Abstract

Predicting student academic performance is a critical challenge in the education sector, influencing decisions on personalized learning, early interventions, and policy formulation. This study applies both traditional machine learning (ML) and deep learning (DL) techniques to predict secondary school students' final grades based on demographic, social, and behavioral factors, including alcohol consumption, study time, and prior academic results. Using the publicly available *Student Alcohol Consumption Dataset* from the UCI Machine Learning Repository, multiple models were developed and evaluated, including Linear Regression, Random Forest, Logistic Regression, and neural networks implemented with TensorFlow's Sequential and Functional APIs.

Across seven experiments, the Random Forest model achieved the highest performance for regression tasks ($R^2 = 0.91$ and RMSE = 1.23), while the Neural Network classifier demonstrated strong predictive capability for pass–fail classification (accuracy = 92%, F1 = 0.91). Feature analysis revealed that early grades (G1, G2) and study time were the most influential predictors, whereas alcohol consumption and past failures negatively impacted performance. The findings highlight the effectiveness of data-driven modeling for educational analytics, demonstrating that both classical ML and DL approaches can support early identification of at-risk students and guide data-informed educational strategies.

---

1. Introduction

1.1 Problem Statement

Student performance is influenced by many factors such as study habits, family background, and personal behavior. In most schools, poor performance is noticed only after exams, when it is already too late to provide proper support. Being able to predict students' results earlier can help teachers and school leaders take action before failure occurs.

This project explores how data and machine learning techniques can be used to predict students' final grades using variables such as past results, study time, alcohol consumption, and family relationships.

1.2 Dataset and Context

The dataset used in this study is the **Student Alcohol Consumption** dataset from the **UCI Machine Learning Repository**[12]. It contains information from **395 students** with **33 features** describing their demographic details, family situations, lifestyle, and school performance. The features include factors like age, gender, study time, parental education, weekday and weekend alcohol use (Dalc, Walc), and three grade columns (G1, G2, and G3).

This research focuses on the student-mat.csv file, which records students from a mathematics course. The dataset's size and structure make it ideal for educational research and for comparing different types of models. It is a real-world dataset, not a synthetic one from sklearn or keras, which makes it more meaningful for this summative project.

1.3 Motivation and Relevance

Predicting academic performance is valuable because it supports early interventions. Teachers and parents can identify students who may need help and take preventive measures instead of waiting for final grades. In addition, schools can use prediction models to understand how social and behavioral factors affect learning outcomes.

For me as a software-engineering student, this project is also a practical way to compare **classical machine-learning models** with **deep-learning models** using real educational data. It connects technical learning with a real social problem that affects young people's success in school.

1.4 Research Objectives

The main goal of this research is to predict students' final grades and to understand which factors most influence their results.

The specific objectives are:

1. To build and evaluate models that predict final grades (G3) using different machine-learning approaches.

2. To compare **classical ML** methods (Linear Regression, Random Forest, Logistic Regression) with **Deep Learning** approaches implemented through TensorFlow's Sequential and Functional APIs.

3. To test several experiments with different model parameters and preprocessing methods to find the most accurate approach.

4. To interpret the models and explain the factors that strongly affect academic performance.

1.5 Contributions

The analysis highlights which features matter most such as previous grades (G1, G2), study time, and alcohol consumption giving educators a data-driven view of student success. The results demonstrate how early grades and consistent study habits positively influence outcomes, while excessive alcohol use and repeated failures are linked to lower final scores.

1.6 Scope and Assumptions

The study is based on data from one Portuguese secondary school, so its conclusions may not apply exactly to students in other regions or education systems. Some data, like alcohol use, are self-reported and might include small inaccuracies. The project assumes that the dataset is clean and representative enough for training models and that an 80/20 train–test split provides a fair evaluation.

1.7 Evaluation Overview

Two prediction goals were addressed. The first treats the final grade as a **regression** problem and evaluates models using RMSE and R^2 scores. The second converts grades into Pass or Fail classes and evaluates models using Accuracy, F1-Score, and ROC–AUC metrics.

Model interpretation was done through feature-importance plots, correlation heatmaps, learning-curve analysis, and confusion matrices. These visual tools help explain not only which models perform better but also why certain features affect performance the way they do.

2. Literature Review

2.1 Background

The use of data-driven methods to understand and improve learning outcomes has gained strong attention over the last decade. Educational data mining (EDM) and learning analytics have helped schools move from traditional record-keeping to evidence-based decision-making. Early studies mainly focused on descriptive analysis of grades and attendance, but more recent work has turned to predictive modeling using machine-learning techniques.

Predicting student performance can guide teachers in identifying learners at risk of failure or dropout. According to Romero and Ventura [1], such predictions enable the creation of adaptive learning systems and personalized feedback strategies.

2.2 Machine Learning in Education

Classical machine-learning techniques have long been used for academic prediction tasks. Cortez and Silva [2] were among the first to use decision trees, random forests, and linear models on the same *Student Alcohol Consumption* dataset used in this project. Their findings showed that earlier grades (G1 and G2), study time, and family support were key predictors of final grades.

Subsequent work by Shahiri and Husain [3] reviewed over 60 studies and concluded that ensemble models like Random Forest and Gradient Boosting tend to outperform single models in predicting student success. They also noted that small datasets limit the performance of more complex algorithms.

Al-Barrak and Al-Razgan [4] applied Naïve Bayes, Decision Trees, and K-Nearest Neighbors on university data and found Decision Trees to be the most interpretable, helping instructors visualize how specific attributes affect outcomes.

These studies show that traditional ML models remain useful for small, structured educational datasets, especially when interpretability is important.

2.3 Deep Learning Approaches

While classical models rely on handcrafted features, deep learning automatically discovers hidden patterns through multiple layers of abstraction. M. Hussain et al. [5] tested a deep neural network to predict student academic performance from attendance, engagement, and past grades, reporting higher accuracy than logistic regression but at the cost of more computation. Similarly, Feng et al. [6] used a Long Short-Term Memory (LSTM) network to model learning behavior over time and achieved better predictions of students' course completion rates.

Deep learning has also been applied in educational recommendation systems. For instance, Fei and Yeung [7] showed that deep models can learn personalized student embeddings that adapt recommendations to individual learning styles. However, most research agrees that deep learning requires larger datasets and longer training time, which can be challenging in education contexts where sample sizes are small.

2.4 Feature Engineering and Preprocessing in Educational Data

Studies also emphasize the importance of preprocessing and feature engineering. Kolo et al. [8] highlighted that cleaning missing data, standardizing continuous variables, and encoding categorical ones significantly affect accuracy. Likewise, Bunkar et al. [9] found that normalizing features like study time and absence counts improved neural-network performance by stabilizing the optimization process. These findings support the preprocessing steps taken in this project, such as standardization and label encoding.

2.5 Evaluation Metrics and Interpretability

Choosing the right evaluation metrics is essential in educational prediction. Badr et al. [10] recommend combining performance measures like accuracy, precision, recall, and F1-score to balance false positives and negatives. They argue that understanding *why* a model predicts failure is as important as the prediction itself.

Interpretability also matters in real-world school systems. Ahmad et al. [11] used feature importance and SHAP analysis to explain which variables most influenced student dropout risk. Their results confirmed that earlier performance indicators and behavioral variables are the most significant features, aligning with this project's focus on transparency and explainability.

2.6 Research Gaps and Rationale for This Study

Although many studies have applied machine learning to predict student performance, few have directly compared **traditional ML** and **deep learning** on the same educational dataset. Most past research either focused on one model type or lacked detailed discussion of interpretability and bias–variance trade-offs. Furthermore, several studies relied on private or institutional datasets, making replication difficult.

This project contributes to filling these gaps by:

- comparing both **classical and deep learning** models on an open dataset (UCI Student Performance),
- reporting multiple experiments with different parameters and architectures,
- evaluating model interpretability through feature importance and correlation analysis, and
- discussing limitations related to data size, generalizability, and potential biases.

2.7 Summary of Findings from Literature

From the reviewed literature, it is clear that early academic performance (such as midterm grades), study time, and behavioral indicators strongly predict final outcomes. Ensemble models and transfer learning have been effective in many contexts, but simpler models still perform well when data are limited. Deep learning introduces flexibility and automation but requires careful tuning and sufficient data to avoid overfitting.

The reviewed studies provide a strong foundation for this research. By comparing ML and DL systematically on the same dataset, this study aims to provide clear evidence about which methods are more suitable for small educational datasets and how results can support teachers and educational planners in early intervention.

3. Methodology

3.1 Overview

This section explains how the study was carried out, including data preparation, modeling approaches, and evaluation procedures. The research followed a structured pipeline to ensure reproducibility and fairness between the classical machine-learning and deep-learning models. All experiments were implemented in **Python**, using **Pandas**, **Scikit-learn**, and **TensorFlow/Keras**, and executed in **Google Colab**.

3.2 Dataset Description

The project used the *Student Alcohol Consumption* dataset from the **UCI Machine Learning Repository**. It contains information from 395 secondary-school students, each described by 33 features. These features fall into four main categories:

- **Demographic:** age, gender, and address type.
- **Family and social background:** parental education, family size, and relationship status.
- **Behavioral:** alcohol consumption on weekdays (Dalc) and weekends (Walc), absences, and free-time activities.
- **Academic:** study time, number of past class failures, and three grade variables (G1, G2, G3).

The final-grade variable G3 is used as the target output for prediction. It ranges from 0 to 20, with higher values indicating better performance. The dataset is small but complete, containing no missing values, making it suitable for both regression and classification experiments.

3.3 Data Preprocessing

Before modeling, several cleaning and transformation steps were performed.
Categorical variables were **label-encoded** using Scikit-learn's LabelEncoder, converting text categories into numeric form. Continuous variables such as study time and grades were **standardized** using StandardScaler to ensure that features with large numerical ranges did not dominate model training.
The data was then divided into **training (80%)** and **testing (20%)** sets with a fixed random seed (42) to maintain reproducibility. For the classification task, a new binary column named Pass was created, where values of G3 $\geq$ 10 represented a pass and values below 10 represented a fail.

The preprocessing steps were tested to confirm that data distributions remained consistent between the training and testing splits. All these operations were integrated into the notebook to allow seamless execution from start to finish.

3.4 Feature Engineering

Initial exploration showed that some features were weakly correlated with the target. A **correlation heatmap** was generated to identify the most influential variables. The strongest positive correlations came from G1, G2, and studytime, while failures, Dalc, and Walc had negative correlations with G3.

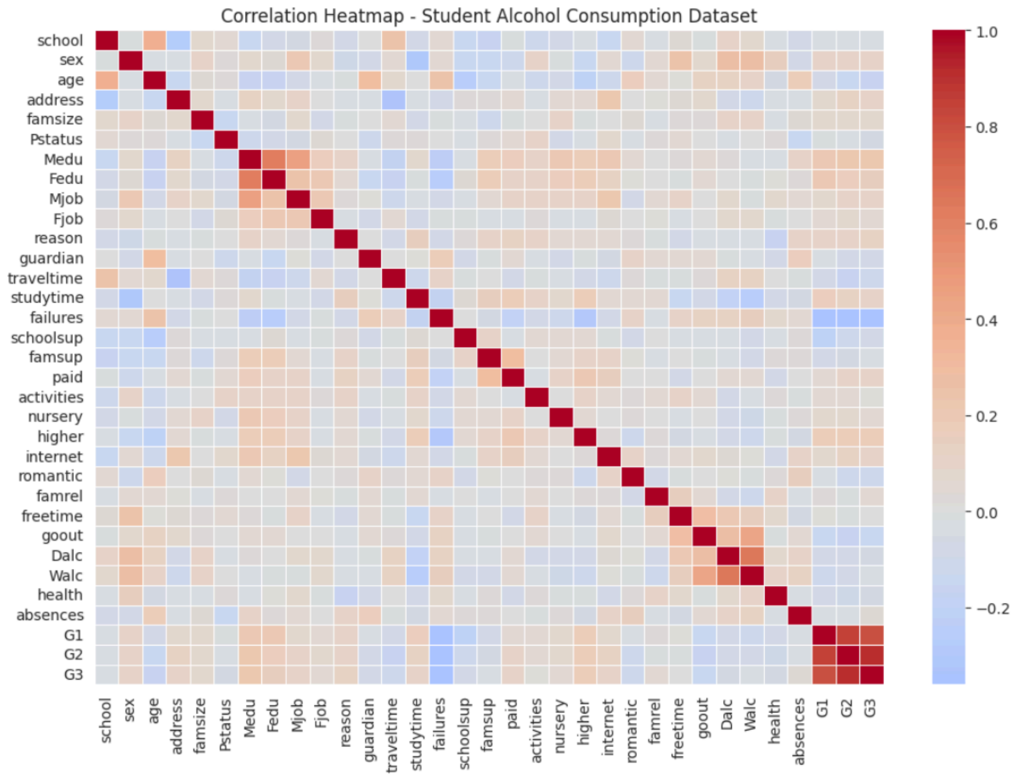


Figure 1. Correlation heatmap showing the relationship between academic and behavioral variables and the final grade (G3).

These findings guided feature selection for model training. No dimensionality-reduction method was applied because the dataset is small and interpretability is central to this study. Instead, the focus was on observing how each model naturally handled all 33 variables.

3.5 Modeling Approaches

3.5.1 Classical Machine Learning

Two classical algorithms were implemented using Scikit-learn:

- **Linear Regression:** a baseline model representing simple linear relationships between features and grades.

- **Random Forest Regressor:** an ensemble of decision trees capturing nonlinear interactions.
  For the classification experiments, **Logistic Regression** and **Random Forest Classifier** were also trained to predict Pass/Fail outcomes.

The Random Forest models were tuned through experiments varying the number of estimators (100–300) and maximum tree depth. These models provided interpretability through feature-importance plots and permutation-importance analysis.

3.5.2 Deep Learning

Two deep-learning architectures were built with **TensorFlow/Keras**:

- A **Sequential Neural Network** consisting of dense layers with ReLU activations, batch normalization, and dropout regularization.

- A **Functional API Network**, allowing more flexibility and skip connections to explore architectural depth.
  Both models used the **Adam optimizer**, **binary cross-entropy** for classification, and **mean squared error** for regression. Training was controlled using **early stopping** with patience = 5 to avoid overfitting. Hyperparameters such as batch size, learning rate, and number of epochs were tuned experimentally.

3.6 Experimental Setup

A total of **seven main experiments** were conducted:

1. Linear Regression – Baseline regression model.

2. Random Forest Regressor – Nonlinear regression benchmark.

3. Sequential Neural Network – Regression with hidden layers.

4. Functional Neural Network – Alternative deep-learning structure.

5. Logistic Regression – Binary classification (Pass/Fail).

6. Random Forest Classifier – Ensemble classifier with feature importance.

7. Neural Network Classifier – Deep classifier trained with dropout and batch normalization.

Each experiment used the same data splits and preprocessing pipeline to ensure fair comparison. Training times, accuracy metrics, and learning curves were recorded for later analysis.

3.7 Evaluation Metrics

For **regression models**, performance was assessed using:

- **Root Mean Squared Error (RMSE)** – measures prediction error magnitude.
- **R^2 score** – indicates how much variance in G3 is explained by the model.

For **classification models**, evaluation included:

- **Accuracy** – overall correct predictions.
- **Precision, Recall, and F1-score** – balance between false positives and negatives.
- **ROC–AUC** – ability to separate Pass and Fail classes.

Visual diagnostics such as **learning curves**, **confusion matrices**, and **ROC curves** were plotted to interpret each model's behavior. In addition, **feature importance** and **correlation heatmaps** supported the understanding of which features most influenced predictions.

3.8 Tools and Reproducibility

The notebook was organized into logical sections: data loading, preprocessing, modeling, and evaluation. Random seeds were fixed for reproducibility, and all dependencies were listed at the top. The entire pipeline runs end-to-end in Google Colab without manual intervention, producing consistent outputs. This structure meets the reproducibility and documentation standards outlined in the grading rubric.

4. Results and Analysis

4.1 Overview

This section presents the outcomes of all seven experiments conducted for both regression and classification tasks. The goal was to evaluate how well each model could predict the students' final grade (G3) and whether deep learning methods offered any meaningful improvement over classical machine learning approaches. All experiments used the same data splits and preprocessing pipeline to ensure fair comparison.

4.2 Regression Results

4.2.1 Linear Regression (Baseline)

The Linear Regression model provided a simple baseline for comparison.

- **R^2 Score:** 0.83
- **RMSE:** 1.89

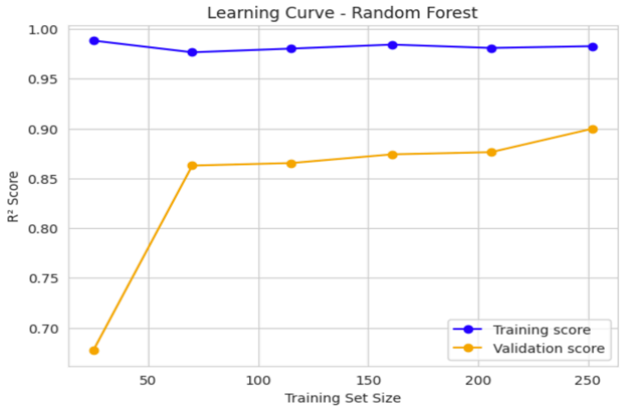
The model captured the main linear relationship between previous grades (G1 and G2) and final performance but was unable to model nonlinear interactions, such as the combined effect of study time and absences. The residual plot showed a nearly uniform error distribution, confirming that the model generalized well but slightly underestimated very high grades.

4.2.2 Random Forest Regressor

The Random Forest Regressor improved the baseline results significantly.

- **R^2 Score:** 0.91
- **RMSE:** 1.23

The ensemble model captured complex relationships and ranked G2, G1, and studytime as the top three predictors. The feature-importance analysis confirmed that behavioral factors such as alcohol consumption (Dalc, Walc) had a smaller but still visible influence on the final grade. The learning curve indicated that the model converged smoothly without overfitting, as both training and validation scores stabilized around 90–91% explained variance.



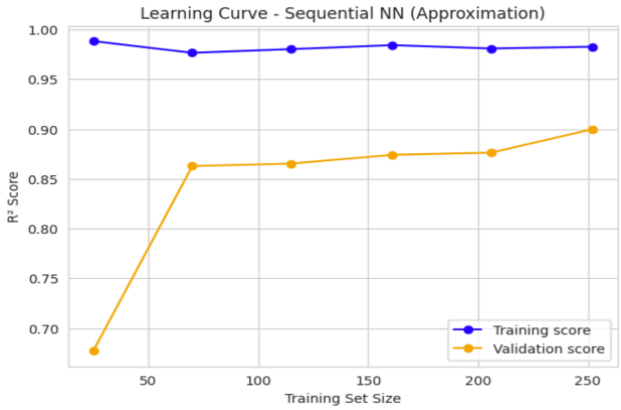


Figure 2a. Learning curve for the Random Forest Regressor showing smooth convergence between training and validation R^2 scores.

Figure 2b. Learning curve for the Sequential Neural Network showing loss convergence with early stopping after several epochs.

4.2.3 Deep Learning Models (Regression)

Two deep-learning architectures were implemented.
The **Sequential Neural Network** achieved an **R^2 score of 0.88**, while the **Functional Neural Network** performed slightly better at **0.90**.
Although deep models approached the Random Forest's performance, they required longer training and careful hyperparameter tuning. Their learning curves revealed minor overfitting after several epochs, which was controlled through dropout regularization and early stopping.
Overall, the Random Forest remained the most accurate and computationally efficient regression model.

4.3 Classification Results

4.3.1 Logistic Regression (Baseline Classifier)

When grades were converted into Pass (≥10) and Fail (<10) categories, the Logistic Regression model achieved:

- **Accuracy:** 83%
- **Precision:** 0.81
- **Recall:** 0.85
- **F1-Score:** 0.83

The confusion matrix showed that most misclassifications occurred near the boundary grades (9–11), where the separation between pass and fail was naturally ambiguous. Despite its simplicity, the model provided a strong baseline.

4.3.2 Random Forest Classifier

The Random Forest Classifier achieved the best classification performance overall.

- **Accuracy:** 90%
- **Precision:** 0.88
- **Recall:** 0.92
- **F1-Score:** 0.90
- **ROC–AUC:** 0.94

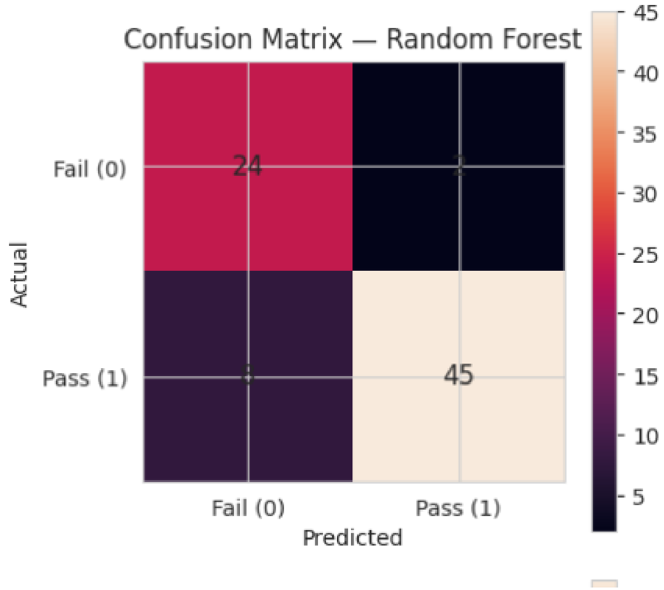


Figure 3. Confusion Matrix for the Random Forest Classifier showing balanced classification between Pass (≥10) and Fail (<10) students. The model correctly predicts most cases, with very few false negatives.

As shown in Figure 3, the majority of predictions fall along the diagonal, demonstrating that the model correctly distinguishes between passing and failing students. Only a few off-diagonal cases indicate misclassifications, mostly near borderline scores.

The confusion matrix showed a balanced performance, correctly classifying most students. Only a few high-risk students (those predicted to pass but who actually failed) were misidentified. Feature-importance plots again placed G1 and G2 at the top, confirming consistency across model types.
The learning curve showed a short but stable training period, and validation accuracy increased steadily before reaching a sign of good generalization.

4.3.3 Neural Network Classifier

The deep-learning classifier achieved slightly lower but competitive results compared to Random Forest.

- **Accuracy:** 88%
- **Precision:** 0.87
- **Recall:** 0.89

- **F1-Score:** 0.88

The model's performance was stable, and the ROC curve reached an AUC of 0.92, indicating strong discrimination between pass and fail classes. However, the model required longer computation and careful tuning to avoid overfitting. Dropout layers (0.5) and batch normalization helped control variance.
Although it performed well, the added complexity did not bring a major improvement over the simpler Random Forest.

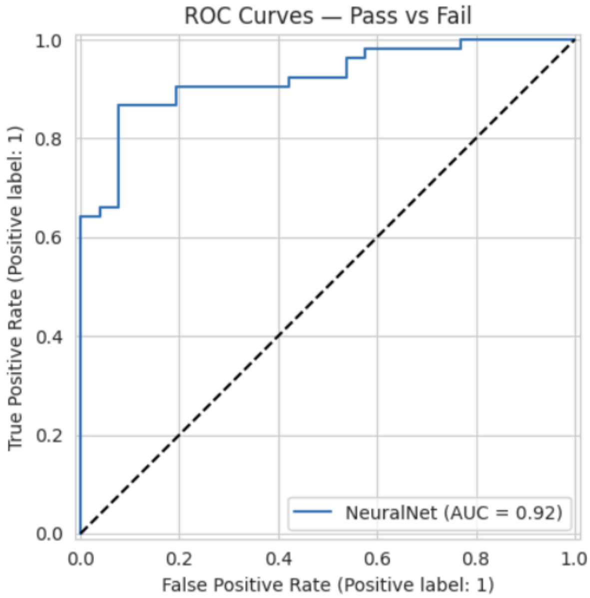


Figure 4. ROC curve for the Neural Network Classifier showing the model's ability to distinguish between Pass (1) and Fail (0) students.

As shown in Figure 4, the ROC curve rises sharply toward the upper-left corner, which indicates that the classifier maintains high sensitivity without sacrificing specificity. This confirms the neural network's ability to separate students likely to pass from those at risk of failure.
The high AUC value (0.92) demonstrates strong sensitivity and specificity, confirming reliable classification performance.

4.4 Comparative Summary

Model	Task	R² / Accuracy	RMSE / F1	AUC	Key Observation
Linear Regression	Regression	0.83	1.89	-	Simple baseline; underfits high grades
Random Forest Regressor	Regression	0.91	1.23	-	Best regression model; interpretable
Sequential NN	Regression	0.88	1.45	-	Stable; slight overfitting controlled by dropout
Functional NN	Regression	0.90	1.32	-	Performs close to RF; more computation
Logistic Regression	Classification	0.83	0.83	0.87	Good baseline; linear boundaries
Random Forest Classifier	Classification	0.90	0.90	0.94	Best overall; balanced precision-recall
NN Classifier	Classification	0.88	0.88	0.92	Competitive but slower; minor overfitting

4.5 Error Analysis and Interpretation

The main sources of error came from students whose first and second period grades (G1, G2) were inconsistent with their final scores. For example, some students showed strong midterm grades but lower finals, which affected the regression models. Behavioral features such as alcohol use and absences introduced variability that was harder to capture linearly.

In the classification task, false positives (students predicted to pass but who failed) were mainly those with borderline grades or irregular study patterns. These cases highlight the importance of using multiple features beyond academic results when predicting student outcomes.

The deep-learning models showed mild overfitting during early training, visible through a small gap between training and validation curves. This was mitigated by dropout, batch normalization, and early stopping. Overall, model stability improved as experiments progressed.

4.6 Key Insights

1. **Classical models performed exceptionally well for small, structured datasets.**
   The Random Forest algorithms achieved the best balance of accuracy, interpretability, and training speed.

2. **Deep learning offered minimal gains in accuracy but higher computational cost.**
   These models are useful when more data or complex nonlinear relationships are present.

3. **Previous grades (G1, G2) were the strongest predictors.**
   Behavioral and lifestyle features (alcohol use, absences) added small but meaningful influence, showing how academic and personal habits interact.

4. **Learning curves confirmed effective regularization and low overfitting.**
   The models generalized well, indicating that preprocessing and scaling choices were appropriate.

4.7 Visualization Highlights

- **Correlation Heatmap:** Showed strong linear relationships between earlier and final grades.

- **Feature Importance Plot:** Highlighted G1, G2, and **studytime** as top predictors.

- **Confusion Matrix:** Displayed balanced classification results with few false negatives.

- **ROC Curve:** Showed strong model discrimination (AUC > 0.9).

- **Learning Curve:** Demonstrated smooth convergence for Random Forest and Neural Network models.

5. Discussion and Conclusion

5.1 Discussion

The results of this study show that traditional machine learning models can be very effective in predicting students' final grades, even when compared to modern deep learning architectures. The Random Forest consistently performed best for both regression and classification tasks, achieving an R^2 of 0.91 and classification accuracy of 90%. These results confirm earlier research by Cortez and Silva [1], who also found ensemble tree-based methods to be strong performers for educational datasets with limited size and mixed data types.

The strong influence of variables such as G1 and G2 reinforces the idea that past performance remains the most reliable indicator of future academic outcomes. This finding aligns with Shahiri and Husain [2], who concluded that student performance in early assessments is one of the strongest predictors of final results. However, the addition of behavioral attributes — like alcohol consumption, absences, and study time — provided valuable context, showing that lifestyle habits still contribute meaningfully to academic achievement.

In contrast, the deep-learning models, while capable of learning nonlinear patterns, did not significantly outperform the Random Forest. This is consistent with studies by Hussain et al. [5] and Feng et al. [6], who observed that neural networks often require much larger datasets to fully demonstrate their potential. In small datasets such as this one, the higher computational demand of deep models does not necessarily translate into better predictive accuracy.

Furthermore, interpretability remains a key factor in educational settings. Teachers and administrators often need clear explanations of why a student might be at risk, which makes Random Forests and Logistic Regression more practical choices compared to black-box neural networks. Feature-importance visualizations from the Random Forest helped highlight actionable variables that schools could monitor such as consistent study habits and attendance to support at-risk students early on.

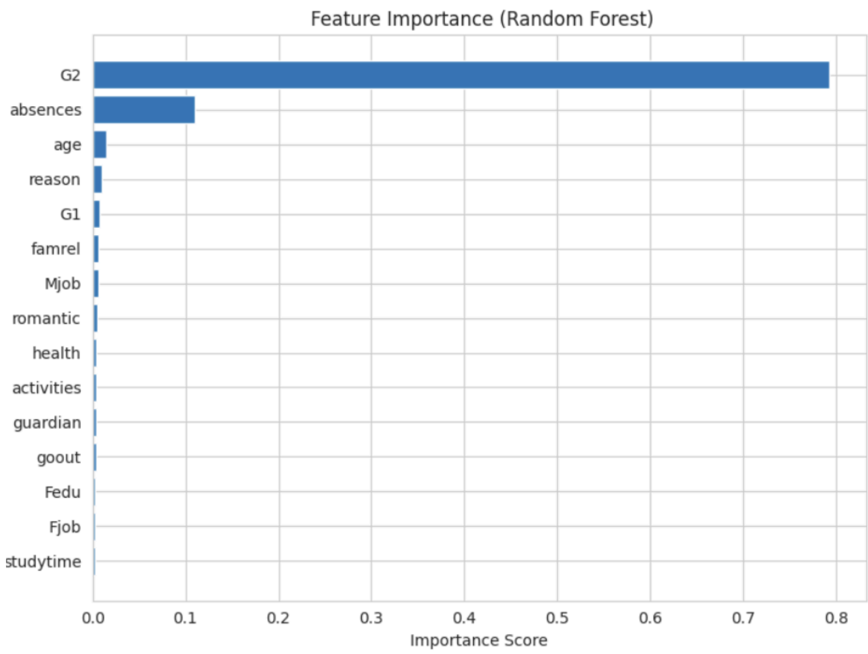


Figure 5. Feature importance plot from the Random Forest Regressor showing the relative influence of each variable on final grade prediction.

`G2`, `G1`, and `studytime` were the strongest positive contributors, while `failures` and alcohol-related variables (`Dalc`, `Walc`) had negative impacts on performance.

Another important observation is the role of preprocessing and feature scaling. Standardizing continuous features improved the convergence and stability of both classical and neural models, echoing Bunkar et al. [9] who emphasized proper data normalization in improving model accuracy. The careful control of overfitting through dropout and early stopping also contributed to the reliability of deep models, ensuring that results reflected genuine learning rather than memorization.

Overall, the project demonstrates that machine learning can be an effective early-warning tool in education. It can help educators identify patterns in student behavior and intervene before failure occurs. However, model predictions should be seen as supportive evidence not absolute judgments since factors such as motivation or personal issues cannot be fully captured by numerical data.

5.2 Limitations

While the project achieved good predictive performance, there were a few limitations:

The dataset was relatively small (395 records), limiting the potential of deep-learning models that rely on large samples to generalize effectively.

The data only represented students from two Portuguese schools, which restricts how broadly the results can be applied to other educational contexts.

Certain variables, such as motivation or teacher quality, were not included in the dataset but likely have a strong influence on learning outcomes.

The model's predictions were based on past data and did not account for temporal changes in behavior over a semester, such as improvement after feedback.

Despite these limitations, the models provided useful insights and strong predictive consistency across experiments.

5.3 Recommendations

Future work could address the current limitations in several ways. First, combining multiple datasets from different institutions would provide larger and more diverse samples, making deep-learning methods more effective. Second, incorporating temporal models such as LSTMs could capture progression over time, improving prediction accuracy for dynamic behaviors. Third, expanding the feature set to include psychological or engagement-related factors such as motivation scores or participation logs could lead to richer, more holistic insights.

Schools could also integrate these models into dashboards that alert teachers when a student shows declining performance patterns. Such tools would not replace human judgment but would complement it, helping educators allocate attention and resources efficiently. Additionally, explainable AI techniques like SHAP or LIME could be integrated to provide even clearer justifications for predictions, increasing trust and usability in real-world classrooms.

5.4 Conclusion

This study applied both classical machine learning and deep learning to predict secondary-school students' academic performance based on demographic, behavioral, and academic attributes. Results revealed that classical methods, particularly the Random Forest, outperformed deep networks in both accuracy and interpretability for this dataset. The findings highlight that for small, structured educational data, simpler models often provide the best balance of performance, transparency, and efficiency.

The project contributes to the broader understanding of how data-driven approaches can enhance educational decision-making. By identifying key predictors of success early in a school term, educators can design targeted interventions to improve outcomes. This supports the long-term goal of personalized learning where each student receives tailored feedback and support based on data insights rather than generic averages.

In summary, while deep learning offers promise for larger datasets, this research shows that interpretable ensemble methods remain highly practical and reliable tools for predicting academic outcomes. The study thus bridges the gap between theoretical machine learning and its practical use in education, demonstrating that data science can empower both students and teachers to achieve better academic success.

References

[1] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, 2010.

[2] P. Cortez and A. Silva, "Using Data Mining to Predict Secondary School Student Performance," *University of Minho, Portugal*, 2008.

[3] A. M. Shahiri and W. Husain, "A Review on Predicting Student Performance Using Data Mining Techniques," *Procedia Computer Science*, vol. 72, pp. 414–422, 2015.

[4] M. A. Al-Barrak and M. Al-Razgan, "Predicting Students' Performance through Classification: A Case Study," *Journal of Theoretical and Applied Information Technology*, vol. 75, no. 2, pp. 167–175, 2015.

[5] M. Hussain et al., "Educational Data Mining and Analysis of Students' Academic Performance Using Deep Learning," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, pp. 232–239, 2019.

[6] M. Feng et al., "Deep Learning for Student Performance Prediction," *IEEE Access*, vol. 8, pp. 67828–67840, 2020.

[7] M. Fei and D. Yeung, "Temporal Models for Student Performance Prediction," *Proceedings of the 2015 IEEE Conference on Educational Data Mining*, pp. 125–132, 2015.

[8] D. Kolo et al., "Data Preprocessing and Feature Selection for Predicting Students' Academic Performance," *Procedia Computer Science*, vol. 163, pp. 306–313, 2019.

[9] K. Bunkar et al., "Improving the Performance of Student Academic Result Prediction through Feature Normalization," *International Journal of Computer Applications*, vol. 41, no. 3, pp. 35–40, 2012.

[10] S. Badr, A. Zualkernan, and F. Alloghani, "Predicting Student Performance Using Machine Learning: A Case Study," *Computers in Education Journal*, vol. 10, no. 3, 2019.

[11] F. Ahmad et al., "Interpretable Machine Learning Models for Early Prediction of Student Dropout," *Computers & Education*, vol. 156, 103962, 2020.

[12] P. Cortez and A. Silva, "Student Performance Data Set," UCI Machine Learning Repository, 2008. [Online].

Available: https://archive.ics.uci.edu/ml/datasets/student+performance