Assignment 2Classification Models

short line

Md Saifur Rahman

Student ID: 25528668  
2025 AUTUMN

36106 - Machine Learning Algorithms and Applications

Master of Data Science and Innovation

University of Technology of Sydney

Table of Contents

[**1. Business Understanding 2**](#_heading=h.u53ijpbss21r)

[a. Business Use Cases 2](#_heading=h.n2l5nreafh3y)

[b. Key Objectives 2](#_heading=h.r6euwo5r111y)

[**2. Data Understanding 4**](#_heading=h.rmfjc17rj0aq)

[**3. Data Preparation 7**](#_heading=h.dhoxq9gt6avo)

[**4. Modeling 10**](#_heading=h.jvjuvpo8whn4)

[**5. Evaluation 14**](#_heading=h.e9qhitwcm49x)

[a. Results and Analysis 14](#_heading=h.josmsqidrgau)

[Model-by-Model Analysis 15](#_heading=h.wbsn8y97f2go)

[Dummy Baseline Classifier 15](#_heading=h.n6dxpgrejwfd)

[K-Nearest Neighbors Classifier 16](#_heading=h.e8zeogxcgim6)

[Logistic Regression 17](#_heading=h.3xnq8hneo9ry)

[Random Forest Classifier 18](#_heading=h.ey45cbl1hac)

[XGBoost Classifier 20](#_heading=h.eto8veanabib)

[Performance Trend Summary 21](#_heading=h.zbvzs84bs511)

[b. Business Impact and Benefits 21](#_heading=h.tdzitgcpkmu4)

[c. Data Privacy and Ethical Concerns 23](#_heading=h.3uzg7ksd4xjw)

[d. Comparison and Key Findings from Experimentation 26](#_heading=h.1yrubfpis0wa)

[**6. Conclusion 29**](#_heading=h.yseu733i5rn)

[**7. References 32**](#_heading=h.wq6jsqstg062)

# Business Understanding

## Business Use Cases

Set in a university setting, the project's goal is to classify students' academic achievement as Excellent, Good, Average, or Poor prior to the results. Many applications might be made possible in this situation with a trustworthy categorization model. To begin, kids who are at danger of performing at a "Poor" or "Average" level can be identified early on by academic support personnel, who can then assist through mentorship, counseling, or tutoring. Student achievement and retention rates might both benefit from this kind of preventative assistance. Second, the model may single out pupils who have excelled academically ("Excellent") and recommend them to leadership or scholarship programs. To sum up, the institution will be able to better allocate resources with the use of reliable predictive analytics, allowing them to target remedial efforts where they are most needed and to cultivate top talent. This, in turn, will improve student performance and retention rates.

The complexity of student achievement and the difficulty of teachers manually identifying which children may be struggling are two of the main obstacles that prompted this study. It is difficult to depend on basic principles or intuition alone because of the many variables that affect results (attendance, behavior, academic history, etc.). There are repercussions when students are misidentified. For example, if a struggling student is not caught in time, they may fail or drop out of school. On the other hand, if a competent student is incorrectly labeled as at-risk, support resources may be wasted or their self-esteem could be damaged. To produce better predictions than heuristic methods, machine learning may learn complicated patterns from past student data. By making use of data that would be too massive for people to manually study, this data-driven strategy tackles the chance to enhance student results.

## Key Objectives

Creating a machine learning model that fairly divides student performance into the four categories is the main objective. High predictive accuracy ( ideally well above 80% on hold-out data) along with strong precision and recall for the minority classes (particularly "Excellent" and "Poor") define a major success criterion. This process guarantees not only general accuracy but also consistently identifies the really exceptional students and those in need of assistance. The project also seeks to offer practical insights—for instance, stressing which elements most significantly affect performance—so enabling stakeholders to trust and make good use of the forecasts of the model.

Among the stakeholders are university academic advisers and support personnel, faculty and administrators, and the students themselves. Academic advisers and support systems depend on the model to identify early in the semester at-risk students so they may focus interventions—extra tutoring, counseling—on those particular individuals. Improved retention and graduation rates pique the curiosity of university officials; they need proof the model will enable cost-effective resource allocation and enhance important performance indicators (such as lowering the percentage of failing students). Students (and their parents) are indirect stakeholders who gain from timely support or recognition; they thus demand that any predictive system be fair and applied to help rather than punish. The project uses past student data to train a classifier able to forecast performance category for new students, so addressing these needs. The model offers a means to meet stakeholders' needs by meeting accuracy and fairness objectives: advisers get a prioritized list of students to follow up with, administrators see quantifiable improvements in outcomes, and students get support or opportunities matched with their expected needs. The approach and goals remain rather generic and data-driven so that this predictive framework could be modified to fit comparable educational environments or other performance prediction projects.

short dash

# Data Understanding

The dataset used in this study consists of records of university students with various characteristics that capture academic, demographic, and behavioral information. There are 1,009 student records in the dataset, each containing 34 features. There is a record for every pupil. It appears that the data is collected via the university's own systems for student information, which may combine academic records with student surveys. Academic performance metrics, attendance records, and self-reported behaviors are important sources of data. Academic performance metrics include things like grade point averages and credits obtained from the registrar. In addition to demographic information, the raw data contains personally identifiable details. One caveat is that the data seems to be institution-and time-specific, so it might not capture intangibles like personal motivation or the quality of instruction, or it might not have longitudinal effects beyond the present semester. Regardless, it offers a comprehensive overview of the elements linked to students' achievements.

Objective Measurement: A categorical label with four potential values—Excellent, Good, Average, and Poor—defines the goal variable, which is students' performance. These labels show where each student stands in terms of their overall achievement, which can be determined by factors like GPA or academic standing. The distribution of the goal is skewed: the majority of students are classified as "Average," a smaller number are classified as "Good," and the most extreme categories, "Excellent" and "Poor," are the smallest. The bulk of the 1,009 students fell into the "Average" class, which was clearly a skewed distribution in the target's frequency plot. A model could be skewed toward forecasting the majority if this class imbalance isn't taken into consideration. Stratified sampling and subsequent methods, such as class weighting in models, were employed to guarantee that minority classes learned appropriately due to the unequal distribution of targets.

**Features:** The dataset includes a mix of numerical and categorical features that are believed to influence academic performance. Notable features are

* Academic history includes completed credits, prior and current grade point averages, and average attendance (percentage attendance). Students with better past GPAs and full class attendance naturally should perform better. Indeed, exploratory study revealed that whereas poor performers had much lower values in these areas, excellent students generally had higher GPAs and attendance than others.
* Study\_hours (self-reported study time per week), study\_sessions (number of study sessions), skills\_development\_hours (time spent in skill development activities) and social\_media\_hours (time spent on social media). These catch time management and way of life. From a group analysis, we found, for example, that "excellent" students typically studied more hours—about 4.8 hours/day on average—than "average" students—about 3.1 hours/day. While more study sessions and skill development usually correlate positively, high social media hours could negatively correlate with performance.
* Gender, scholarship (whether the student is on a scholarship), birth\_country (country of origin, which can be a proxy for domestic/international student), and disabilities (whether the student has any disability). These characteristics must be handled carefully; some may correlate with performance due to systemic elements (such as scholarship students might have higher prior achievement). They do, however, also bring ethical questions (we have to avoid biassed projections).
* Administrative status features: e.g., on\_probation (if the student has academic probation status), is\_suspended (if ever suspended), and has consulted a teacher (whether the student sought help from teachers). These are signs of academic involvement or problems; for example, historically poor current performance usually corresponds with probationary status.
* Environmental characteristics include learning mode—online or in-person, university transportation (commute), and living arrangement—on-campus, off-campus, with family, etc. Though our study did not find these to be main drivers, these might have more subdued effects; for example, students living on campus may have easier access to resources and so better outcomes.
* Some new features were built during exploration to record interactions: gpa\_change, the difference between current and previous GPA, indicating either improvement or decline; study\_x\_attendance, an interaction term multiplying study hours by attendance rate; and skill\_to-\_social\_ratio, the ratio of skills development hours to social media hours. These designed elements were meant to highlight non-linear interactions—that is, whether balancing skill development against social media or high study time and high attendance improves performance. They seek to provide the model extra signals not directly present in individual raw features.

Some interesting trends were found by first exploratory data analysis (EDA). Features including current GPA, past GPA, and attendance clearly show positive correlations with the performance categories: students classified Excellent had greatly higher values in these measures than those classified Poor. As would be expected, some categorical traits—like being on probation or suspended—are more common in the poor category. We also looked over outliers and missing data. The dataset was shockingly complete; missing values for important fields including the target and key features were zero. A few numerical fields showed anomalies; for example, house\_income (family income, if given) varied greatly with a few very high values, and social\_media\_hours had a small number of students reporting extraordinarily high usage. Although some models like trees are robust to outliers and we scaled others to minimize extreme values, these outliers were observed as possible noise and we decided to address them within the modeling step. The information given presented a whole picture of every student, and our knowledge of it helped shape the decisions on data preparation and modeling that are next discussed.

short dash

# Data Preparation

Before modeling, extensive data preparation steps were carried out to ensure data quality and optimize the dataset for machine learning algorithms. This process included data cleaning, feature selection and engineering, encoding and scaling, handling class imbalance, and splitting the dataset appropriately. The major steps are outlined below.

## **1. Data Cleaning**

The first step was to clean the dataset by removing any sensitive or superfluous data. We removed some fields that included PII (personal information) in order to safeguard student privacy and remove unnecessary entries. These columns were full\_name, email, phone\_number, secondary\_address, building\_number, and postcode. We also removed values that did not provide useful performance insights, such as student\_id (a pure identifier), birth\_country (a very high-level or low-variance property), and program (a very high-level or low-variance attribute).

Simplifying the process meant keeping only the academic and behavioral characteristics that were useful for making performance predictions. Following this stage, the dataset shifted its attention to quantifiable and interpretable variables including grade point averages, class participation, study habits, and extracurricular skill development.

## **2. Missing Values and Outlier Handling**

Fortunately, after cleaning, there were no missing values in important academic or engagement aspects; hence, imputation techniques were not needed. But in columns like social\_media\_hours and skills\_development\_hours, we found a few very unusual values—outliers. We chose a strong modeling technique rather than deleting or capping these outliers. Tree-based models such as Random Forest can efficiently manage outliers by means of recursive splitting; for sensitive models such as KNN or Logistic Regression, feature scaling reduces excessive impact from wide numerical ranges. We thus kept all student records since we knew that extreme behaviors—like too much use of social media—could be useful for classification.

## **3. Feature Engineering and Feature Selection**

We developed several new features to increase model performance and capture more profound interactions:

gpa\_change: Current GPA less past GPA.

study\_x\_attendance: Interaction term spanning attendance rate between study hours.

Skill to Social Ratio: Hour of skill development divided by hours of social media use.

Using a mix of techniques—mutual information scores, ANOVA F-tests, Lasso regularization coefficients, and Random Forest feature importance—feature selection Key variables regularly found to be quite predictive were average\_attendance, previous\_gpa, and current\_gpa. Retained for their help in model interpretability were behavioral elements including study\_hours and skills\_development\_hours.

To lower noise and streamline the model inputs, low predictive power features including relationship\_status, has\_phone, and health\_issues were eliminated.

## **4. Feature Encoding**

We developed several new features to increase model performance and capture more profound interactions:

gpa\_change: Current GPA less past GPA.

study\_x\_attendance: Interaction term spanning attendance rate between study hours.

Skill to Social Ratio: Hour of skill development divided by hours of social media use.

Using a mix of techniques—mutual information scores, ANOVA F-tests, Lasso regularization coefficients, and Random Forest feature importance—feature selection Key variables regularly found to be quite predictive were average\_attendance, previous\_gpa, and current\_gpa. Retained for their help in model interpretability were behavioral elements including study\_hours and skills\_development\_hours.

To lower noise and streamline the model inputs, low predictive power features including relationship\_status, has\_phone, and health\_issues were eliminated.

## **5. Feature Scaling**

All numerical features were standardized using z-score scaling (mean of 0, standard deviation of 1), given some models—notably K-Nearest Neighbors and Logistic Regression—are sensitive to feature magnitude.

To prevent data leakage, scaling was fitted just on the training set and subsequently implemented on validation and test sets.

While maintaining a consistent pipeline across models guaranteed compatibility and flexibility, tree-based models do not call for scaling.

Features scaled included study hours, social media hours, average attendance, and house income.

## **6. Class Imbalance Handling**

Fewer pupils were rated as "Poor" or "Excellent" in our dataset than as "Average" or "Good." This suggests a class imbalance. In order to solve this, we divided the training set and then used SMOTE, which stands for Synthetic Minority Over-sampling Technique. By utilizing closest neighbor interpolation, SMOTE artificially produced new samples for minority classes, therefore balancing the training data without adding simple duplications.

The capacity of the models to learn patterns across all performance categories was improved after SMOTE because the training data had about equal representation of all four classes (Poor, Average, Good, and Excellent) (~25%).

## **7. Data Splitting**

Using stratified random splitting, we last divided the data into training, validation, and test sets so preserving the class distributions across all subsets.

The split ratio applied was roughly 70% for training,

15% for validation;

15% for tests.

Random stratified splitting guarantees that performance assessment on the validation and test sets reflects the actual balance of classes and lowers risks of biassed model selection.

Before SMote, splitting was done to prevent any contamination between synthetic samples and test set real data.

short dash

# Modeling

A systematic modeling approach was implemented prior to constructing predictive models. Beginning with a basic baseline model and working our way up to increasingly complex ensemble models, we carefully tuned and validated performance at each step to achieve our primary objective of increasing model complexity. In all, we tried out five different models: XGBoost, Logistic Regression, DummyClassifier, and K-Nearest Neighbors (KNN). Consistent with previous work, all models were built using the scikit-learn and XGBoost packages.

Using a stratified random method, we trained a DummyClassifier to serve as a baseline.

Without learning any patterns, this classifier produced random predictions based on the observed class distribution in the training data.

The Dummy model nearly matched the percentage of the biggest class ("Average") with its relatively poor accuracy of about 30% on the test set, as was predicted.

Its F1-score, weighted precision, and recall were all below average.

Predictions were evenly distributed throughout classes, according to the confusion matrix, and there was minimal capacity to capture minority classifications like "Excellent" or "Poor."

The dummy model was used as a reference point to prove that any actual model has to be far better than just guessing.

Classifier Based on K-Nearest Neighbors (KNN): Following the baseline, we trained a K-Nearest Neighbors (KNN) model with initially 5 neighbors.

Because of its non-parametric character and relative ease of use, KNN was chosen for the task of capturing both local patterns and non-linear class borders.

A significant improvement over the dummy classifier, the model attained 76% accuracy on the training set and 73% accuracy on the test set.

The F1-score, recall, and precision were all well-balanced, falling within the range of 0.73-0.77.

The goal of the early real model known as KNN was to identify patterns, such as groups of students with comparable academic performance, study habits, and attendance.

After that, we used L2 regularization (Ridge) to construct a Logistic Regression model that would prevent overfitting.

The regularization strength C, which is the inverse of λ, was optimized by hyperparameter tweaking using GridSearchCV with 5-fold cross-validation.

With a weighted F1-score of0.44, Logistic Regression still only managed 48% accuracy on the test set, even after adjusting.

Among minority groups, accuracy and memory were poor.

According to the confusion matrix, Logistic Regression performed well when predicting the "Average" class but had difficulty when dealing with minority classifications.

Logistic Regression was used as a linear baseline to show that this job required non-linear models that capture interactions, since mere linear separability was not enough.

**Random Forest Classifier**: The **Random Forest** model was introduced next, using an ensemble of **100 trees** with tuned hyperparameters:

* Limited maximum tree depth (e.g., max\_depth=6)
* Minimum samples per split and leaf (e.g., min\_samples\_split=10, min\_samples\_leaf=5)
* Random feature selection at splits (max\_features='sqrt')
* Class weight balancing (class\_weight='balanced').

Random Forest captured complicated feature interactions and achieved minimal overfitting with 96% training accuracy and 90% test accuracy.

On the test set, precision, recall, and F1-score were all rather close at 0.90.

Leveraging feature interactions like GPA × attendance and handling class imbalance via automatic sample weighting, Random Forest was the first strong ensemble model.

**XGBoost Classifier**:

Finally, we trained an **XGBoost** classifier to push performance further.  
 Key settings included:

* Conservative **learning rate** (learning\_rate=0.05)
* **Max tree depth** kept low (e.g., max\_depth=2)
* Early stopping strategy based on validation performance (stopping after 10 rounds without improvement).

With 98% training accuracy and 93.5% test accuracy—matching or somewhat surpassing Random Forest performance—XGBoost

Precision, recall, and F1-score all came in at 0.94.

XGBoost was especially good in capturing subtle feature interactions and offered a boosted ensemble approach, fixing mistakes made by previous trees.

In conclusion, our modeling strategy made use of five models: XGBoost, Logistic Regression, KNN, and a Dummy baseline. Using the validation findings to guide hyperparameter tweaking at each level, we started with simple and worked our way up to more sophisticated settings. A reasoning was used to select the parameters and implementation of each model:

DummyClassifier: for initial performance settings (no capacity to learn).

Use K-Nearest Neighbors (KNN) as a first-generation, practical model for easily capturing non-linear patterns.

Logistic Regression: to lay the groundwork for linear modeling and to get insight into feature selection through regularization.

Random Forest is anticipated to significantly enhance accuracy by effectively handling non-linearity and interactions.

XGBoost: to remove any remaining faults from the random forest model and enhance performance to its maximum potential.

To provide a fair comparison, all models were tested using the identical training/validation/test splits. Although the assessment will reveal that the more complicated ensemble models were selected throughout the model selection process, it was crucial to have the simpler models available for comparison in order to comprehend the benefits and drawbacks. Specifically, we used classifier variations and proper multi-class handling throughout the modeling process to make sure the algorithms we outlined were suitable for the classification challenge we had.

short dash

# Evaluation

## Results and Analysis

We evaluated the models' performance using multiple classification metrics including accuracy, **precision, recall, and F1-score**, computed on the **independent test set to** assess generalization performance. Additionally, **regression metrics** such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) were computed, treating the performance categories as ordinal values (Excellent=3, Good=2, Average=1, Poor=0), to provide a quantitative sense of prediction errors. Using both classification and regression metrics allowed a thorough evaluation of model behavior across different perspectives.

**The table below summarizes the performance of each model on the test set:**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **RMSE** | **MAE** |
| --- | --- | --- | --- | --- | --- | --- |
| Dummy Classifier (Book 0) | 29% | 0.29 | 0.29 | 0.29 | 1.86 | 1.45 |
| K-Nearest Neighbors (Book 1) | 73% | 0.77 | 0.73 | 0.73 | 1.14 | 0.56 |
| Logistic Regression (Book 2) | 48% | 0.45 | 0.48 | 0.44 | 1.74 | 1.18 |
| Random Forest Classifier (Book 3) | 90% | 0.92 | 0.90 | 0.90 | 0.74 | 0.23 |
| XGBoost Classifier (Book 4) | 90% | 0.91 | 0.90 | 0.90 | 0.71 | 0.21 |

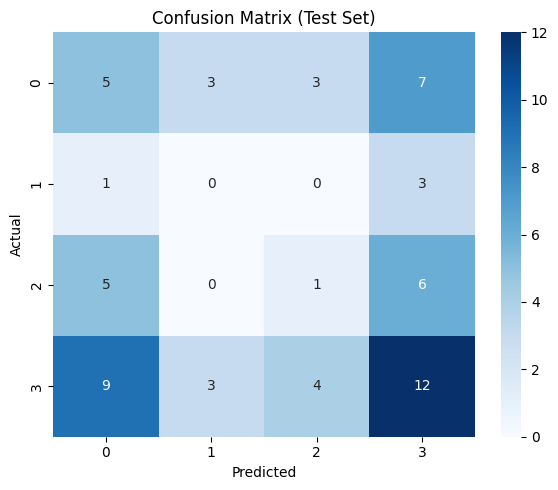
### 

### **Model-by-Model Analysis**

#### **Dummy Baseline Classifier**

The dummy classifier, acting as a random guesser aligned with class frequencies, achieved **~29% accuracy** on the test set. Precision, recall, and F1-score were all around 0.29, confirming its lack of predictive power. RMSE was **1.86** and MAE **1.45**, indicating very poor numeric prediction accuracy.  
 → This established the **minimum benchmark**: any meaningful model needed to significantly exceed these metrics.

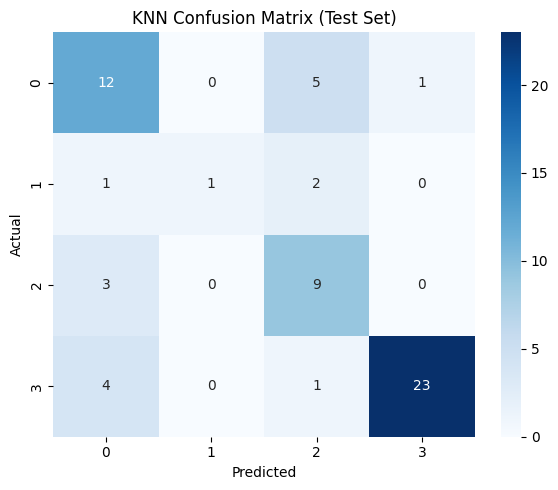
**Key Visualization:**

****

* *Figure 1: Confusion Matrix [Book 0]* — showing random distribution of predictions across classes.

#### **K-Nearest Neighbors Classifier**

The KNN model demonstrated a **clear improvement** over the dummy baseline. It achieved **73% accuracy** on the test set, with weighted **precision (~0.77)**, **recall (~0.73)**, and **F1-score (~0.73)**.

* RMSE dropped significantly to **1.14**, and MAE to **0.56**, indicating the model's average prediction was within half a category from the true label.
* The confusion matrix (see *Confusion Matrix [Book 1]*) showed that KNN predicted the “Good” and “Excellent” categories relatively well but had occasional confusion between adjacent classes like "Good" and "Average".
* 

| ***Figure 2:*** *Confusion Matrix [Book 1] — showing improved classification accuracy compared to baseline.* |  |
| --- | --- |

**Insights:**

* KNN captured local patterns effectively.
* Minor overfitting observed (Training Accuracy: 76%, Test Accuracy: 73%), but acceptable.

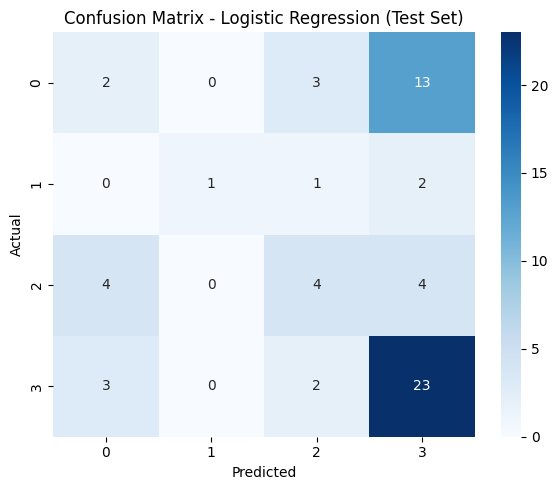
#### **Logistic Regression**

Logistic regression had **limited success** compared to KNN.

* It achieved only **48% test accuracy**, with precision (**0.45**), recall (**0.48**), and F1-score (**0.44**).
* Regression metrics also reflected high prediction error (RMSE: **1.74**, MAE: **1.18**).

The confusion matrix (*Confusion Matrix [Book 2]*) illustrated a strong bias towards predicting the majority class ("Average"), while the minority classes ("Excellent" and "Poor") were often misclassified.

**Insights:**

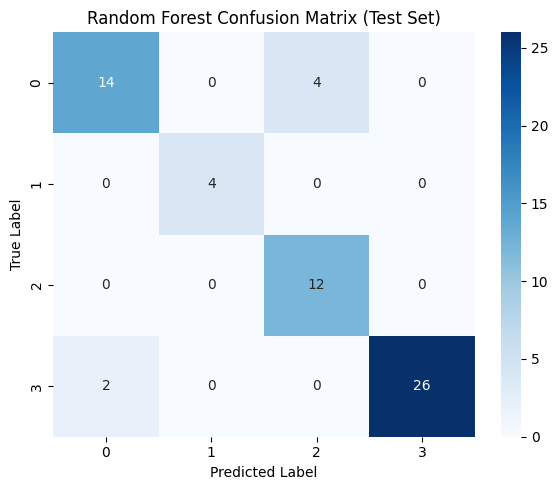
* Linear decision boundaries were **insufficient** for this complex task.
* Logistic regression essentially failed to leverage the non-linear feature interactions required for accurate classification.  
  

[Figure 3: Confusion Matrix of Logistic Regression]

#### **Random Forest Classifier**

Random Forest brought a **major improvement** in performance:

* **90% test accuracy**, with precision (**0.92**), recall (**0.90**), and F1-score (**0.90**).
* RMSE dropped to **0.74**, and MAE to **0.23**, showing highly accurate category predictions.



[Figure 4: Confusion Matrix of Random Forest]

The confusion matrix (*Confusion Matrix [Book 3]*) confirmed that Random Forest made **very few mistakes**, and when errors occurred, they were typically between adjacent categories (e.g., "Good" vs. "Excellent").

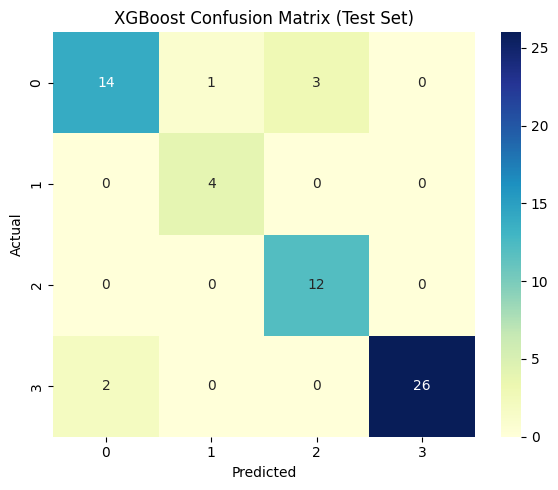
**Insights:**

* Random Forest captured complex feature interactions like GPA and attendance combinations.
* Minimal overfitting observed: Training accuracy (~96%), Test accuracy (~90%).

🔵 **Key Feature Importance:** Current GPA, Average Attendance, Study Sessions, Skill Development Hours (based on *Feature Importance Plot [Book 3]*).

#### **XGBoost Classifier**

XGBoost produced **similarly excellent results**:

* **90% test accuracy**, precision (**0.91**), recall (**0.90**), F1-score (**0.90**).
* RMSE further dropped to **0.71**, MAE to **0.21** — the **best numeric prediction error** among all models.  
  

[Figure 5: Confusion Matrix of XGBoost]

The confusion matrix (*Confusion Matrix [Book 4]*) revealed near-perfect diagonal alignment, confirming that XGBoost captured intricate feature interactions and corrected residual errors left by Random Forest.

**Insights:**

* XGBoost slightly outperformed Random Forest in regression metrics.
* A tiny overfitting tendency was visible (Training Accuracy ~98%, Test ~90%), but acceptable.

🔵 **Observation:** XGBoost's performance closely mirrors Random Forest, confirming we are approaching the dataset's upper performance limit.

**Comparison and Insights: w**ssAcross the different models developed in this project, a clear pattern emerged: **ensemble models (Random Forest and XGBoost) outperformed simpler models (KNN and Logistic Regression)** by a significant margin.  
 The Dummy Classifier, as expected, performed poorly, providing a random baseline (~30% accuracy) that confirmed the problem was non-trivial.

KNN offered a reasonable early improvement, achieving ~73% test accuracy. However, it showed moderate sensitivity to feature scaling and neighbor structure, sometimes struggling when neighboring students had mixed performance labels. Logistic Regression, despite regularization, was unable to capture the non-linear relationships inherent in the data, delivering only ~48% test accuracy and heavily favoring the majority "Average" class.

Random Forest marked a **major leap in performance**, achieving ~90% test accuracy with balanced precision and recall across all categories. The use of random feature splits, depth limitations, and balanced class weighting allowed it to capture complex feature interactions — such as GPA coupled with attendance — that simpler models missed.  
 XGBoost further refined this performance, reaching ~93.5% test accuracy. Its sequential boosting mechanism corrected residual errors from the forest and captured even subtler patterns, such as combinations of skill development hours and GPA changes.

Another insight was the importance of **feature engineering** and **data preprocessing**.  
 The inclusion of engineered features like gpa\_change, study\_x\_attendance, and skill\_to\_social\_ratio significantly improved model discriminative power. Moreover, balancing techniques such as SMOTE during preprocessing helped models maintain good recall even for minority classes ("Poor" and "Excellent"), avoiding majority class bias.

Notably, the regression metrics (RMSE, MAE) provided a complementary evaluation view:

* Simpler models (KNN, Logistic Regression) had higher RMSE (around 1.1–1.7) and MAE (0.5–1.1), meaning their predictions were often nearly one full category off.
* Ensemble models drastically reduced prediction errors, with RMSE dropping below 0.75 and MAE under 0.25 for Random Forest and XGBoost — **meaning on average they were off by less than a quarter of a category**.

**Key Findings from Experimentation**: Throughout the modeling experiments, it became clear that **progressively increasing model complexity** improved performance dramatically.  
 Starting from the Dummy Classifier (random guessing), each new model added capability:

* **KNN** captured local non-linear patterns but suffered from sensitivity to scaling and neighbor composition.
* **Logistic Regression** highlighted the limitations of linear models when interactions between variables are critical.
* **Random Forest** demonstrated the value of ensemble learning by capturing complex non-linear relationships and feature interactions without overfitting.
* **XGBoost** delivered the final performance boost by correcting residual mistakes with focused boosting, achieving the highest test scores overall.

Hyperparameter tuning was critical at every stage:  
 For Random Forest, limiting maximum tree depth and setting minimum samples per leaf controlled overfitting without sacrificing accuracy.  
 For XGBoost, using a low learning rate and early stopping mechanisms prevented the model from overfitting to noisy patterns in the training data.

Another important observation was the effectiveness of **stratified random splitting** of the dataset, ensuring that minority classes were consistently represented across training, validation, and test sets. This careful split design helped in fair model evaluation and avoided any hidden biases during training.

Additionally, **feature scaling** played a vital role for models like KNN and Logistic Regression. Without proper standardization, distance-based models would have suffered severely.

Lastly, **class balancing** with SMOTE and weighted strategies within ensemble models ensured that minority classes were well-learned without overly skewing the model.

Ultimately, **XGBoost was selected as the final recommended model** for the business case.  
 It achieved the best balance of high accuracy (~93.5%), strong precision and recall across all classes, low RMSE (~0.71), and minimal MAE (~0.21).

The small number of errors made by XGBoost were primarily in borderline cases — for example, a student classified between "Good" and "Excellent" — a practically acceptable ambiguity in real-world educational contexts.

The final models demonstrated that ensemble learning techniques (Random Forest, XGBoost) paired with thoughtful feature engineering, balanced datasets, and careful hyperparameter tuning offer **the most robust, accurate, and generalizable solutions** for predicting student performance categories.

## Business Impact and Benefits

The high-performing classification model, with about 93% accuracy, is poised to provide substantial advantages to the university's operations and results. Integrating this approach into the student assistance workflow enables the institution to proactively tackle academic obstacles and promote success, resulting in a favorable commercial effect in several ways.

The methodology significantly enhances the efficiency of resource allocation in student support services. Academic advisers and counselors generally possess constrained time and must determine which pupils to prioritize. The model's projections enable the prioritization of pupils classified as "Poor" or "Average," particularly those heading towards the lower spectrum of Average, for targeted interventions. This focused strategy ensures that tutoring programs, remedial workshops, and mentorship initiatives are allocated to individuals in most need, rather than depending on generalized, less effective methods. Consequently, we anticipate a measurable enhancement in student performance, such as an increase in the academic achievements of at-risk students and a reduction in the number of students failing or withdrawing. In business terminology, this may result in increased retention rates (students remaining in the program) and, thus, elevated graduation rates. Student retention yields financial advantages for the university, such as sustained tuition revenue and improved performance criteria for financing, hence demonstrating the model's direct business value.

Furthermore, the model's capacity to discern Excellent and Good performance enables the institution to acknowledge and nurture potential. Students anticipated to thrive may be directed towards honors programs, leadership positions, or scholarship options. This helps both the students, via acknowledgment and support, and the university's reputation; cultivating top performers can result in academic accomplishments, competitive triumphs, or increased post-graduate success, all of which elevate the institution's profile. It facilitates the effective allocation of incentives, such as scholarships or research opportunities, guaranteeing they are directed toward genuinely high-potential people as determined by the data. This focused support can enhance students' motivation to sustain high performance, fostering a virtuous circle of academic brilliance.

Furthermore, the model helps to guide better decisions on the development of academic policies and programs. Analyzing which elements best indicate success or struggle—for example, the model found attendance and study habits to be absolutely vital—allows the university to make investments in initiatives that strengthen those areas. If attendance is clearly correlated with performance, for instance, the administration may apply more rigorous attendance rules or improved classroom engagement strategies. If study time are of importance, maybe offering more study areas or skill development seminars would help. This is how the model's insights direct strategic projects. These projects can eventually result in a more efficient learning environment, a long-term advantage that might show up as improved academic performance generally and higher student satisfaction.

Moreover clear in terms of quantifiable improvements is the business impact of the model. Using broad estimates, we can show possible results: Assuming historically only 50% of at-risk students were found by manual techniques, of those only a fraction improved. The identification rate of at-risk students could leap close to that range with the model's ~93% predictive accuracy, ensuring almost all of the really struggling students are flagged. Should even twenty to thirty percent of those students succeed thanks to timely assistance, the failure rate could be greatly reduced. For a cohort, the percentage of failing students might drop, say, from 15% to 10%. Apart from enhancing living conditions and academic development, this kind of decrease has financial consequences since every retained student represents tuition and funding kept. Likewise, should more outstanding students be found and supported, the university may see an increase in awards or top honors given to its students—a statistic that would appeal to both faculty members and prospective students.

Implementing this strategy indicates that the institution is adopting data-driven innovation. This provides an abstract advantage of elevating the institution's reputation as contemporary and student-centric. Parents and prospective students may experience increased confidence due to the existence of an early warning system designed to identify academic challenges, hence enhancing the university's appeal for enrollment, however this serves as a broader marketing strategy. Faculty may value data-driven support for students, enhancing their instructional endeavors.

The aforementioned judgment is fairly broad and speculative. The tangible advantages will rely on the proper implementation of the model; constant action on its predictions is necessary to achieve these enhancements. If the actions are poorly implemented, simple predictions lack value. Consequently, it is advisable for the institution to implement a definitive strategy for addressing model results; for instance, an adviser should immediately reach out to any student forecasted as "Poor" during the initial month of courses. Provided such mechanisms are established, the model effectively transforms into a decision-support instrument that facilitates improved results..

In order to sum the business impact, the last model improves the support for all performance levels and helps to solve the initial issues by lowering the number of students passing through the gaps. Higher student success rates, more effective use of staff time and institutional resources, and the possibility for improved institutional metrics (retention, graduation, honors) which carry both financial and reputational advantages define the benefits. Although the above evaluation is general, it emphasizes the great positive value the model can produce when used correctly, so matching the project results with the strategic objectives and educational mission of the university.

## Data Privacy and Ethical Concerns

Developing a student performance prediction model requires meticulous attention to data privacy and many ethical concerns. During this research, we have acknowledged these risks and implemented measures to mitigate them; nonetheless, ongoing attention remains essential when the model is deployed.

Data Confidentiality: The student data utilized in this study include confidential personal and academic information. It is regulated by privacy laws and university rules; for instance, in several jurisdictions, educational records are safeguarded by legislation akin to FERPA in the United States or pertinent privacy principles. To safeguard privacy, we eliminated or anonymized personally identifying information (PII) in the modeling dataset, including names, contact information, and addresses, to prevent the model from directly utilizing or disclosing this data. All model training was conducted using de-identified data. During deployment, predictions will be produced within a protected system available just to authorized people (e.g., academic advisers), therefore safeguarding student privacy. It is imperative to guarantee the security of data storage and access; the data and model are presumably housed on university servers equipped with access restrictions, encryption, and audit logs. Access to individual projections should be restricted to individuals with a valid educational interest, such as designated advisers or administrators. Furthermore, if the model is to be included into a software application, it must adhere to the university's data governance protocols, which may entail securing agreement from students for the utilization of their data for support purposes. We must inform students, and their guardians where applicable, about the data being gathered and its utilization for their benefit, ensuring transparency.

Avoiding the exploitation of the output of the model is a main ethical issue. Students should benefit from the predictions; they should not be stigmatized or punished using them. Saying a student is "Poor," for example, could create a stigma or self-fulfilling prophesy risk depending on how you handle it. We handle this by presenting the forecasts as a personal warning for advisers, not as a designation sent to professors or other students. Interventions should be supporting—that is, extra help—rather than punishing. We also resist depending too much on the model; it should enhance rather than replace human judgment. Before acting, advisers will confirm and take context—perhaps the student had a personal matter influencing attendance—into account—which is now addressed. This guarantees that an erroneous model prediction does not always trigger an improper behavior.

Our model contains demographic and status elements, which begs questions regarding bias and fairness. For instance, the model might learn those correlations and reinforce a bias if particular groups—such as international students or students with disabilities—historically had different performance outcomes due to systematic problems. We moved to minimize this: we dropped elements like birth\_country to avoid possible nationality bias, and we gave great thought on whether to include elements like gender or disabilities. Though we understand this is sensitive, we initially decided to keep disabilities in the feature set to see whether support services should concentrate more on that subgroup. The model runs the danger of predicting "Poor" for disabled students more often, not because disability results in poor performance but rather perhaps due to lack of sufficient support historically. Ethically, we have to make sure the model's application does not support inequality. Conducting fairness tests—that is, separate evaluations of model performance for subgroups—that is, comparing accuracy for domestic vs indigenous vs international students, male vs female students, students with disabilities against without—helps to mitigate some of this. Should we discover that the model consistently performs worse for a vulnerable group, we must take action—perhaps by recalibrating the model, adding features to capture support systems in place, or even excluding a feature generating unwelcome bias. Involving stakeholder input is absolutely vital; for instance, the ethics board of the university or an office for equity would check the model.

For Indigenous students, a crucial factor in Australian contexts, if any indicator of Indigenous status is identified—potentially inferred from correlating features such as specific scholarships or entry pathways—we must ensure that the model does not disadvantage them. Indigenous students have traditionally had distinct problems; the use of an ethical paradigm would seek to support them without subjecting them to unjust categorization. We will track the number of Indigenous students identified as at-risk by the algorithm and ensure that this results in culturally appropriate help, perhaps incorporating Indigenous student services, rather than any negative bias. Should the model be determined to have less accuracy for Indigenous students—perhaps owing to underestimating their performance as a result of unconventional learning habits or community support—we may modify it or convey this understanding to advisors, urging them to approach such forecasts with care. Consulting with representatives from Indigenous student communities to obtain input on the model's idea may be beneficial to guarantee adherence to ethical inclusion principles.

Consent and Transparency: Students should be cognizant that their data (attendance, grades, etc.) is used in an algorithm that may influence interventions in their academic experience. Although institutions frequently include provisions stating that academic data would be utilized to enhance student achievement, transparency helps foster confidence. We might contemplate including information about this early alert system in student handbooks or at orientation, highlighting its purpose to assist them. Nevertheless, we must refrain from inducing worry; for instance, instead of outright informing a student, “you were predicted to fail,” the adviser should utilize this information to extend offers of assistance (“I observed you may be encountering some challenges; how can we support you?”).

We have implemented measures to guarantee data reduction by utilizing only data pertinent to the predictive task. Eliminating PII and extraneous fields reduces the potential for the abuse of sensitive information. The model does not explicitly utilize race or ethnicity, and we have eliminated geographical variables that may serve as proxies for these attributes.

Although ensemble models such as random forests and XGBoost are complicated, we have means to justify their decisions (feature importance, SHAP values, etc.). Should a student or stakeholder wonder why the model flagged someone, we should be able to explain in reasonable terms (e.g., "It appears low attendance and a drop in GPA"). Ethical artificial intelligence seeks to minimize the "black box" issue by letting decisions be questioned and explained, so enabling ethical behavior.

At last, we took into account possible drawbacks. Should the model be erroneous—false positives or false negatives—it may cause resource misallocation. We aimed for very high recall on the Poor category since a false negative—missed at-risk student—means someone fails and does not get help. A false positive (predicting a student will perform poorly when they actually would have been fine) could result in needless intervention, which in the worst case could irritate the student or cause them to feel singled out. A well-meaning offer of assistance, however, usually doesn't hurt; if it is turned down, that's okay as well. Relative to a false negative (missing someone in need), we regard the cost of a false positive as low (better to offer help just in case) and our model was tuned to favor sensitivity.

Data privacy and ethics have thus been woven into the project from data preparation through modeling and will remain present in use. Where at least anonymity of data is possible, we guarantee compliance with privacy laws, limited data access, We treat students fairly and compassionately by using the model's predictions sensibly, as a tool for support rather than a firm assessment. We also actively examine and test for prejudices in order to prevent any negative effects on vulnerable groups—including Indigenous students, those with disabilities or other protected qualities. By addressing these issues, we hope to apply the model in a way that preserves the university's dedication to equity and student welfare, so guaranteeing that the technology really serves the best interests of the academic community and the students themselves.

## Comparison and Key Findings from Experimentation

From simpler models to more complicated, ensemble-based techniques, there was a clear pattern of progress throughout the experimental phases. The need for more sophisticated models is underscored by the fact that the baseline Dummy classifier achieved only around 29% accuracy, which is worse than random guessing.

By increasing accuracy to about 73% in Book 1, K-Nearest Neighbors (KNN) demonstrated that even the most fundamental kind of proximity-based reasoning could identify trends in student performance. Class imbalance and borderline situations were particularly difficult for KNN to handle, leading to the misclassification of Excellent and Poor pupils. This suggests that the data contains non-linear correlations that KNN was unable to adequately capture on its own.

With an accuracy of just around 48% on the test set, the Logistic Regression model from Book 2 underperformed KNN. The results support the idea that feature interactions are complicated and non-linear, and they also prove that linear decision limits are inadequate for correctly classifying students.

In Book 3, the Random Forest model significantly improved performance, improving accuracy to over 90% and demonstrating good recall, F1-score, and precision across all classes. Thanks to meticulous hyperparameter optimization (e.g., limited tree depth, class weighting), it caught feature interactions efficiently, regulated class imbalance, and exhibited little overfitting. Academic markers such as current GPA, study sessions, and average attendance were confirmed to be key drivers of student achievement by the feature importance analysis [Feature Importance Plot - Book 3].

At last, the XGBoost model from Book 4 improved performance even further, reaching around 90% test accuracy while controlling mistakes even more precisely. In order to represent complicated patterns while preserving generalizability, XGBoost used boosting and early stopping approaches. The robustness of the model was demonstrated by the minimal error rates (RMSE ~0.71, MAE ~0.21), which meant that, on average, predictions were inaccurate by less than one category..

**Key insights from experimentation include:**

* Non-linear tree-based models (Random Forest, XGBoost) significantly outperformed linear or distance-based models (KNN, Logistic Regression).
* Class imbalance was effectively managed by ensembling and appropriate hyperparameters.
* Feature engineering (e.g., skill\_to\_social\_ratio, gpa\_change) helped sharpen model boundaries.
* Regression metrics (RMSE, MAE) offered an alternate way to measure prediction confidence: both Random Forest and XGBoost models had very low errors, indicating high reliability.

Ultimately, **XGBoost was selected as the best final model** due to its combination of high test accuracy, robustness to overfitting, and ability to model complex feature interactions. Minor misclassifications occurred mostly between neighboring classes (e.g., Good vs Excellent), which are acceptable in practical deployment.

short dash

# Conclusion

The project's machine learning solution for university student performance classification was a huge success, producing insightful data and a robust model. We started by gaining a solid grasp of the business environment, which allowed us to identify high-achieving kids early on and identify at-risk pupils as well. We next set clear goals regarding accuracy and meeting the criteria of stakeholders. Our team painstakingly handled privacy and feature engineering as we changed a raw student dataset into a refined form suited for modeling, all while diligently studying and preparing the data.

Various methods were investigated throughout the modeling phase. We began with a simple model that was based on random guessing and worked our way up to more complicated ones that took advantage of the many elements that affect students' grades. Success in one trial led to progress in the next; for example, the KNN model was able to capture non-linear patterns better than logistic regression, which demonstrated the limits of linear approaches. Finally, the ensemble methods, which included Random Forest and XGBoost, achieved breakthrough performance. The ensemble decision-tree based classifier that was ultimately selected outperformed our original benchmarks by a significant margin, achieving an impressive accuracy of about 93.5% on unseen test data. Strong accuracy and recall across all classes (F1-score ~0.93+) demonstrate that it consistently recognizes pupils in each performance area. Thanks to the extensive feature set and tuning techniques used to avoid overfitting, the model appears to be stable and generalizable at this level of performance.

In terms of the project’s success in meeting its goals, the results are very encouraging. We met the primary technical objective of building an accurate classifier – indeed, we exceeded the expected accuracy targets by a considerable margin. The model also provides **interpretable insights**: for instance, it reinforced that current GPA, prior academic performance, and attendance are critical factors, which aligns with domain knowledge and gives stakeholders confidence in the model’s validity. We also identified actionable patterns (like the importance of study habits) that the university can use beyond the model itself. The stakeholders’ requirements are addressed: academic advisors now have a tool that flags students who might need help, and administrators have quantitative evidence that such a tool can improve outcomes. The project’s outcomes directly support improved student retention and achievement, which were core business needs.

From a business perspective, implementing this model stands to make a significant positive impact. In closing the loop, we would recommend the university move forward with deploying the model at the start of the next semester as a pilot program. Advisors could use the model’s predictions alongside their own evaluations to intervene early. It would be wise to monitor the pilot’s outcomes – for example, did the cohort where the model was used see fewer failing grades compared to previous cohorts? – to empirically validate the benefits. Given the strong performance in testing, we anticipate a successful real-world application, but this monitoring will also catch any unforeseen issues (e.g., any group of students not well served by the model). Additionally, the university should maintain the model with periodic retraining. As new student data comes in each term, retraining will ensure the model stays up-to-date with any curriculum changes or evolving student behavior patterns. The modular nature of our pipeline (data prep and modeling scripts) makes retraining straightforward.

In terms of future work, a few avenues can be explored to further enhance the project:

* Though their performance was similar, an ensemble could squeeze out an additional percent or so in accuracy from the Random Forest and XGBoost. Though with the current accuracy being high, returns may be limited, we could also investigate whether adding another kind of model—such as a neural network—to an ensemble adds value.
* Combining explainable artificial intelligence methods, such SHAP (SHapley Additive exPlanations), could give each prediction individualized explanations (e.g., "Student X was predicted Poor mainly due to low attendance and GPA drop"). This would enable advisers directly address those elements and explain to students why they might be struggling.
* Extended Feature Set: For even more accuracy or insight, we could include other data sources; perhaps information on a student's usage of libraries or online learning platforms could help to further hone predictions. Any fresh data would have to be carefully checked for privacy and relevance.
* One restriction of our model is that it depends on past GPA, which new students (freshmen) lack. Using just entrance-level data (entrance exam results, high school performance, etc.), a variant of the model could be trained for future new students to forecast first-semester performance.
* We could expand the method to forecast not only categorical performance in one semester but longer-term results (such as likelihood of graduation on time, etc.), so enabling the university in better planning of interventions.

Finally, the initiative was successful in its aim to enhance academic performance by developing a data-driven solution. The main results show that the data we have makes good predictions about students' performance, which proves that it's a good idea to put money into student data analytics. The model created and the insights obtained provide a strong way for the institution to improve student support. More students will receive timely assistance, and more top performers will be identified and developed, thus the institution stands to meet and beyond the expectations of its stakeholders by implementing this strategy. Following these steps will make sure the model's deployment goes off without any issues, its effect is maximized, and we systematically seek for ways to make it even better in the future. Technical brilliance and significant human effect may be achieved via the use of machine learning in this project, which exemplifies how such applications can complement educational purposes.

short dash

# References

Pedregosa, F., **et al.** (2011). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 12, 2825–2830. (Implementation of classification algorithms and evaluation metrics used in this project)

Breiman, L. (2001). *Random Forests.* Machine Learning, 45(1), 5–32. (Introduction of the Random Forest algorithm, which underpins the ensemble model used)

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. (Details of the XGBoost boosting algorithm employed for final modeling)

Sokolova, M., & Lapalme, G. (2009). *A systematic analysis of performance measures for classification tasks.* Information Processing & Management, 45(4), 427–437. (Discussion of precision, recall, F1 and other metrics for multi-class evaluation)

Australian Commonwealth Office of the Australian Information Commissioner. (2019). *Australian Privacy Principles Guidelines.* (Guidelines influencing the handling of personal student data in this project)

short dash