Ollscoil Teicneolaíochta an Oirdheiscirt
South East Technological University

Student Performance Predictor:

A Machine Learning Approach to Identifying At-Risk Students

Name: Abdulaziz Hameed Aloufi

Student ID: C00266252

Module: Data Science & Machine Learning Portfolio

Email: mr.alofi19@gmail.com

GitHub: https://github.com/Abdulaziz1313

Date: 28/11/2025

Contents

1. Introduction

Educational institutions increasingly rely on data to better understand student performance and to identify learners who may be at risk of failing. Early identification of such students can allow lecturers and support staff to provide timely interventions, personalised feedback, and additional resources.

This project aims to develop a data-driven machine learning system that predicts whether a student will pass or fail based on demographic, social, and academic features. Using the UCI Student Performance dataset, the project explores how factors such as study time, absences, parental education and previous academic history influence final outcomes.

The work presented in this portfolio follows the typical data science lifecycle: data understanding and exploratory data analysis (EDA), preprocessing and feature engineering, model selection and evaluation, and model explainability using feature importance and SHAP values.

2. Dataset Description

The dataset used in this project is the Student Performance dataset from the UCI Machine Learning Repository. It contains information on secondary school students from Portuguese schools, including demographic, social and academic attributes, along with final grades.

For this project, the mathematics dataset (student-mat.csv) is used. After loading, the dataset contains 395 rows (students) and 33 original features, plus an additional derived label (pass_fail). The target label is created from the final grade G3: students with G3 >= 10 are labelled as pass (1) and the rest as fail (0). To avoid data leakage, G1, G2 and G3 are dropped from the feature set before training.

The dataset includes demographic and social features (school, sex, age, address, family size, parental education), study-related attributes (studytime, failures, absences), and school-related support indicators. There are no missing values, but there is some imbalance between pass and fail classes, which is considered during model evaluation.

3. Data Preprocessing

Preprocessing steps include:

- Creating the binary target column pass_fail from G3.
- Dropping G1, G2 and G3 to prevent information leakage.
- Encoding categorical variables such as school, sex, address and family size using LabelEncoder

so that models can process them as integers.

- Splitting the data into training and test sets using an 80/20 split with stratification on pass_fail to

preserve the original class distribution.

These steps ensure that the model is trained on clean, numeric data and that evaluation on the test set is fair and representative.

4. Exploratory Data Analysis (EDA)

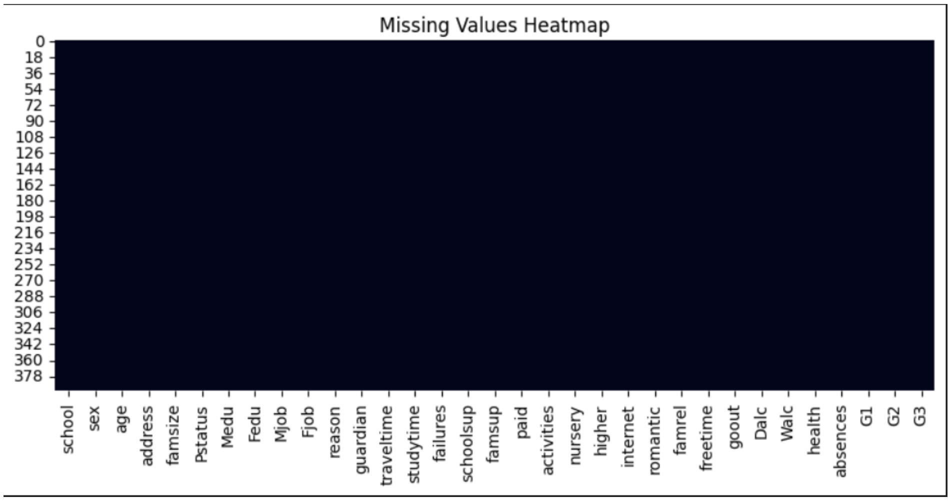


Figure 1

Exploratory Data Analysis was used to understand the structure of the dataset and the main relationships between variables. A first step was to check data quality. The missing values heatmap in Figure 1 shows a solid block with no gaps, confirming that there are no missing values in any feature, so no imputation was required.

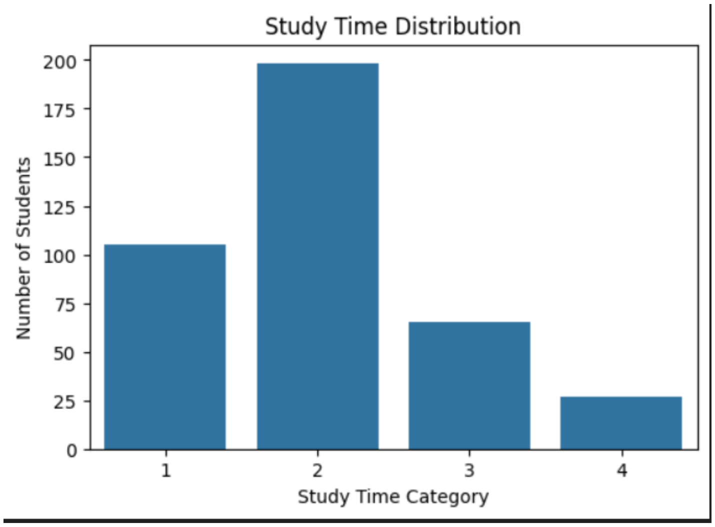


Figure 2

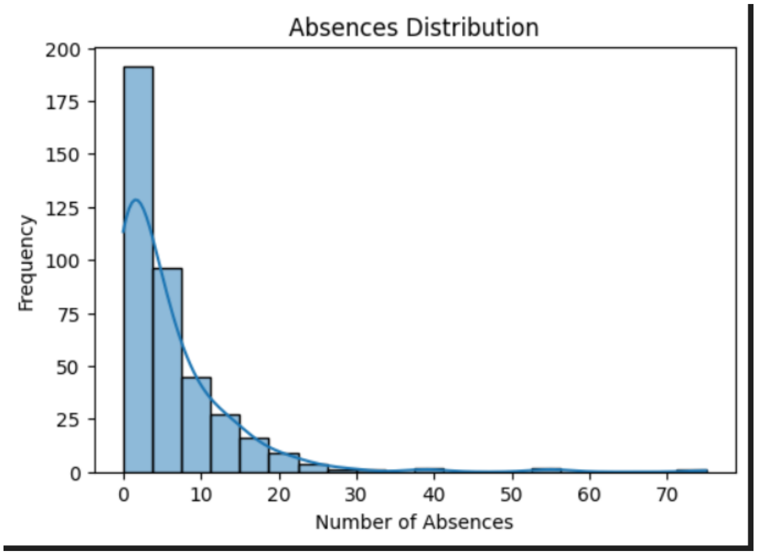


Figure 3

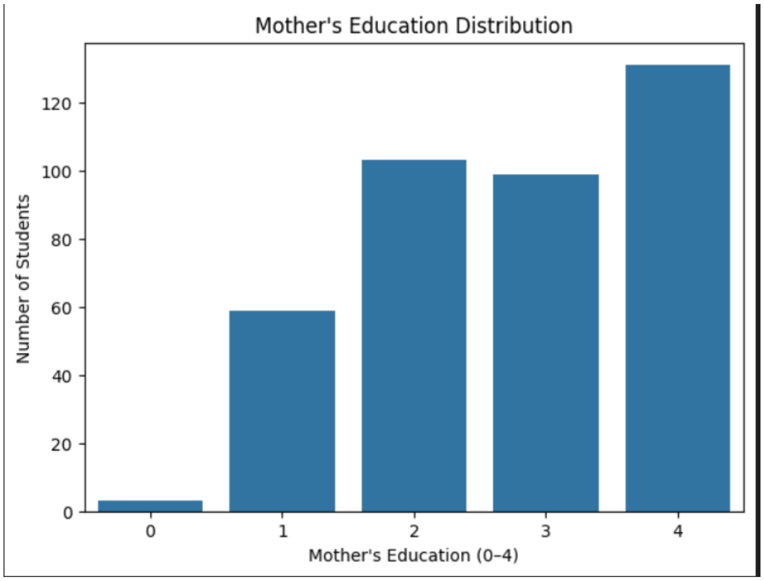


Figure 4

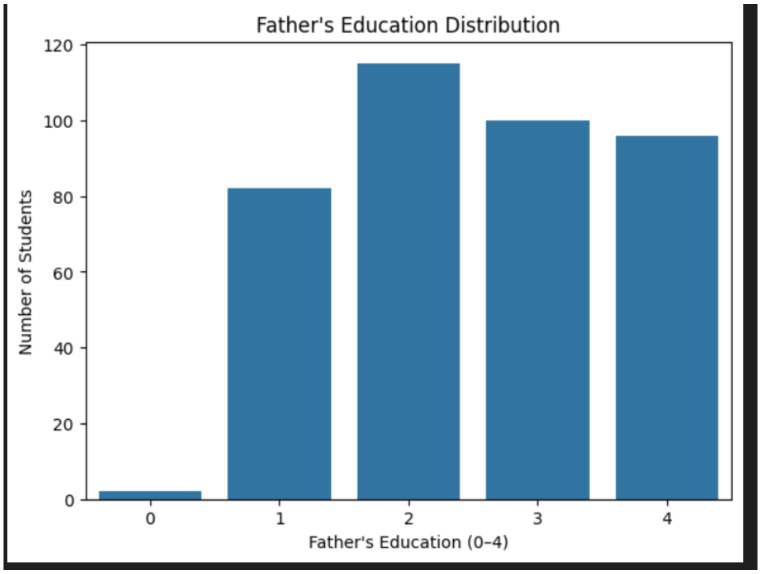


Figure 5

The distribution plots in Figures 2–5 summarise key input variables. The study time distribution in Figure 2 shows that most students report low to moderate study time (categories 1 and 2), while relatively few are in the highest category (4). The absences histogram in Figure 3 is highly right-skewed: most students have only a few absences, but a small group have very high absence counts. Figures 4 and 5 show the distributions of mother's and father's education levels; the majority of parents have education levels 2–4, with very few at level 0, suggesting a generally moderate to high parental education background.

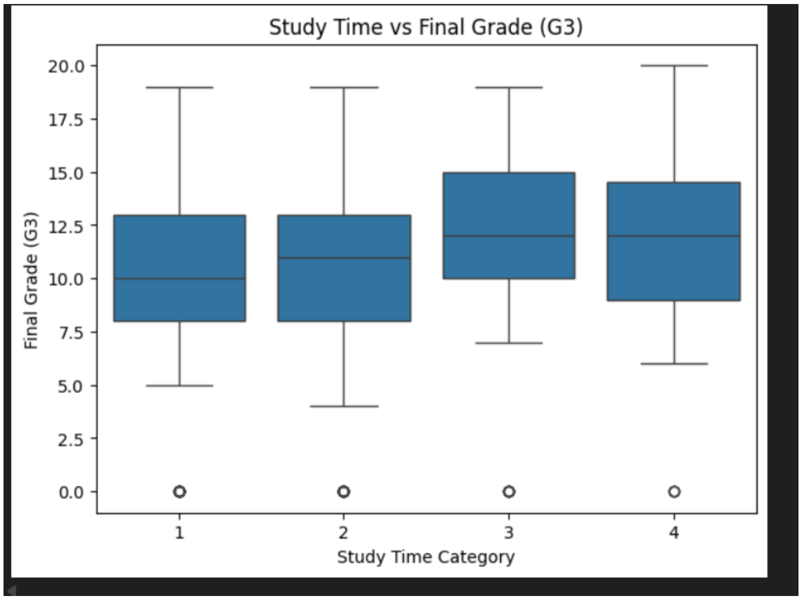


Figure 6

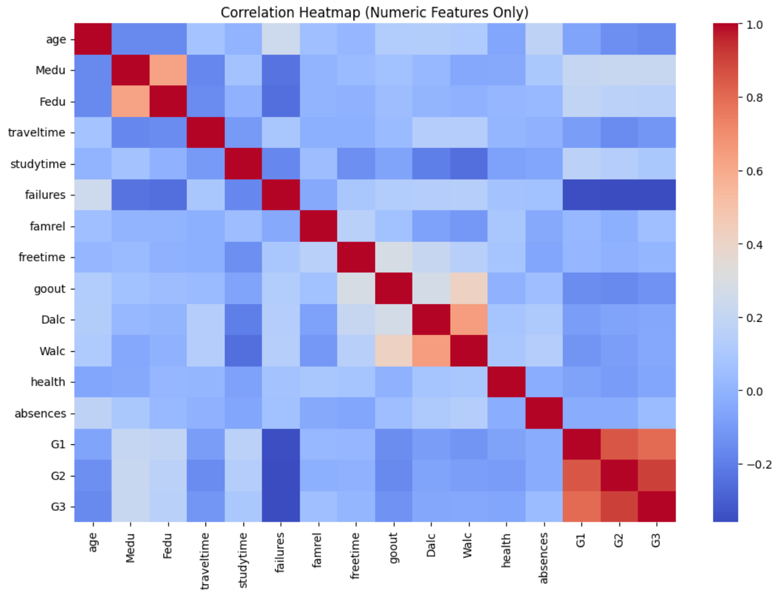
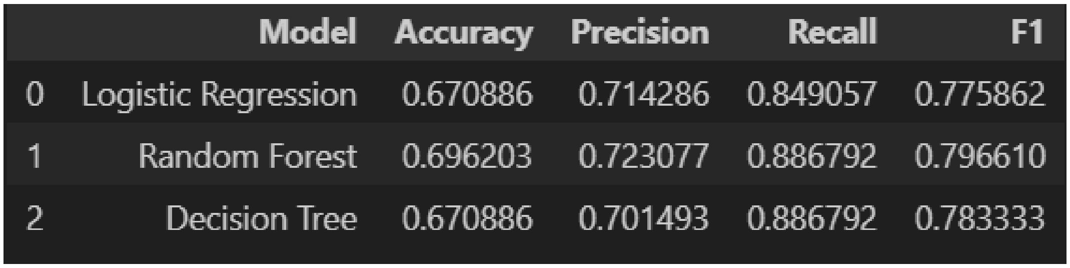


Figure 7

Relationships between features and the final grade are examined in Figure 6, which plots G3 against studytime. Median grades tend to increase with higher study time categories, indicating that students who study more hours per week generally achieve better results. Finally, the correlation heatmap in Figure 7 (computed before dropping G1, G2 and G3 from the feature set) confirms strong positive correlations between the period grades and the final grade, and negative correlations between failures and G3, with a weaker negative relationship between absences and G3.

Overall, the EDA results suggest that absences, past failures, studytime and parental education are important predictors of student performance, and these findings guided the later modelling and explainability steps.

5. Model Development



	Model	Accuracy	Precision	Recall	F1
0	Logistic Regression	0.670886	0.714286	0.849057	0.775862
1	Random Forest	0.696203	0.723077	0.886792	0.796610
2	Decision Tree	0.670886	0.701493	0.886792	0.783333

Table 1

Three supervised learning models are implemented and compared: **Logistic Regression**, **Random Forest**, and a **Decision Tree** classifier. Logistic Regression serves as a simple linear baseline model. Random Forest, an ensemble of decision trees, is well suited to tabular data and provides built-in feature importance scores. The Decision Tree model offers interpretability but is more prone to overfitting than the ensemble.

All models are trained on the same training set created using an 80/20 train–test split with stratification on the pass_fail label. Categorical variables are encoded using LabelEncoder, and the same preprocessed feature matrix is used for all three models. A helper function is implemented to compute **accuracy**, **precision**, **recall**, and **F1-score** for each model, allowing direct comparison of their performance. The resulting metrics are summarised in **Table 1**.

6. Evaluation and Results

Each model is evaluated on the held-out test set using accuracy, precision, recall and F1-score. **Table 1** presents a side-by-side comparison of these metrics for Logistic Regression, Random Forest and the Decision Tree. Overall, the **Random Forest** classifier achieves the best trade-off between accuracy and F1-score, outperforming both the linear Logistic Regression baseline and the single Decision Tree model.

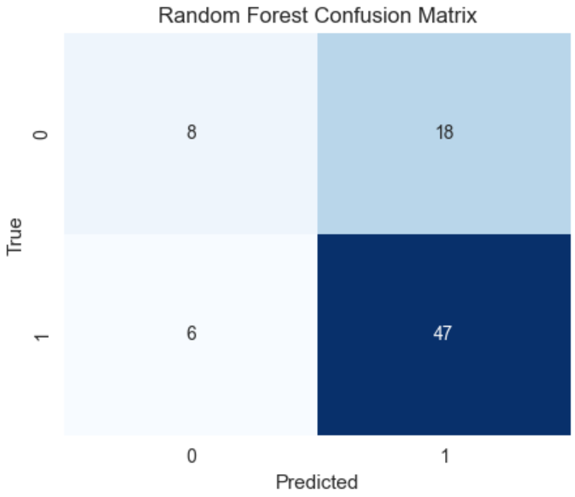


Figure 8

To further analyse performance, **Figure 8** shows the confusion matrix for the Random Forest model. The matrix indicates that the model correctly identifies most students who pass as well as a large proportion of those who fail, with relatively few false negatives (failing students incorrectly predicted as pass), which is important in an early-warning context.

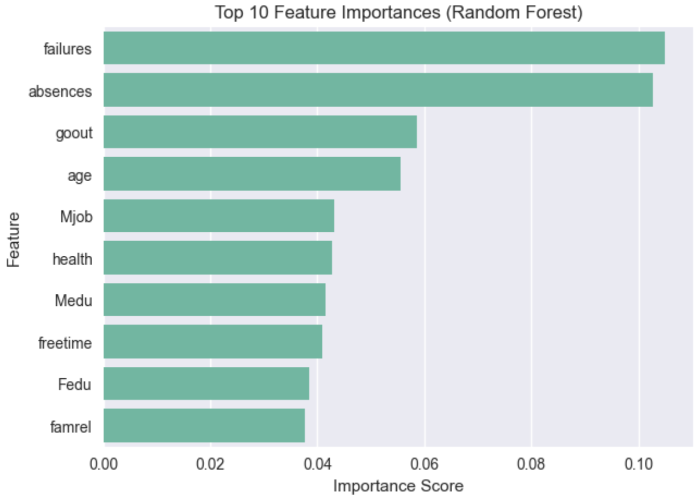


Figure 9

Finally, the Random Forest's impurity-based feature importance scores are plotted in **Figure 9**, which displays the top ten most influential features. Absences and past failures appear as the strongest predictors, followed by studytime and parental education variables. These results are consistent with the earlier EDA findings and provide an initial interpretation of which factors drive the model's decisions.

Based on these results, the Random Forest classifier is selected as the **final model** for the portfolio. It is saved to disk as best_model.pkl using joblib.dump, so that it can be reused later (for example, in a dashboard or for SHAP-based explainability) without retraining.

7. Model Explainability

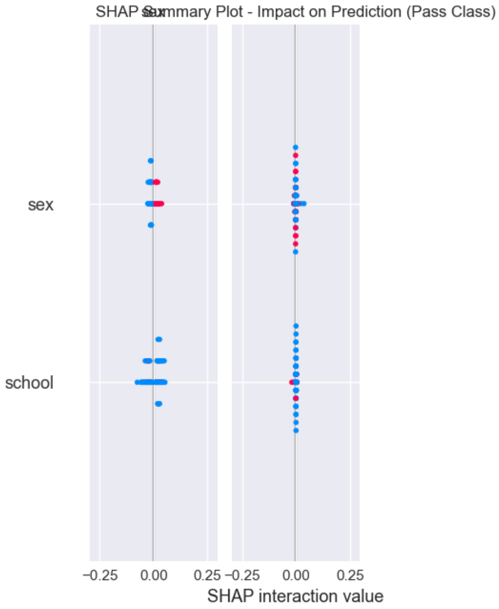


Figure 10

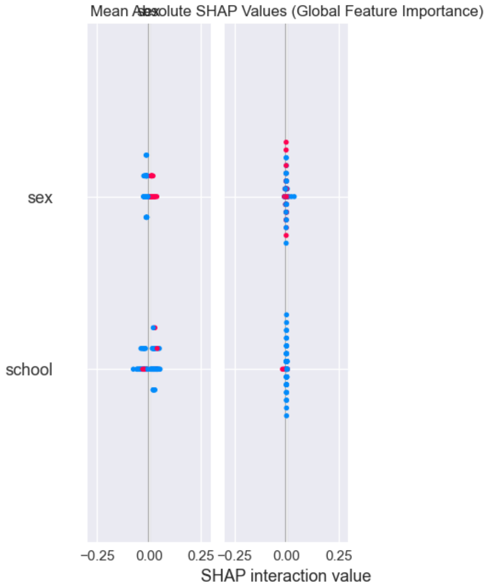


Figure 11

Model explainability is critical in the educational context, because teachers and support staff need to understand **why** a student is predicted to pass or fail. This project uses two complementary approaches to interpret the final Random Forest model:

- **Random Forest feature importance** – The model's feature_importances_ attribute is used to rank predictors globally. The top ten features are shown in **Figure 9**. Absences and past failures have the highest importance scores, followed by studytime and parental education variables. This indicates that these features contribute most to the model's decision-making.
- **SHAP (SHapley Additive exPlanations)** – A TreeExplainer is applied to the trained Random Forest model and SHAP values are computed for a sample of test instances. The **SHAP summary plot** in **Figure 10** shows, for each feature and each student, whether higher values push predictions towards pass or fail. The **mean absolute SHAP values** in **Figure 11** provide a global view of feature impact across the dataset.

These techniques confirm that **higher absences and more past failures strongly increase the likelihood of a fail prediction**, while **greater studytime and higher parental education levels tend to support pass predictions**. The consistency between the Random Forest feature importance plot (Figure 9), the SHAP visualisations (Figures 10 and 11), and the earlier EDA results increases confidence that the model behaves in a reasonable and interpretable way.

8. Discussion and Future Work

The results demonstrate that machine learning can effectively identify students at risk of failing using relatively simple features. However, the dataset is limited to a specific region and school system, and does not capture all possible influences on performance such as teaching quality or socio-economic factors.

Future work could include hyperparameter tuning of the models, exploring additional algorithms such as Gradient Boosting or XGBoost, predicting exact grades using regression models and building an interactive dashboard where staff can input student data and view predictions and explanations in real time.

9. Conclusion

This project presents an end-to-end data science and machine learning pipeline for predicting student performance. It covers data understanding, preprocessing, exploratory analysis, model development, evaluation and explainability.

The Random Forest model provides strong predictive performance and, together with feature importance and SHAP analysis, offers transparent insights into why certain students are at risk. The

work demonstrates practical skills in Python, scikit-learn and SHAP, and highlights how data-driven

approaches can support more informed decision-making in education.

10. References

Dataset (student-mat.csv)

UCI Machine Learning Repository (2014) *Student Performance Data Set*. University of California, Irvine, Machine Learning Repository. Available at: https://archive.ics.uci.edu/dataset/320/student+performance

Main student performance paper (same authors as dataset)

Cortez, P. and Silva, A.M.G. (2008) ‘Using data mining to predict secondary school student performance’, in Brito, A. and Teixeira, J. (eds.) *Proceedings of the 5th Future Business Technology Conference (FUBUTEC 2008)*. Porto, Portugal: EUROSIS, pp. 5–12.

Random Forests algorithm

Breiman, L. (2001) ‘Random forests’, *Machine Learning*, 45(1), pp. 5–32.

scikit-learn (the ML library you used)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É. (2011) ‘Scikit-learn: Machine learning in Python’, *Journal of Machine Learning Research*, 12, pp. 2825–2830.

SHAP (explainability)

Lundberg, S.M. and Lee, S.-I. (2017) ‘A unified approach to interpreting model predictions’, in *Advances in Neural Information Processing Systems (NeurIPS 2017)*, 30, pp. 4765–4774.