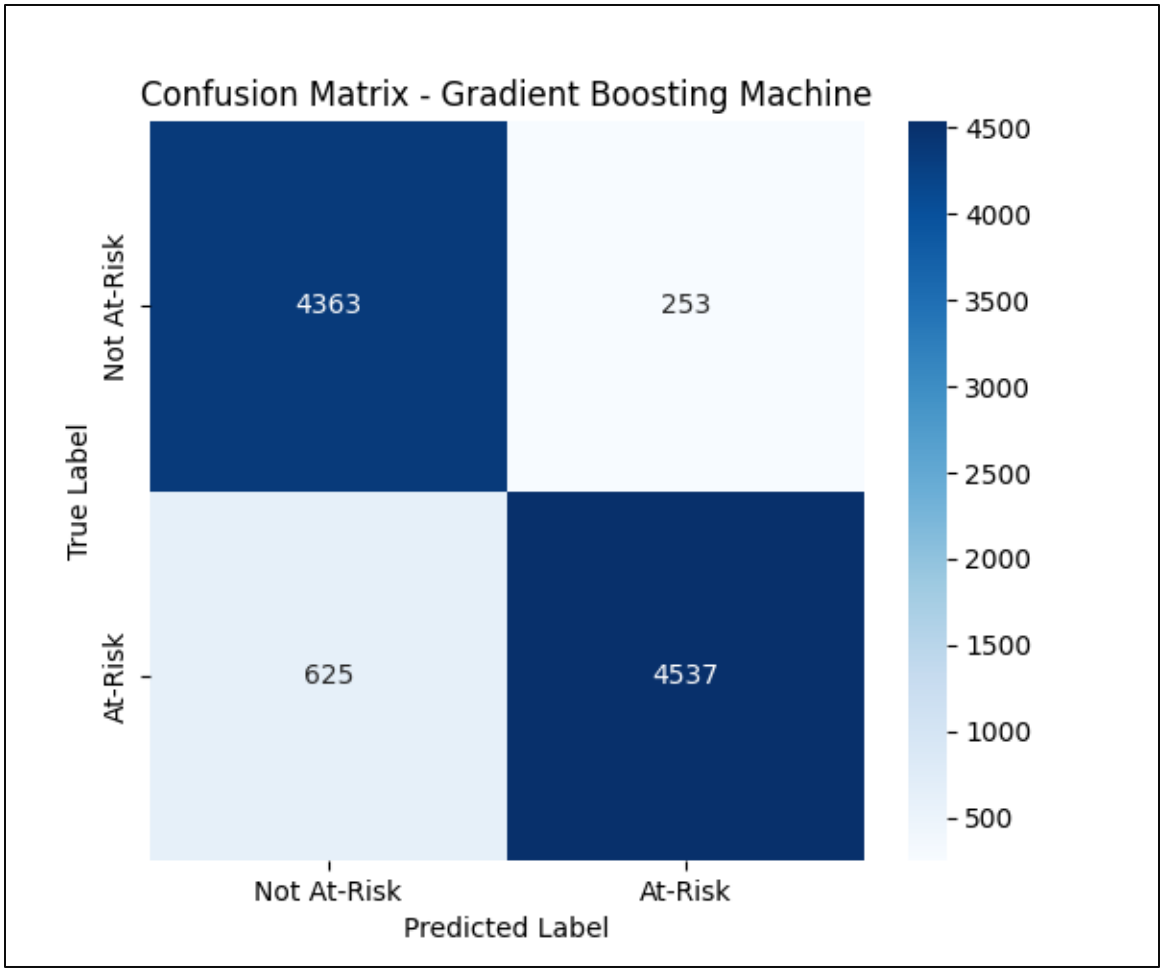

Figure 1: Confusion Matrix for Gradient Boosting Model

Figure 2 presents the Receiver Operating Characteristic (ROC) curve. The curve's strong shift to the top-left corner and the high Area Under the Curve (AUC) value demonstrate the model's excellent ability to discriminate between at-risk and not-at-risk students across all classification thresholds.

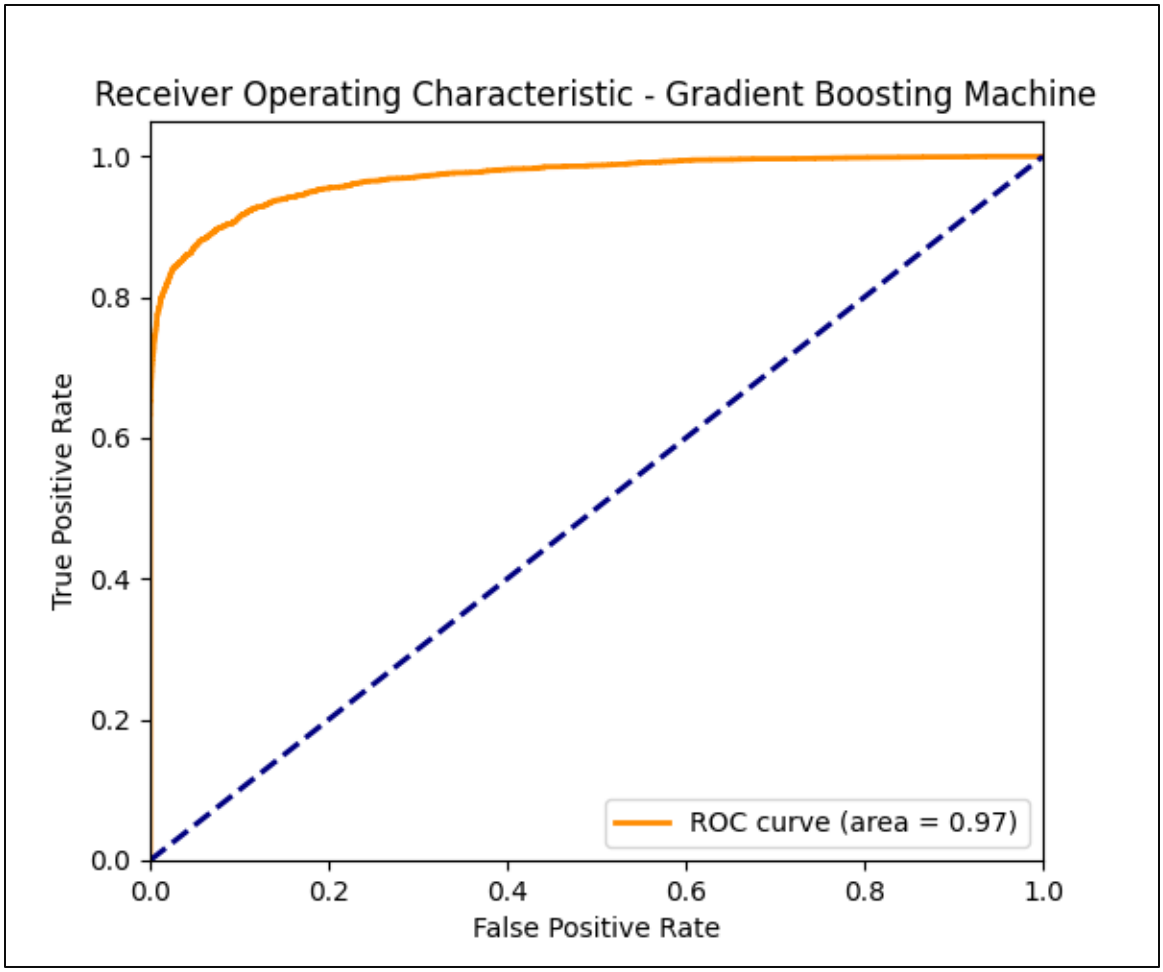

Figure 2: ROC Curve for Gradient Boosting Model

.2. Feature Importance Analysis

To understand what drives the model's predictions, we extracted the feature importance scores from the Gradient Boosting classifier. This analysis reveals which features the model found most useful for making its decisions, providing critical insight for interpretation and actionability.

Figure 3 plots the mean decrease in impurity (Gini importance) for the top 15 most important features. The results are highly informative:

- **Assessment Scores are Key:** The student's average_assessment_score was the strongest predictor. This makes intuitive sense and matches findings from other studies [9].
- **Engagement Matters:** Features like total_clicks and avg_clicks_per_day were also among the top predictors. This proves that how a student interacts with the online learning system is a very strong signal of their success, confirming our earlier analysis [2].
- **A Warning Sign:** The demographic feature imd_band (socioeconomic status) was also a top feature. This is a critical finding. It means the model is likely using socioeconomic status to make predictions. If used without care, this could lead to a unfair system where students from poorer backgrounds are automatically labeled as at-risk, not because of their ability, but because of their background [4]. This finding directly justifies the need for the bias check we propose next.

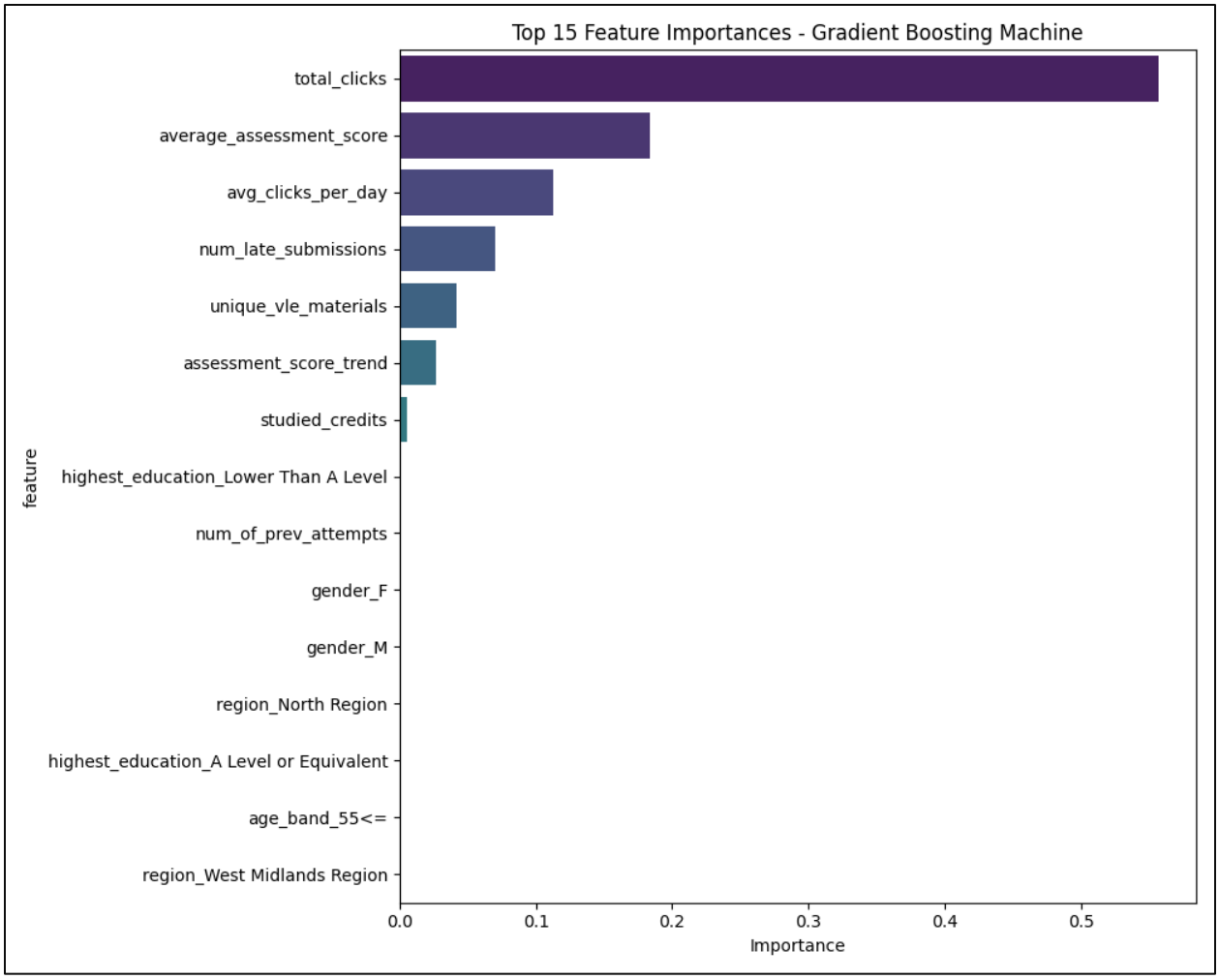

Figure 3: Top 15 Feature Importances from the Gradient Boosting Model

5. Discussion

This study successfully developed a high-accuracy predictive model for identifying at-risk students, thereby answering the primary research question. However, in line with our established framework [1, 2], we must interpret these results beyond mere performance metrics.

5.1. Interpretation of Findings and Ethical Implications

The model's high performance, particularly its precision, suggests it could be highly effective in targeting support resources. The feature importance analysis (Figure 3) clarifies the source of this predictive power: a combination of **assessment scores** and **VLE engagement metrics**.

However, this success is potentially double-edged, as discussed in [2].

- **The Source of Predictive Power:** The model's accuracy is driven by the very features we engineered (total_clicks, avg_clicks_per_day). This confirms our prior analysis [2]. However, as critically noted, VLE clickstream data is not a pure measure of engagement. It can reflect

privilege (stable internet access, dedicated study time) and **learning style** just as much as academic diligence. A model rewarding high click counts could therefore systematically disadvantage students with limited connectivity or caregiving responsibilities.

- **The Risk in Features:** The notable importance of imd_band is a clear red flag. It indicates that the model is likely learning associations between socioeconomic status and academic outcome. If used naively, this could lead to a self-fulfilling prophecy where students from disadvantaged backgrounds are disproportionately flagged as at-risk, not because of their potential, but because of their background [4]. This **empirically observed risk** is why we explicitly excluded demographics from intervention targeting in our methodology and why a fairness audit is the mandatory next step.

5.2. Limitations and Future Work: A Research Trajectory

This study has limitations that directly chart the course for our future work.

1. **Bias Audit:** The foremost limitation is the absence of a formal fairness audit. As a direct continuation of this work, we will apply the methodology from [1] and the critical lens from [2] to analyze model performance (precision, recall) across demographic subgroups (e.g., by imd_band, age_band). This is essential to determine if the model's high overall performance masks disparate impacts on specific groups.
2. **Hyperparameter Tuning:** Performance could potentially be improved further through systematic hyperparameter optimization.
3. **Causal Intervention:** This study focuses on prediction. The logical next step is to implement the proposed personalized intervention framework and measure its causal impact on student outcomes through a controlled experiment.

This paper represents the second step in a planned trilogy: (1) establishing a methodological and ethical foundation [1, 2], (2) empirical technical implementation (this study), and (3) rigorous bias mitigation and intervention analysis (future work).

6. Conclusion

This research provides a robust, empirically validated machine learning pipeline for early detection of at-risk students using the OULAD dataset. The Gradient Boosting model achieved outstanding performance (91.02% accuracy, 94.72% precision), demonstrating the powerful predictive signal contained in student engagement and assessment data.

However, in consciously adhering to the ethical framework established in our previous work, we conclude that high accuracy is necessary but not sufficient for responsible AIEd. The features that drive this model's success are the same ones that carry inherent risks of bias and surveillance. Therefore, we contend that the development of predictive models must be intrinsically linked with their ethical audit.

This study delivers the high-performance model; it now imperative that we subject it to the rigorous fairness analysis it requires. By doing so, we can strive to move from simply building accurate predictors to developing trustworthy, equitable, and truly supportive tools that enhance education for all students.

References

1. M. Y. AlFaress, M. Janarthanan, and C. K. Dixit, "A novel approach for data preprocessing and feature engineering in educational data mining," *SPAST Reports*, vol. 1, no. 1, 2024. [Online]. Available: https://spast.org/techrep/article/view/5278
2. M. Y. AlFaress, M. Janarthanan, and C. K. Dixit, "The impact of data feature selection on AI-based student performance prediction: An analysis of educational use cases," SGS Engineering & Sciences, vol. 1, no. 2, 2025. [Online]. Available: https://spast.org/techrep/article/view/5415
3. A. Essa and H. Ayad, "Improving student success using predictive models and data visualisations," Research in Learning Technology, vol. 20, 2012, Available: https://doi.org/10.3402/rlt.v20i0.19191
4. R. S. Baker and A. Hawn, "Algorithmic bias in education," International Journal of Artificial Intelligence in *Education*, vol. 32, no. 4, pp. 1052–1092, 2022. [Online]. Available: https://doi.org/10.1007/s40593-021-00285-9
5. Á. F. Agudo-Peregrina, Á. Hernández-García, and F. J. Iglesias-Pradas, "Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning," Computers in Human Behavior, vol. 31, pp. 542–550, 2014. [Online]. Available: https://doi.org/10.1016/j.chb.2013.05.031
6. C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 10, no. 3, p. e1355, 2020. [Online]. Available: https://doi.org/10.1002/widm.1355
7. E. Howard, M. Meehan, and A. Parnell, "Contrasting prediction methods for early warning systems at undergraduate level," The Internet and Higher Education, vol. 37, pp. 68–75, 2018. [Online]. Available: https://doi.org/10.1016/j.iheduc.2018.02.001
8. J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics dataset," Scientific Data, vol. 4, no. 1, p. 170171, 2017. [Online]. Available: https://doi.org/10.1038/sdata.2017.171
9. P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance" in Proceedings of the 5th Annual Future Business Technology Conference, 2008, pp. 5–12. Available:https://www.researchgate.net/publication/228780408_Using_data_mining_to_predict_secondary_school_student_performance
10. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," ACM *Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021. [Online]. Available: https://doi.org/10.1145/3457607