

Assignment

2

Classification Models

Fatemeh Elyasifar
Student
2025 AUTUMN

ID:

25589351

36106 - Machine Learning Algorithms and Applications
Master of Data Science and Innovation
University of Technology Sydney

Table of Contents

1.	Business Understanding	2
a.	Business Use Cases	2
b.	Key Objectives	2
2.	Data Understanding	4
a.	Data Sources and Limitations	4
b.	Key Variables and Relevance	4
c.	Exploratory Data Analysis (EDA)	5
3.	Data Preparation	8
a.	Data Cleaning	8
b.	Feature Engineering	8
c.	Data Splitting	9
d.	Transformation	9
4.	Modeling	10
a.	Experiment 1: Support Vector Classifier (SVC)	10
b.	Experiment 2: Decision Tree	10
c.	Experiment 3: Extra Trees	11
d.	Experiment 4: Extra Trees with Feature Improvements	11
5.	Evaluation	12
a.	Results and Analysis	12
b.	Business Impact and Benefits	13
c.	Data Privacy and Ethical Concerns	14
6.	Conclusion	15
7.	References	16

1. Business Understanding

a. Business Use Cases

Academic underperformance remains a pressing issue in many educational systems, which can affect students later in life. A recent OECD report highlights that over 20% of students across member countries fail to meet baseline proficiency in key subjects, limiting future academic and career opportunities. My project aims to build a classification model that predicts student performance categories based on academic, behavioral, and demographic indicators, which assists educators identify students at risk and provide support.

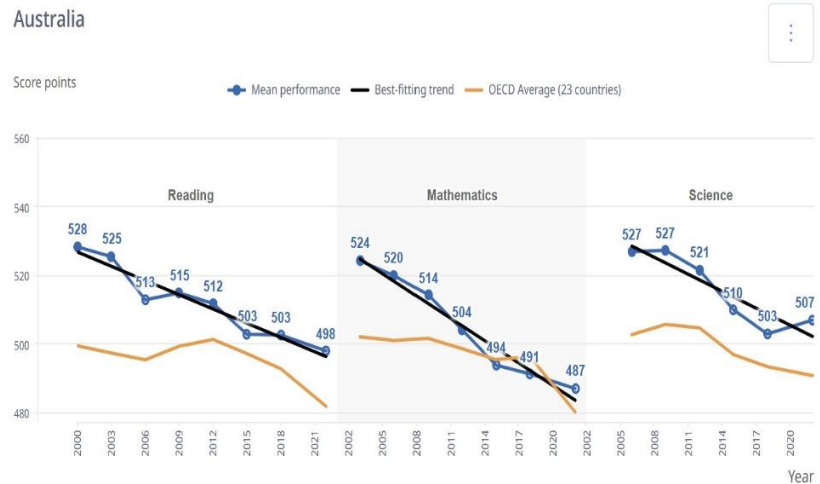


Figure 1. Trends in Australian student performance in Reading, Mathematics, and Science (2000–2022), compared to OECD average.

• Who This Project Helps:


- **Student Support Officers** can identify at-risk students early and provide timely support.
- **Academic Advisors** can guide students with targeted interventions.
- **University Staff** can enhance strategies to improve student outcomes.

b. Key Objectives

The goal of this project is to build a forecasting model to identify at-risk students with poor or average academic performance, helping them avoid failure and reach their full potential.

• Key Objectives:

- Develop a machine learning model that predicts student performance at the end of the semester using academic, behavioral, and demographic indicators. The model should be reliable and generalize well to unseen data.
- Identify critical features that significantly influence student performance.
- Achieve high and interpretable performance metrics, such as F1-score and Recall, with a target of above 80%, ensuring the model is both effective and not overfitting.

- 
- Identify the stakeholders and their requirements.
 - Explain how the project aims to address these requirements.

- **Stakeholders and Requirements**

- **University Staff** use performance insights to strengthen student success initiatives and guide academic planning.
- **Student Support Officers** rely on early risk detection to prioritize outreach and deliver targeted support, helping prevent the misuse of limited resources.
- **Academic Advisors** use predictions to personalize guidance and recommend appropriate learning resources.

By incorporating machine learning algorithms into performance prediction, the project meets stakeholders' requirements by providing interpretable and actionable insights based on current student data.



2. Data Understanding

- The dataset used in this project consists of student information collected from various states in Australia. It was provided in the form of structured CSV files, containing 1,009 rows and 45 columns, including academic, behavioral, and demographic data.

a. Data Sources and Limitations

- The dataset was provided by Anthony as part of the 36106 assignment and appears to have been extracted or adapted from records related to the Australian Department of Education. Given the nature of the dataset, there are several potential limitations:
 - Missing or incorrect values in certain features, which may affect model accuracy.
 - Potential biases in the data, especially in the target variable, make it difficult to deploy the predictive model in real-world settings.
 - Unknown data collection methodology, which makes it difficult to assess how representative or up to date the data truly is.

b. Key Variables and Relevance

- target – The label indicating student performance level, classified into four categories: excellent, good, average, and poor. It is the outcome the model aims to predict.
- previous_gpa – Reflects prior academic performance, strong indicator of future success.
- average_attendance – Higher attendance often correlates with better academic outcomes.
- social_media_hours – May signal distractions; excess use can reduce study focus.
- study_hours – Indicates time spent on learning, generally linked to higher performance.
- completed_credits – Represents academic progress; students with more completed credits may have stronger performance and consistency.

c. Exploratory Data Analysis (EDA)

EDA helped extract patterns and understand data better:

The target variable is highly imbalanced — nearly 50% of students fall under the 'Poor' category, while only 5.5% are labeled 'Excellent'. This imbalance may bias models and reduce recall for minority classes.

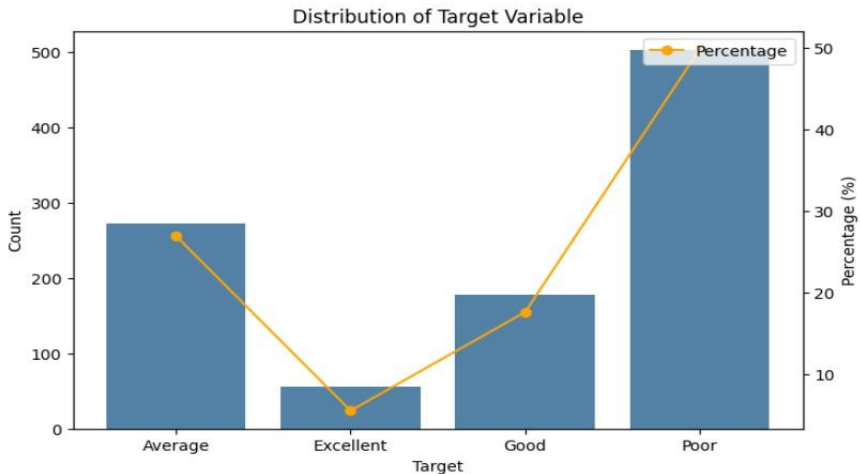


Figure 2. Target Distribution

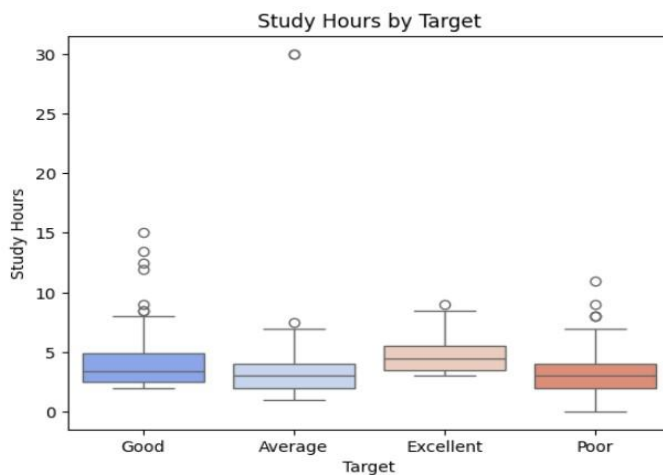


Figure 3. Study Hours by Target

Study hours generally increase with performance level. 'Excellent' and 'Good' students study more, but high variability and outliers suggest study quality or other factors may affect outcomes.

Higher attendance is common among top-performing students. The 'Poor' group showed more variability, with some extremely low attendance. The ceiling effect is noted, with many students reporting 100%.

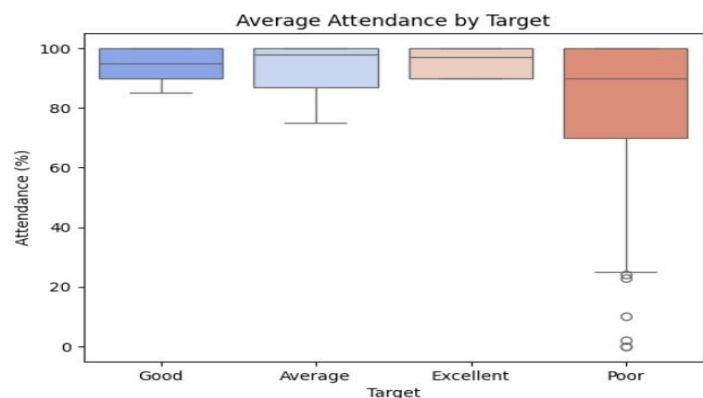


Figure 4. Average Attendance by Target

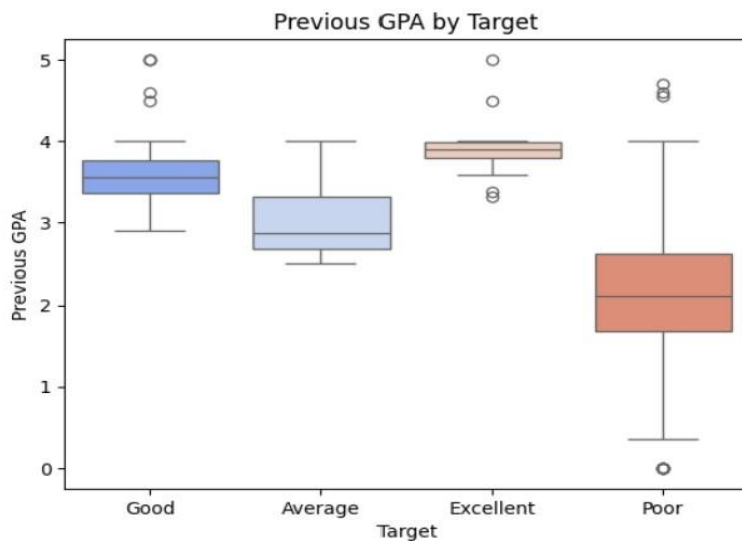


Figure 5. Previous GPA by Target

Previous GPA strongly correlates with performance. 'Excellent' students had higher, consistent GPAs; 'Poor' students showed wide variability. Outliers suggest GPA alone isn't always reliable.

Current GPA is a right-skewed continuous feature, with most students scoring between 3.0 and 4.0. Outliers near 0 may indicate poor performance or data errors.

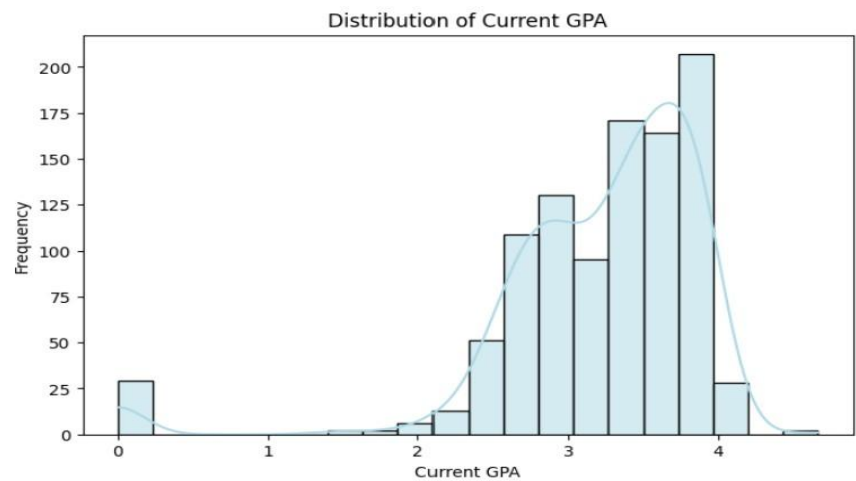


Figure 6. Distribution of Current GPA

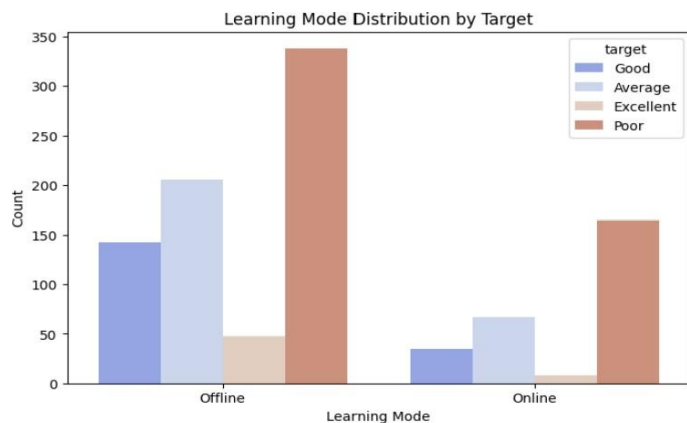


Figure 7. Learning Mode Distribution by Target

The 'learning_mode' feature is a categorical variable with two values: Online and Offline. Most students in the 'Poor' and 'Average' categories are in offline mode, while fewer high-performing students are in online mode, suggesting a possible association between learning mode and academic performance.

Study_Hours: Generally, increases with performance level. 'Excellent' and 'Good' students study more, but high variability and outliers suggest study quality or other factors may affect outcomes.

Social_Media_Hours: Weak negative trend with performance. Higher usage is linked to lower performance, especially among 'Poor' students.

Age: Shows no clear trend with performance or study habits. Age distribution is uneven and discrete, limiting its predictive power.

Completed_Credits: May reflect academic progress and consistency. Students with more completed credits tend to fall in higher performance categories.

Correlation Analysis:

Key positive predictors: previous_gpa (0.69), current_gpa (0.46), average_attendance (0.35). Negative predictor: social_media_hours (-0.40). Weak or irrelevant features (e.g., student_id, postcode) were dropped.

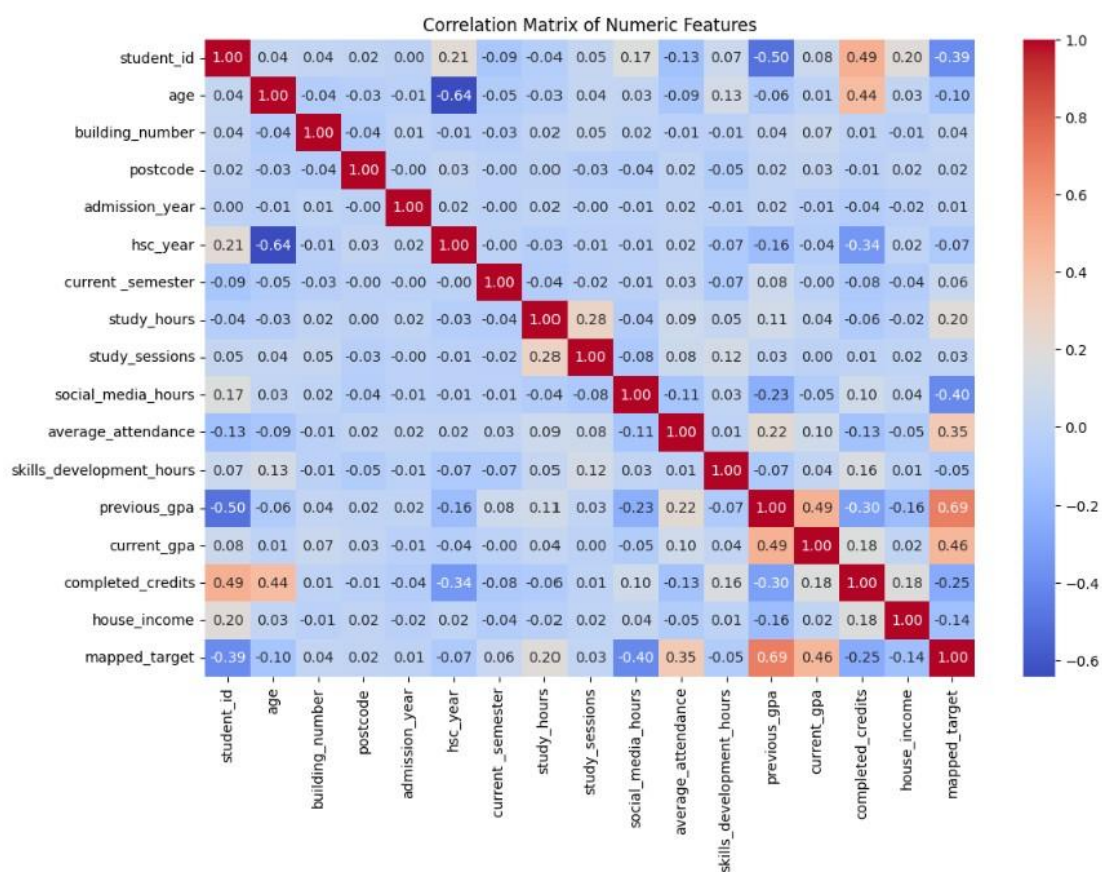


Figure 8. Correlation Matrix

3. Data Preparation

In preparing the dataset for modeling, several key steps were taken to ensure data quality. I implemented a pipeline for data cleaning, feature engineering, and transformation. These steps helped make the data consistent, meaningful, and representative of real-world student performance.

a. Data Cleaning

After narrowing the dataset to a selected subset of relevant features determined during the feature selection process, I addressed three core cleaning tasks:

- **Data Types:** I converted object columns to strings for consistent encoding of categorical data and kept numerical columns as floats to maintain their analytical value.
- **Errors:** I fixed minor inconsistencies to ensure smooth preprocessing. I removed an extra space in the column name 'current _semester' to allow proper referencing. I also replaced an incorrect value (2022.0) in the 'current_semester' column with an estimated value calculated as $(2023 - \text{admission_year}) * 2$, based on the assumption of two semesters per year. Additionally, I corrected a data entry error where 'admission_year' was recorded as 22022.0, likely due to a repeated keystroke. These corrections helped improve data consistency and reduced the risk of calculation errors during modeling.
- **Missing Values and Duplicates:** No duplicate records were found. I filled missing values in 'skills' and 'area_of_interest' with their respective modes to preserve all rows and maintain consistency for encoding and modeling, avoiding potential bias or training errors.

b. Feature Engineering

To improve model performance, I engineered six new features that add meaningful insights and help the model capture key patterns:

- **gpa_change** – Tracks academic improvement or decline by comparing previous and current GPA.
- **academic_gap_years** – Measures years between high school and university; replaces hsc_year.
- **study_efficiency** – Estimates focus per study session, offering more depth than total hours.
- **resource_access_score** – Combines has_phone and has_laptop to reflect tech access.
- **academic_status_level** – Merges probation/suspension status into a single risk indicator.
- **attendance_flag** – Flags students with attendance below 75% as potentially at risk.

c. Data Splitting

Since students are admitted at different times, I chose to split the dataset based on the `admission_year` to reflect a real-world, time-based scenario. By sorting the data chronologically, the model is trained on older student records and evaluated on more recent ones. This strategy prevents data leakage and allows for a realistic assessment of how well the model generalizes to future cohorts. The dataset was split into 70% training (older admissions), 20% validation (more recent), and 10% testing (latest admissions), ensuring evaluation on truly unseen data.

d. Transformation

- **Categorical Features:** I initially used one-hot encoding for `skills` and `area_of_interest` to avoid ordinal bias, but this created 145 features. After testing, I switched to label encoding to reduce dimensionality, as my best model didn't rely on any ordinal relationships.
- **Boolean Features:** I mapped them to 0 and 1 since they only had two distinct values.
- **Scaling:** I used `StandardScaler` to scale numerical inputs to mean 0 and variance 1, which was especially important for scale-sensitive models like Support Vector Classifier (SVC).
- **Handling Imbalanced Target Variable:** After two experiments, I applied Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic examples for underrepresented classes. This improved class balance and helped the model better capture patterns in the minority categories like Excellent.



4. Modeling

For this project, I applied several machine learning algorithms to predict student performance categories, including SVC, Decision Tree, and Extra Trees. These models were chosen for their strengths in classification tasks. My primary focus was on maximizing the F1-score (balancing precision and recall) and recall on the validation set, while ensuring model interpretability and generalizability. High recall was particularly important to identify as many students in need of support as possible.

a. Experiment 1: Support Vector Classifier (SVC)

In the first experiment, I used SVC with grid search to tune C, kernel, and gamma. The best configuration (C = 0.1, class_weight = 'balanced', gamma = 'scale', and kernel = 'linear') yielded strong F1 and recall scores on the validation set. I also applied cross-validation to assess generalizability. While SVC performed well, it fell short of achieving the target score of 80 defined in the business objective. Therefore, I moved on to a non-parametric model like Decision Tree to better capture non-linear patterns in the data.

b. Experiment 2: Decision Tree

In the second experiment, I used the Decision Tree classifier to model student performance, tuning class_weight, criterion, max_depth, and min_samples_split via grid search. The best setup (class_weight = 'balanced', criterion = 'entropy', max_depth = 5, and min_samples_split = 20) produced reasonable F1 and recall scores on the validation set. However, the unusually strong performance raised concerns about overfitting or possible data leakage. I also tested the best grid search setup and observed similarly high results, which led me to explore advanced sampling techniques like SMOTE and experiment with ensemble models like Extra Trees to improve generalization and reduce overfitting.

	mean_test_score	std_test_score	params
0	0.953042	0.025906	{'class_weight': 'balanced', 'criterion': 'gini', 'max_depth': 10, 'min_samples_split': 2}
1	0.953042	0.025906	{'class_weight': 'balanced', 'criterion': 'gini', 'max_depth': 15, 'min_samples_split': 2}
2	0.953042	0.025906	{'class_weight': 'balanced', 'criterion': 'gini', 'max_depth': 20, 'min_samples_split': 2}
3	0.950629	0.026646	{'class_weight': 'balanced', 'criterion': 'entropy', 'max_depth': 10, 'min_samples_split': 5}

Figure 9. Best Grid Search Result

27	0.909341	0.027225	{'class_weight': 'balanced', 'criterion': 'entropy', 'max_depth': 5, 'min_samples_split': 20}
28	0.876274	0.036238	{'class_weight': 'balanced', 'criterion': 'gini', 'max_depth': 5, 'min_samples_split': 5}
29	0.875974	0.036194	{'class_weight': 'balanced', 'criterion': 'gini', 'max_depth': 5, 'min_samples_split': 2}
30	0.874965	0.036876	{'class_weight': 'balanced', 'criterion': 'gini', 'max_depth': 5, 'min_samples_split': 10}
31	0.867134	0.037043	{'class_weight': 'balanced', 'criterion': 'gini', 'max_depth': 5, 'min_samples_split': 20}

Figure 10. Selected Hyperparameters with F1 Score of 91

c. Experiment 3: Extra Trees

To address class imbalance, I applied SMOTE, which improved recall for minority classes like *Excellent*. I then used the Extra Trees classifier and tuned hyperparameters using grid search. The best configuration (class_weight = 'balanced', criterion = 'entropy', max_depth = 20, min_samples_leaf = 2, and n_estimators = 200) achieved strong validation performance and better generalization than the Decision Tree model. However, the results were similar to those of SVC, so I created additional features and switched to label encoding to reduce dimensionality and reapply the best-performing model.

d. Experiment 4: Extra Trees with Feature Improvements

After reducing features from 145 to 32 through feature engineering and label encoding, I reapplied the Extra Trees classifier using the configuration class_weight = 'balanced', criterion = 'entropy', max_depth = 20, min_samples_leaf = 2, and n_estimators = 200. This setup, combined with the improved feature set, delivered more reliable and interpretable results, successfully meeting the target score with strong F1 and recall scores and better generalization to unseen data. The reduced dimensionality also improved training efficiency without sacrificing performance.

	mean_test_score	std_test_score	params
0	0.939553	0.018822	{'class_weight': 'balanced', 'criterion': 'entropy', 'max_depth': 20, 'min_samples_leaf': 2, 'n_estimators': 200}
1	0.937455	0.016301	{'class_weight': 'balanced', 'criterion': 'entropy', 'max_depth': 20, 'min_samples_leaf': 2, 'n_estimators': 100}
2	0.936600	0.019466	{'class_weight': 'balanced', 'criterion': 'gini', 'max_depth': 20, 'min_samples_leaf': 2, 'n_estimators': 200}
3	0.932398	0.019456	{'class_weight': 'balanced', 'criterion': 'gini', 'max_depth': 20, 'min_samples_leaf': 2, 'n_estimators': 100}
4	0.919060	0.019992	{'class_weight': 'balanced', 'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 2, 'n_estimators': 200}
5	0.917413	0.022251	{'class_weight': 'balanced', 'criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 2, 'n_estimators': 100}

Figure 11. Best Hyperparameters for Extra Trees



5. Evaluation

a. Results and Analysis

The performance of each model was evaluated using F1 and recall scores, with a focus on accurately and consistently identifying students with poor and average performance to support early intervention.

Model	F1 Score (Validation)	Recall (Validation)	F1 Score (Test)	Recall (Test)
DummyClassifier (Baseline)	0.36	0.37	-	-
SVC	0.73	0.72	0.72	0.72
Decision Tree	0.94	0.94	0.94	0.94
Extra Trees	0.81	0.81	0.82	0.81

- Baseline Model (Dummy Classifier): The F1 score of 36 and recall of 37 set a low benchmark.
- SVC: The model performed moderately across validation and test sets, suggesting reasonable generalization. However, the F1 and recall scores did not meet the project's threshold of 80, highlighting its limitations in capturing complex non-linear patterns.
- Decision Tree: With the best hyperparameters, the model achieved an unusually high score of 94, well above the project goal of 80. However, this raises concerns about overfitting. Despite outperforming SVC, the model may struggle to generalize to unseen data, making it unsuitable for real-world deployment.
- Extra Trees: This model delivered the most reliable and balanced performance, matching the target score of 80. It also achieved F1 and recall scores of 81 on the test set, indicating strong generalizability. Feature importance analysis further confirmed the impact of age, along with other key predictors.

Extra Trees emerged as the best model, meeting the target F1 and recall scores while demonstrating high predictive accuracy on unseen data.

b. Business Impact and Benefits

- Extra Trees had the strongest impact in addressing academic underperformance: With F1 and recall scores of 0.82 and 0.81 on test data, the model meets project goals and reliably identifies students at risk.
- It enables university staff to intervene early, especially for 'Poor' and 'Average' performers (recall: 0.88 and 0.67), reducing the chance of long-term academic failure.
- Support officers can use the model's insights to allocate support resources more effectively.
- Academic Advisors use predictions to recommend learning resources.
- Misclassifications are minimal and mostly occur between adjacent categories, limiting negative impacts.

Compared to other models, Extra Trees's performance enables better alignment with goals and business use cases. However, this model was trained on a comparatively narrow range of students, focusing only on 1,408 students (with some synthetic data generated through SMOTE) across Australia. As such, this should be seen as an initial proof of concept rather than a production-ready solution. Broader data collection, a more current dataset, and further testing will be necessary to ensure its effectiveness on a scale.

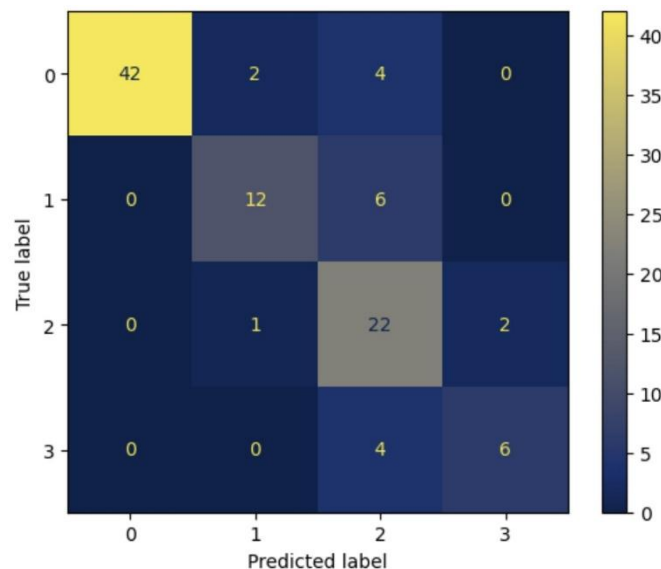


Figure 12. True Labels vs. Predicted Labels



c. Data Privacy and Ethical Concerns

- **Data Privacy:** This project involved student data containing personal identifiers such as email addresses, phone numbers, and residential addresses. To ensure privacy, all identifiable information was removed in line with ethical and data protection standards.

Ethical Considerations:

- **Imbalanced Target Variable:** The dataset had class imbalance, leading to potential bias and unequal accuracy. SMOTE was used to improve recall for minority classes, though synthetic data may not fully reflect real-world complexity.
- **Equity and Generalization Risks:** The model was trained on a limited dataset of 1,408 students, which may reduce generalizability and fairness, especially for underrepresented groups. This limitation could lead to misclassifications, potentially impacting student motivation and academic development. Future work should incorporate broader, more current datasets and include fairness assessments across all groups.



6. Conclusion

- This project aimed to develop a predictive model to identify student performance in the upcoming semester using both parametric and non-parametric classification techniques. Among the models evaluated, Extra Trees proved to be the most effective, achieving a test F1 score of 82 and a recall of 81, successfully meeting the project objective of maintaining scores around 80. These results demonstrate the model's ability to provide consistent and accurate predictions within university settings.
- Key insights include:
 - SVC performed reliably but was limited in capturing complex non-linear relationships.
 - The Decision Tree model showed signs of overfitting, highlighting the need for ensemble methods like Extra Trees to improve generalization.
 - Extra Trees effectively handled data complexity and delivered strong performance without overfitting.
 - GPA, attendance, and study hours method were identified as key predictors of student performance.
 - Misclassifications were minimal and mostly occurred between adjacent categories, helping to reduce negative impact.
- The project successfully addressed the stakeholders' requirements by producing an interpretable and generalizable prototype model focused on identifying at-risk student performance for early academic intervention. However, the model was trained on a relatively narrow dataset of 1,408 students and did not account for broader educational contexts or diverse student populations. For real-world deployment, the next steps should include:
 - **Expand the dataset** to include more students, institutions, and recent data for improved generalizability.
 - **Testing the model across different academic terms or years** to ensure robustness under changing educational trends and student behaviors.
 - **Exploring additional ensemble methods** to further enhance predictive performance

This exploratory model serves as a strong foundation for building data-driven decision-support tools for the education sector; however, it requires the above considerations before deployment to confirm the model's reliability and ensure its effectiveness in real-world use.



7. References

- Anthony. (2025). *students_performance.csv* [Dataset]. The dataset provided by the professor.
- OECD. (2023, December 5). *PISA 2022 Results (Volume I and II) – Country Notes: Australia*. OECD Publishing. https://www.oecd.org/en/publications/pisa-2022-results-volume-i-and-ii-country-notes_ed6fbcc5-en/australia_e9346d47-en.html

