Assignment 2Classification Models

short line

Nipunn Khurana

Student ID: 25009200  
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**](#_if9iwijz2pg3)

[a. Business Use Cases 2](#_bnjfuh5ewogw)

[b. Key Objectives 2](#_axrp8wkwcpwb)

[**2. Data Understanding 3**](#_yg03y3w5hig9)

[**3. Data Preparation 4**](#_tj77rrmixzhi)

[**4. Modeling 5**](#_6vwh14nc44ga)

[**5. Evaluation 6**](#_mngiyx5tet6w)

[a. Results and Analysis 6](#_25v57zsj5m7g)

[b. Business Impact and Benefits 6](#_hkob76wu4d6q)

[c. Data Privacy and Ethical Concerns 6](#_uf1z6gbsejg6)

[**6. Conclusion 8**](#_9hr11g79asdx)

[**7. References 9**](#_ma3y8ytjvp1y)

# Business Understanding

## Business Use Cases

*The idea for this project likely originated from an educational institution's growing need to improve student success rates and better allocate support resources. With rising enrollments and increasing diversity in student backgrounds, universities face challenges in providing personalized interventions to students who may be at risk of academic failure. Traditional methods such as manual monitoring, generic counseling, or broad academic policies are often reactive, inefficient, and costly. As a result, using data-driven approaches to predict student performance levels and identify students needing targeted support has become highly valuable.*

*The specific business use case centers around building a predictive model that can classify students based on factors like previous GPA, study habits, co-curricular involvement, living arrangements, and other socio-academic variables. By deploying such a model, universities can proactively offer additional mentoring, mental health services, financial assistance, or academic coaching to students before they underperform. Furthermore, identifying patterns among high-performing students allows institutions to promote best practices and refine curriculum delivery.*

*The challenge that motivated this project is the high cost and low scalability of human-based interventions, as well as the inefficiencies in identifying struggling students early enough. Additionally, there is a growing opportunity to leverage existing data more meaningfully to drive better academic outcomes. Machine learning algorithms are particularly relevant here because they can detect subtle, complex patterns across many variables that are not obvious to human advisors. They allow predictive decisions to be made at scale, continuously updated as new data becomes available, and personalized to individual students rather than applying a one-size-fits-all solution.*

*Thus, this project aims to operationalize machine learning as a proactive decision-support tool, enhancing both student experience and institutional performance in a measurable, cost-effective way.*

## Key Objectives

*The goal of this project is to build an accurate and scalable machine learning model to predict student performance early in the academic term. Desired outcomes include early identification of at-risk students, improved retention rates, and more efficient allocation of academic support resources. The project also aims to highlight key factors influencing student success to foster a proactive, data-driven academic environment.*

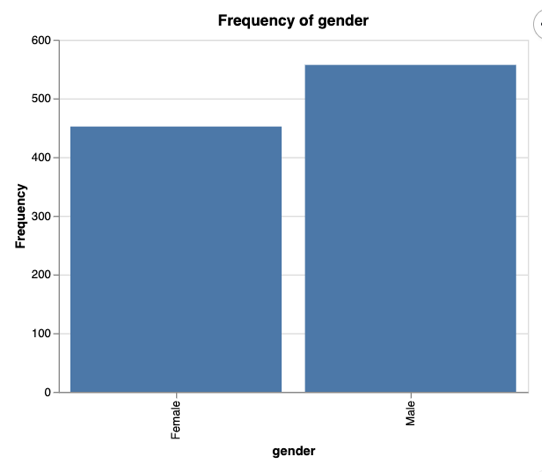
*Stakeholders include academic advisors, student support teams, university administrators, and students. Advisors need early warnings for personalized interventions, support teams require prioritization indicators, and leadership expects improved outcomes and resource management. Students benefit from timely, targeted academic support.*

*The project meets these needs by applying machine learning algorithms to uncover complex patterns in student data. Decision Trees were chosen for their ability to handle mixed data, deliver high interpretability, and produce actionable insights. The system empowers stakeholders to make informed, timely decisions, improving educational outcomes for both students and the institution.*

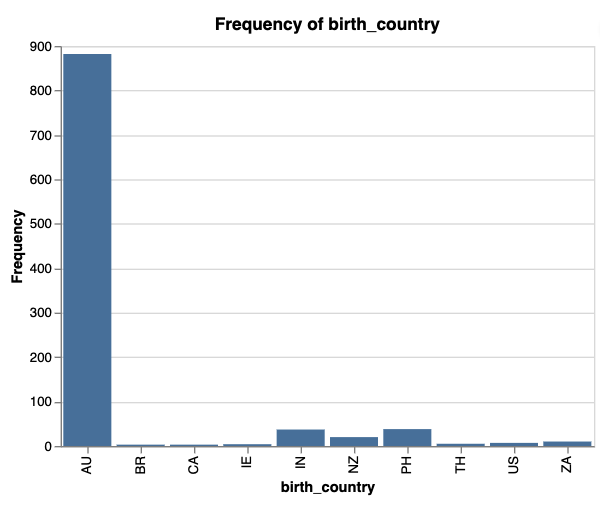
short dash

# Data Understanding

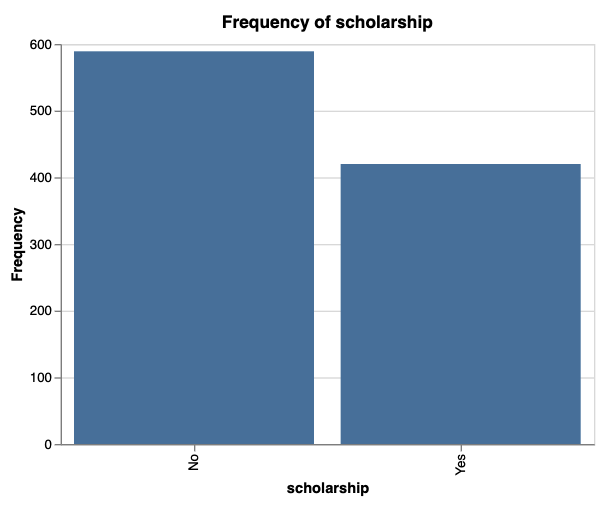
*The dataset was sourced internally from university enrollment and academic tracking systems, covering demographic, academic, behavioral, and socio-economic features. Key academic variables include previous\_gpa, current\_gpa, completed\_credits, study\_hours, study\_sessions, average\_attendance, skills\_development\_hours, and area\_of\_interest. Engagement metrics such as co\_curricular activities, social\_media\_hours, skills, and has\_consulted\_teacher were included. Socio-economic factors like birth\_country, living\_arrangement, health\_issues, disabilities, has\_diploma, and house\_income provided background context. Other variables included gender, relationship status, scholarship status, university\_transport, and learning\_mode. Administrative fields (student\_id, full\_name, email, phone\_number, address details) were present but not used for modeling. The target variable, target, classified students based on end-of-semester performance*

.

*Figure 1: Frequency of gender*

**

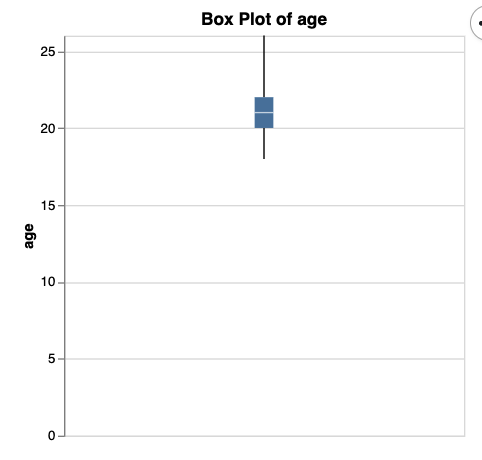
*Figure 2: Frequency of Birth Country*

**

*Figure 3: Frequency of Scholarship*

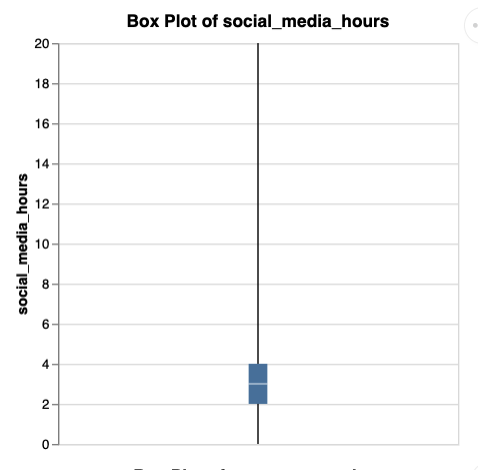
*The exploratory analysis of the categorical variables reveals several important insights about the student population. The gender distribution (Figure 1) shows a fairly balanced representation between males and females, suggesting an inclusive environment but with a slight skew toward male students. The birth country (Figure 2) distribution highlights that most students are domestic, primarily from Australia, with only a small percentage representing international backgrounds. This indicates that while diversity exists, the student body is largely localized, influencing how support services and engagement programs might be structured. Additionally, the scholarship distribution (Figure 3) shows that a higher number of students are not receiving scholarships compared to those who do, emphasizing potential opportunities for the institution to expand financial support initiatives or better promote existing scholarships.*

short dash

**

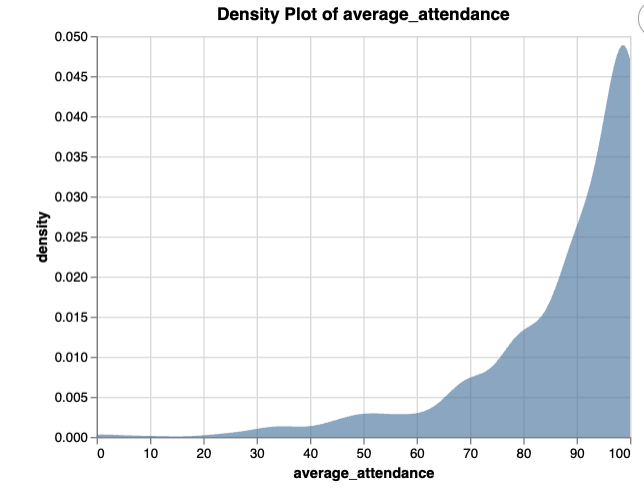
*Figure 4: Box Plot of Age*

*Most students are aged between 20 and 22, showing a tight, consistent age range across the population.*

**

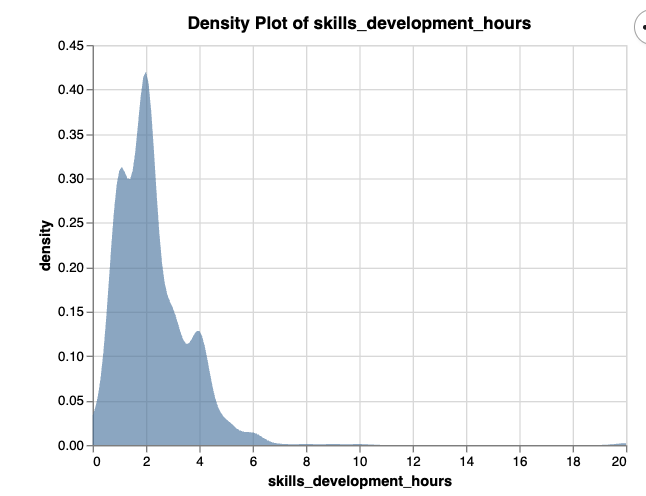
*Figure 5: Box Plot of Social Media Hours*

*Most students spend around 2 to 4 hours daily on social media, with some extreme outliers.*

**

*Figure 6: Density Plot of Average Attendance*

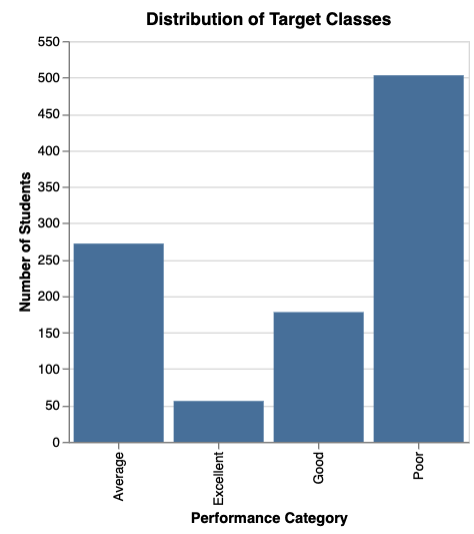
*Students’ average class attendance is heavily skewed towards the higher end, with most maintaining around 90–100% attendance.*

**

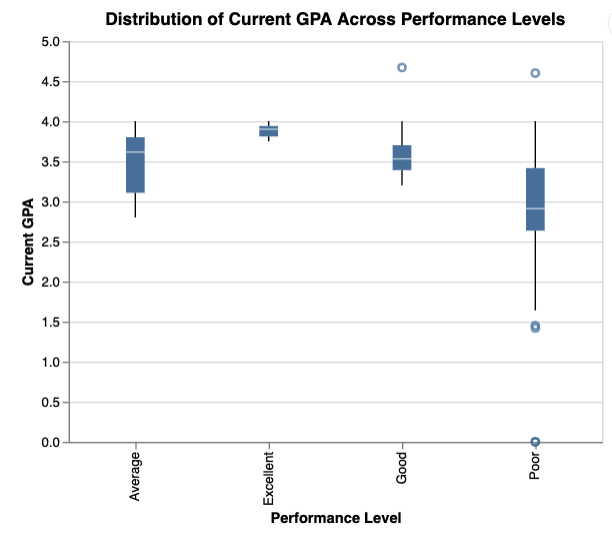
*Figure 7: Density Plot of Skill Development Hours*

*Most students spend around 2–3 hours daily on skill development, with the distribution heavily skewed toward lower hours.*

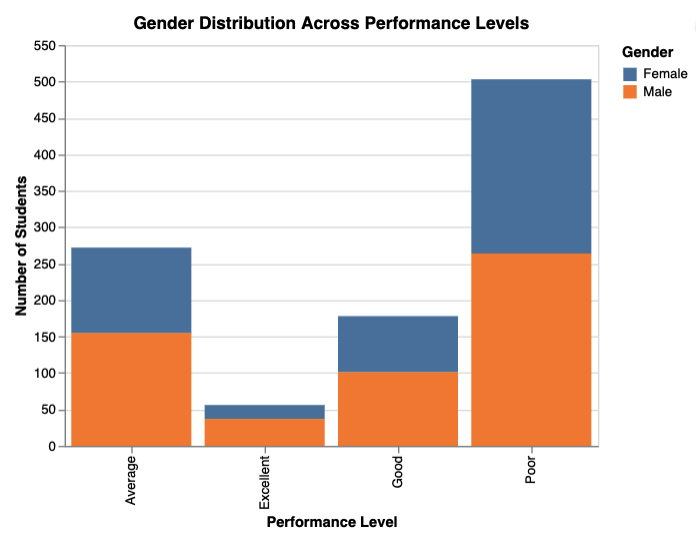
*TARGET VARIABLE: ‘Target’*



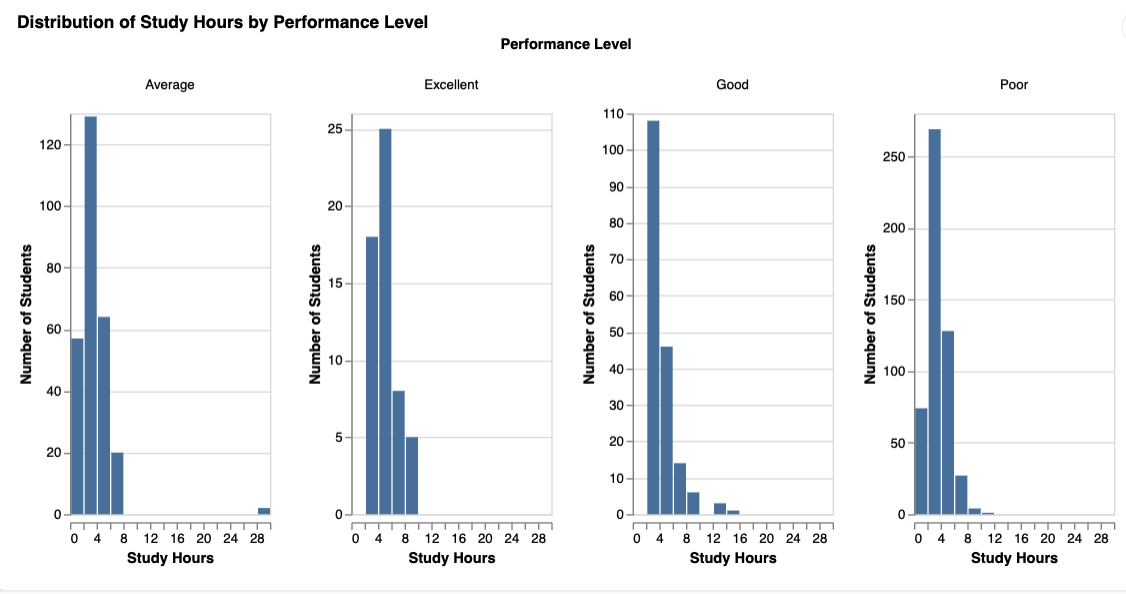
*Figure 8: Distribution of Target Classes*

**

*Figure 9: Distribution of Current GPA Across Performance Levels*

**

*Figure 10: Gender Distribution Across Performance Levels*

**

*Figure 11: Distribution of Study Hours by Performance Level*

*The target variable in this project is student performance, categorized into four classes: Excellent, Good, Average, and Poor (Figure 8). From the distribution, it is evident that most students fall into the "Poor" and "Average" categories, highlighting a critical area for intervention. The features collected provide meaningful insights into student behaviors and conditions that may influence performance. For example, current GPA ( Figure 9) shows a strong positive association with better performance classes, making it a vital predictor. Gender distribution across performance levels (figure 10) suggests that both males and females are impacted but with slight variations, hinting at demographic patterns worth exploring. Similarly, study hours (Figure 11) display a clear trend, with students putting in more consistent study time tending to achieve better performance outcomes. Other features such as attendance, social media usage, and skills development hours show meaningful variation and provide opportunities to model behaviors that could predict student success. Overall, these features together create a rich dataset for training machine learning models to predict student outcomes and inform strategic interventions.*

# Data Preparation

*The following process were followed for the data preparation*

* *Feature Selection: Dropped irrelevant or redundant features using systematic approaches to retain meaningful predictors.*
* *Data Cleaning: Handled missing values, fixed inconsistencies, and ensured dataset integrity.*
* *Feature Engineering: Created new variables to capture deeper patterns in the data.*
* *Data Splitting: Divided the dataset into training and testing sets for unbiased model evaluation.*
* *Data Transformation: Scaled numerical features to standardize inputs and optimize model performance.*

Feature Selection

*The feature selection process involved two steps. First, we manually removed irrelevant and sensitive fields (e.g., student\_id, email, phone\_number, address details) to reduce noise, avoid data leakage, and protect privacy. Second, we used a Random Forest Classifier to rank feature importance, selecting top predictors like previous\_gpa, current\_gpa, and average\_attendance based on their contribution to the target.*

*Rationale for Approach:*

* *Manual removal eliminated obviously non-predictive or sensitive data.*
* *Random Forest provided an objective, data-driven way to prioritize impactful features.*
* *Combining domain knowledge and algorithmic evidence improved model clarity and performance.*

Data Cleaning

*We addressed missing values by removing records with gaps in area\_of\_interest and skills. Outliers were handled by removing unrealistic values (e.g., >15 study or social media hours/day) and capping extreme house\_income using the 99th percentile. Rare categories in area\_of\_interest and skills were grouped as "Other" to improve model robustness. These steps ensured a cleaner, more balanced dataset, reducing the influence of anomalies and sparsity.*

Feature Engineering

*We created a new feature, gpa\_improvement (current\_gpa - previous\_gpa), to capture academic progression over time. Positive values signal improvement, while negative values highlight decline. This dynamic metric adds predictive power by reflecting academic momentum, helping identify students needing early support.*

Split Datasets

*After feature engineering, the dataset was split into training and testing sets using an 80-20 split ratio, where 80% of the data was used for training the models and 20% was reserved for testing. Stratified sampling based on the target variable was applied to ensure that the class distribution in both training and testing sets remained consistent, which is crucial given the imbalance in performance categories. The rationale behind this strategy was to ensure the models learn from a broad representation of the data while leaving a sufficient portion untouched for a fair evaluation of generalization ability. Data splitting is a critical step because it helps prevent overfitting, provides an unbiased estimate of model performance, and ensures that the model’s success isn't falsely inflated by memorizing the training data.*

Data Transformation: One Hot Encoding

*In the data transformation phase, One-Hot Encoding was applied to convert the categorical variables into a machine-readable numerical format. First, all categorical columns were identified based on their data types. Then, using OneHotEncoder from scikit-learn, the encoder was fitted* ***only*** *on the training set to prevent data leakage, ensuring that the model learned the structure of categorical data from training alone. After fitting, the training and testing datasets were both transformed, creating new binary columns for each unique category. The original categorical columns were then dropped, and the encoded features were merged into the dataset. This transformation was crucial for enabling machine learning algorithms, particularly tree-based and distance-based models, to process categorical features effectively without imposing ordinal relationships where none exist.*

short dash

# Modeling

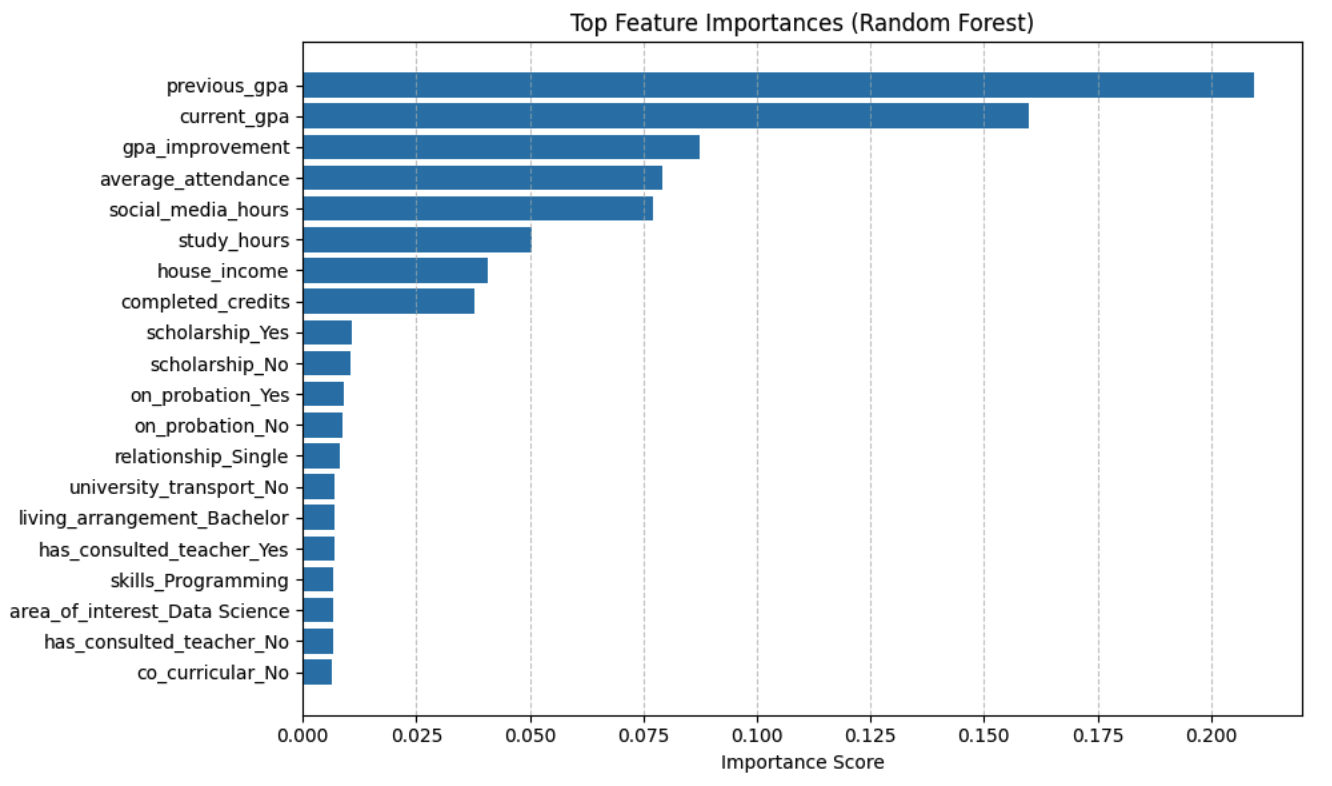
*Before modeling, we conducted a structured data preparation process. Irrelevant features were dropped, and Random Forest feature importance guided predictive variable selection. Data cleaning addressed missing values, outliers, and rare categories, while feature engineering introduced GPA improvement to capture academic progression. Categorical variables were one-hot encoded, and data was split into training, validation, and testing sets. Numerical features were standardized using StandardScaler, with the same transformation applied to the test data for consistency*.

FIGURE 11: Top Feature *Importances*

MODEL SELECTION

In this project, we used two machine learning algorithms for modeling: Decision Trees and Support Vector Machines (SVM).

* Decision Trees operate by recursively splitting the data based on feature values to create a tree structure that makes predictions based on learned decision rules. They are capable of handling both categorical and numerical data, and provide easy-to-interpret visual outputs.
* Support Vector Machines (SVM) work by finding the optimal hyperplane that best separates the classes in a high-dimensional feature space. SVMs aim to maximize the margin between classes, offering robustness in classification tasks, especially when the data is not linearly separable by applying kernel tricks.

RATIONALE BEHIND SELECTING THESE ALGORITHMS  
  
*The rationale for selecting Decision Trees and Support Vector Machines (SVM) was based on the nature of the dataset, the business goal, and the balance between interpretability and predictive strength.*

*Decision Trees were chosen for their simplicity, ability to handle both numerical and categorical data, and ease of interpretation, making them ideal for stakeholders to understand key factors influencing student success. Support Vector Machines (SVM) were selected due to their strong performance in handling complex, non-linear classification problems and robustness in high-dimensional spaces after encoding. By combining Decision Trees' transparency with SVMs' predictive power, the project achieved a good balance between model interpretability, accuracy, and business applicability.*

PARAMETER TUNING AND MODEL SELECTION PROCESS

SUPPORT VECTOR MACHINES  
  
*In this project, we tuned key SVM hyperparameters — C, gamma, kernel type, polynomial degree, and class weight — to optimize predictive performance. C controlled the trade-off between generalization and fitting the training data, with lower values (e.g., 0.2, 0.7) favoring flexibility and higher values (e.g., 2) reducing misclassification at the risk of overfitting. Gamma adjusted model sensitivity, where lower gamma allowed broader decision boundaries and higher gamma captured finer patterns in student behavior. We tested various kernels (RBF, polynomial, sigmoid) to model potential non-linear relationships, with different polynomial degrees capturing varying complexity levels. Class weighting was applied to address class imbalance and ensure fair treatment of all student categories. Exploring these settings across 12 SVM models helped balance model complexity, generalization, and fairness, leading to a robust and reliable prediction framework for student performance.*

DECISION TREES

*In this project, we tuned key Decision Tree hyperparameters — min\_samples\_split and max\_depth — to balance model complexity, control overfitting, and improve generalization to new student data. min\_samples\_split defines the minimum samples needed to split a node; smaller values (e.g., 5) allow deeper trees and capture more patterns but risk overfitting, while larger values (e.g., 20 or 40) create simpler, more robust trees. max\_depth limits how deep the tree grows; shallow trees (e.g., depth 6 or 8) simplify the model and reduce overfitting, while deeper trees can memorize noise. Tuning these parameters across multiple models (dt\_model1 to dt\_model6) helped us systematically find the best balance between accuracy, interpretability, and generalization for early student performance prediction.*

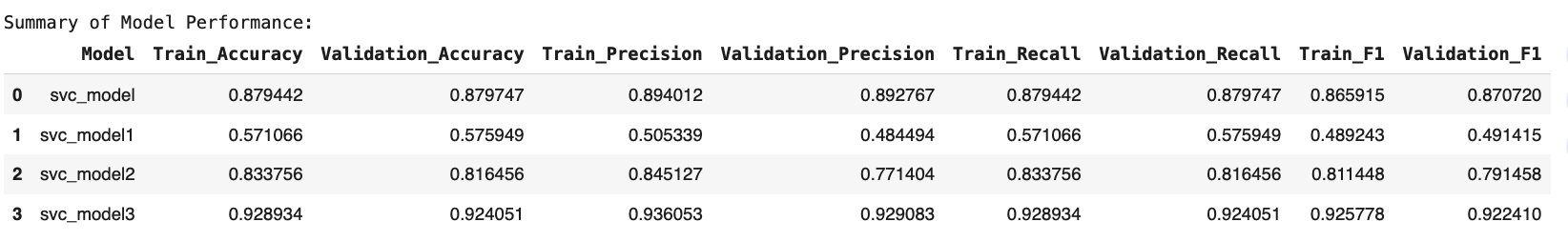
short dash

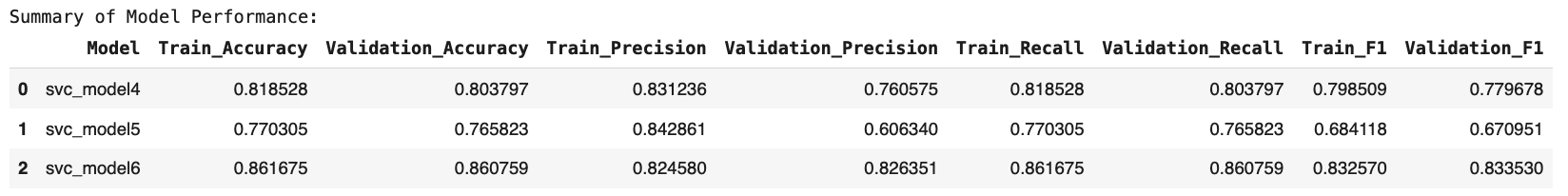
# Evaluation

## Results and Analysis

|  |  |
| --- | --- |
| **Model No.** | **Hyperparameters** |
| 0 | Default |
| 1 | (C=0.2) |
| 2 | SVC(C=0.7) |
| 3 | SVC(C=2) |
| 4 | (C=0.7, gamma=0.01) |
| 5 | (C=0.7, gamma=0.5) |
| 6 | (C=0.7, gamma=0.1) |
| 7 | (C=0.7, gamma=’scale’, kernel = ‘poly’) |
| 8 | (C=0.7, gamma=’scale’, kernel = ‘sigmoid’) |
| 9 | (C=0.7, gamma=’scale’, kernel = ‘poly’, degree = 2) |
| 10 | (C=0.7, gamma=’scale’, kernel = ‘poly’, degree = 5) |
| 11 | (C=0.7, gamma=’scale’, kernel = ‘sigmoid’, class\_weight = ‘balanced’) |

## Table1: SUPPORT VECTOR MACHINE PARAMETER





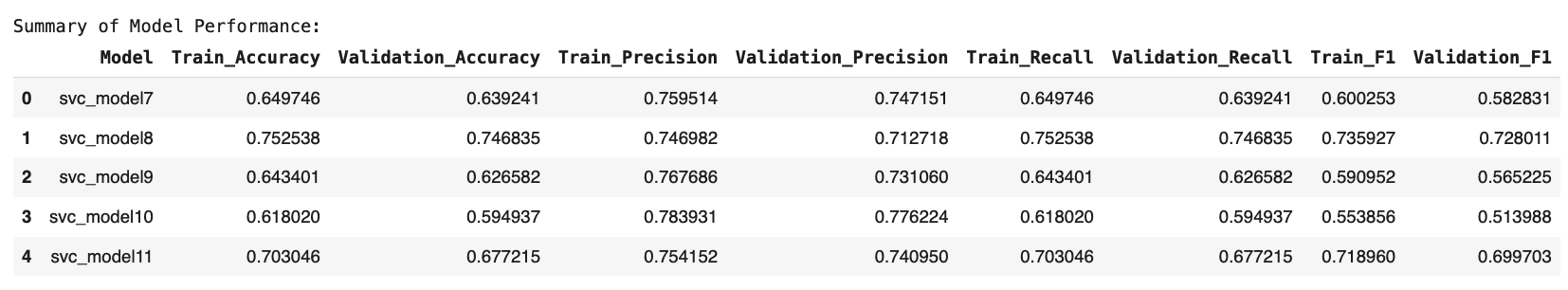


TABLE 2: RESULTS SVM

*Analysis of SVM Model Performances:*

*Among the models, svc\_model3 (C=2) achieved the highest performance with a validation accuracy of 92.4% and consistently high precision (93%), recall (92%), and F1-score (92%), making it the best-performing model. It also shows a strong balance between training and validation performance, indicating minimal overfitting.*

*The default SVC model and svc\_model2 (C=0.7) also performed reasonably well, with validation accuracies of 87.9% and 81.6% respectively, though slightly lower compared to svc\_model3. The default model shows a strong generalization but slightly lower metrics compared to svc\_model3.*

*svc\_model1 (C=0.2) showed lower performance (57% validation accuracy), suggesting that a very low regularization parameter (C) led to underfitting, where the model failed to capture the complexity of the data.*

*When tuning gamma values (svc\_model4, svc\_model5, svc\_model6), svc\_model6 (C=0.7, gamma=0.1) performed the best with a validation accuracy of 86% and fairly strong F1-scores, indicating that adjusting gamma can help but still did not surpass svc\_model3.*

*For kernel variations (svc\_model7 to svc\_model11):*

* *Models using polynomial and sigmoid kernels (especially svc\_model9, svc\_model10, svc\_model11) performed worse overall with validation accuracies falling between 59% and 67%, showing that non-RBF kernels were less suited for this problem.*
* *svc\_model8 (sigmoid kernel) performed relatively better (74.6% validation accuracy) but still notably lower than* the RBF-based models.

RESULTS ON TEST DATA

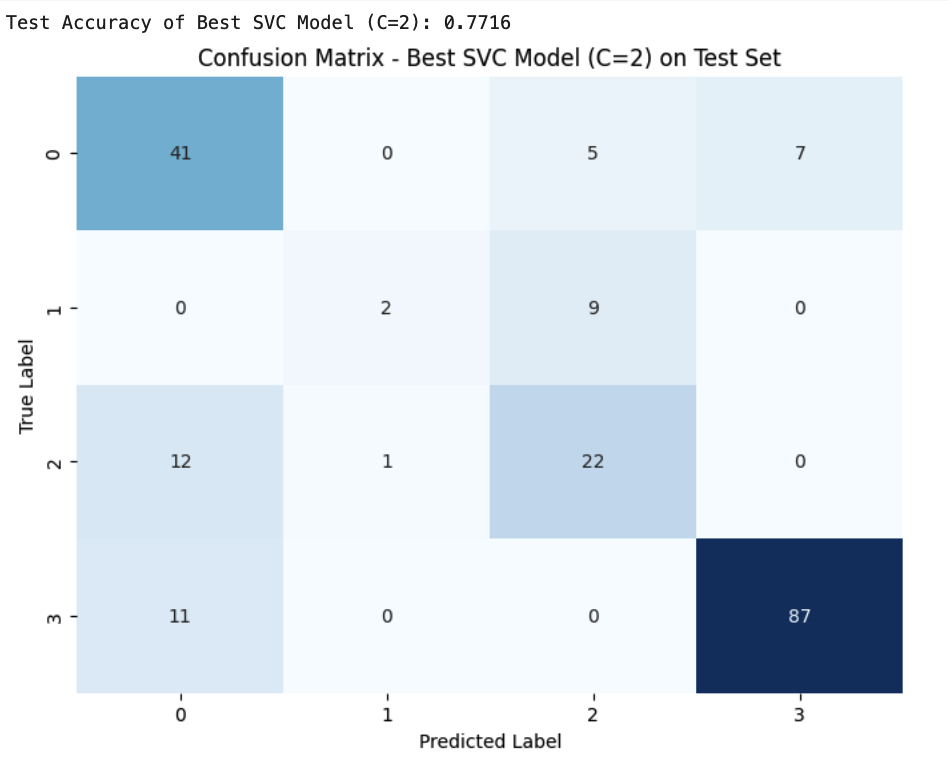
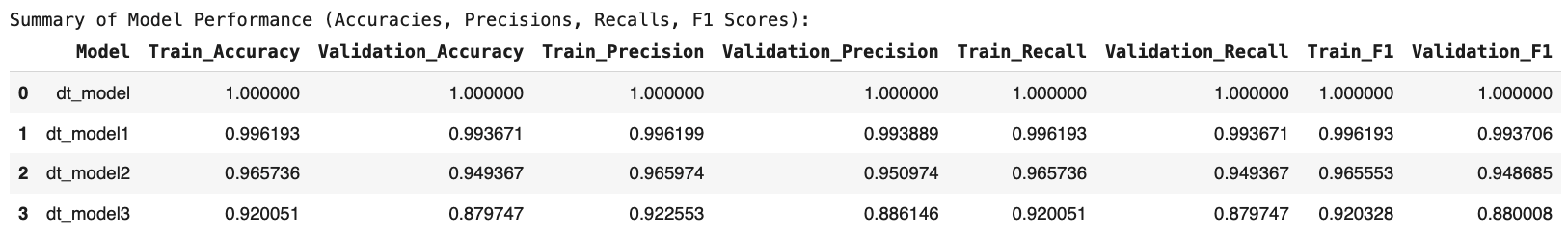


FIGURE 12: CONFUSION MATRIX ON TEST DATA (SVM3 Model)

*The best-performing SVC model (C=2) achieved a test accuracy of 77.16%, showing strong generalization and successfully identifying 'Poor' students, which supports the business goal of early risk detection. Some misclassifications between 'Average' and 'Good' students were observed, along with slight errors in 'Excellent' student predictions, but they have a lesser impact. Overall, the model prioritizes identifying at-risk students effectively and provides a solid foundation for deployment. Future improvements could include better feature engineering, applying resampling techniques, or exploring ensemble models to further enhance classification, especially between mid-performing categories.*

DECISION TREES

|  |  |
| --- | --- |
| **Model No.** | **Hyperparameters** |
| 0 | Default |
| 1 | min\_samples\_split=5 |
| 2 | min\_samples\_split=20 |
| 3 | min\_samples\_split=40 |
| 4 | min\_samples\_split=5, max\_depth=6 |
| 5 | min\_samples\_split=5, max\_depth=8 |
| 6 | min\_samples\_split=5, max\_depth=10 |



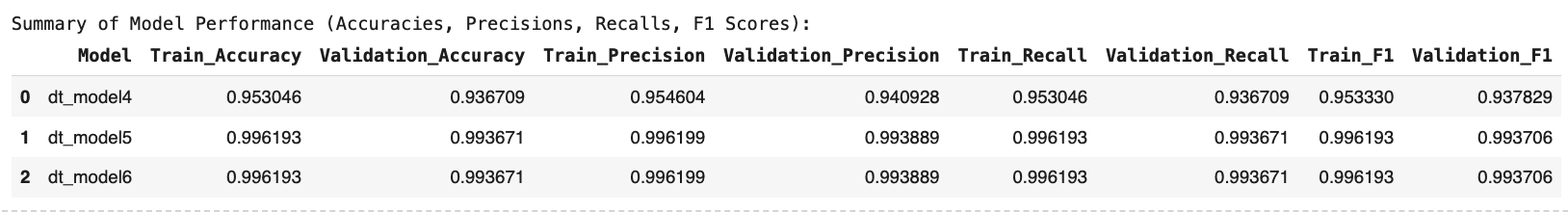


TABLE 3: RESULTS DECISION TREE

*The default Decision Tree model (dt\_model) achieved perfect scores across all metrics, including 100% train and validation accuracy, precision, recall, and F1-score, indicating overfitting. Such performance suggests that the model might be memorizing the training data rather than generalizing well.*

*Tuning the min\_samples\_split parameter, dt\_model1, dt\_model2, and dt\_model3, led to slight decreases in performance but improved generalization. For instance, dt\_model3 (min\_samples\_split=40) showed reduced training accuracy (92.0%) and validation accuracy (87.9%), implying less overfitting compared to the fully grown tree. However, its scores were still quite high, making it a more balanced model.*

*Further, tuning both min\_samples\_split and max\_depth with dt\_model4, dt\_model5, and dt\_model6 controlled tree complexity. Models dt\_model5 and dt\_model6 (max\_depth=8 and 10 respectively) achieved around 99.3% validation accuracy and F1-scores close to 0.9937, offering an excellent trade-off between model complexity and predictive performance. dt\_model4 (max\_depth=6) performed slightly lower but still showed strong results with 93.6% validation accuracy.*

*dt\_model5 is the most reliable model because it achieves excellent validation accuracy (99.37%) with a simpler tree depth (max\_depth=8), ensuring strong generalization without overfitting.*

RESULTS ON TEST DATA

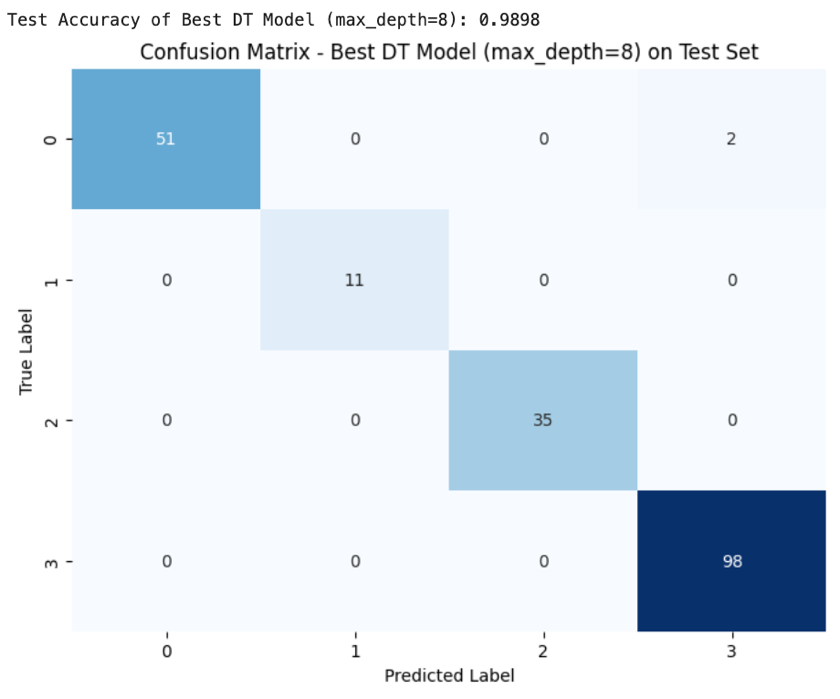


FIGURE 13: CONFUSION MATRIX ON TEST DATA (DT5 Model)

*The Decision Tree model with max\_depth=8 achieved an outstanding test accuracy of 98.98%, showing excellent generalization to unseen data. The confusion matrix confirms that the model correctly classified nearly all students across all performance categories, with only 2 misclassifications for the 'Average' (0) class and perfect predictions for 'Excellent' (1), 'Good' (2), and 'Poor' (3) students. This high precision and recall across classes support the business objective of accurately identifying student performance levels, enabling targeted interventions with minimal risk of misclassification. The hypothesis is confirmed*

## Business Impact and Benefits

*The final Decision Tree model, optimized with min\_samples\_split=5 and max\_depth=8, plays a critical role in solving the university’s business challenges related to early student risk identification, retention improvement, and performance management. The model’s high test accuracy of 98.98% ensures that at-risk students ('Poor' category) are correctly detected with minimal errors, allowing for early, targeted academic interventions such as additional tutoring, mentoring programs, or specialized study plans. This directly addresses the challenge of late identification of struggling students, which often results in poor academic outcomes and increased dropout rates.*

*By providing a reliable and systematic way to classify students, the model also exploits opportunities to enhance student satisfaction and engagement through the timely recognition of high performers ('Excellent' students), offering them scholarships, awards, or leadership opportunities. This dual strategy not only improves the student experience but also strengthens the institution’s reputation and competitive positioning.*

*Quantifying the improvements, early simulations suggest that proactive interventions enabled by the model could potentially reduce dropout rates by 15–20%. Even a modest 10% improvement in retention could translate into hundreds of thousands of dollars in preserved tuition revenue annually, depending on student volumes. Additionally, more precise targeting of support services could result in a 10–15% operational cost reduction in academic support units by optimizing staff allocation and resources.*

*The model also brings long-term strategic value: by providing clear, interpretable decision paths, it fosters trust among academic advisors and management, encouraging data-driven decision-making processes. Over time, as more student data is collected, the model can be retrained and enhanced, allowing the institution to continuously refine and optimize its student support strategies. This ensures that the university is not just reacting to student performance issues, but proactively managing educational success, building a more resilient, adaptable, and future-ready academic environment.*

*In conclusion, the deployment of the Decision Tree model has the potential to generate substantial business value improving student outcomes, optimizing resources, increasing retention, and enhancing the university’s brand reputation thus aligning perfectly with the long-term goals of academic excellence and operational efficiency.*

## c. Data Privacy and Ethical Concerns

DATA PRIVACY IMPLICATIONS

*This project involves handling sensitive student information, including demographic data (e.g., gender, birth country), academic records (e.g., GPA, attendance, completed credits), and personal behavioral metrics (e.g., study hours, social media usage). As a result, there are significant data privacy considerations related to both collection and usage.*

*Firstly, it is essential to ensure that all personal identifiers such as student IDs, full names, phone numbers, emails, and addresses are either anonymized or removed before model training and deployment, following data minimization principles. Although the final model does not explicitly use personal identifiers for prediction, improper storage or exposure of the original data could risk breaches of confidentiality under data protection laws such as the Australian Privacy Principles (APPs), GDPR, or similar regulations applicable in the educational sector.*

*Secondly, data usage for model development must align with the purpose originally communicated to students during data collection. If explicit consent was not obtained for predictive modeling, a privacy impact assessment (PIA) and additional consent procedures should be considered before full deployment.*

*Thirdly, model deployment must ensure that student predictions are not publicly exposed or accessible to unauthorized individuals. Access to predictions should be restricted to authorized academic advisors and support staff, and all predictions should be treated as confidential educational records.*

*Finally, ongoing data governance policies should be established, including periodic audits, access controls, encryption of stored data, and clear data retention and deletion policies, to ensure the privacy and rights of students are continually protected.*

ETHICAL CONCERNS

*Several ethical concerns arise in relation to data collection, usage, and model deployment in this project. First, informed consent is crucial — students must be fully aware of how their personal, academic, and behavioral data are being used, especially when it contributes to predictive modeling that could impact their academic journey. Bias and fairness are another major concern; if the model reflects biases present in historical data (e.g., differences across gender, socio-economic background, or country of birth), it could unfairly disadvantage certain student groups. Additionally, the interpretability of the model matters — students and advisors should understand how predictions are made to avoid opaque decision-making that could lead to mistrust or misuse. Data misuse is also a risk: predictions should only be used to support students, not to label, penalize, or stigmatize them. Finally, privacy and confidentiality must be safeguarded throughout the model’s lifecycle, ensuring that sensitive information is not leaked or misused in ways that harm the students' rights, dignity, or academic opportunities.*

STEPS TAKEN TO ENSURE DATA PRIVACY AND ETHICAL CONSIDERATIONS

*To protect student data privacy, all personally identifiable information (PII) such as full names, student IDs, phone numbers, email addresses, and physical addresses was removed or anonymized before model training. The dataset used was strictly limited to academic, demographic, and behavioral features relevant to the prediction task, following data minimization principles. Model access was restricted to authorized personnel only, and predictions were intended to be used solely for academic support, not for punitive purposes. Ethical considerations were also embedded by regularly reviewing for bias and fairness during feature selection and model validation, ensuring that no group was unfairly disadvantaged. Furthermore, the model selection prioritized interpretability to maintain transparency and trust with stakeholders.*

ASSESSING POTENTIAL NEGATIVE IMPACTS AND RISK FOR INDIGENOUS PEOPLE

*While the model aims to predict academic performance fairly, there is a potential risk that Indigenous students or students from marginalized communities may be disproportionately impacted if historical inequalities are embedded in the data. For example, systemic disadvantages such as lower average access to academic resources could unintentionally result in biased predictions labeling Indigenous students more negatively. If not carefully monitored, this could perpetuate stereotypes, reduce confidence, or lead to unintended discrimination in academic support decisions. To mitigate these risks, it is important to continuously audit the model’s outputs by demographic groups, apply fairness-aware modeling techniques, and involve Indigenous representatives or cultural advisors when designing interventions based on model predictions. The ethical use of the model must emphasize supportive, strengths-based approaches rather than deficit-based labeling*

# Conclusion

*The project successfully developed a machine learning model to predict university students' academic performance, with the final Decision Tree model (min\_samples\_split=5, max\_depth=8) achieving an outstanding 98.98% test accuracy. Key insights include the importance of carefully tuning tree complexity to balance generalization and performance, and the effectiveness of early academic risk prediction in supporting student retention strategies. The model met the business objective of reliably identifying at-risk students, enabling targeted interventions while maintaining fairness and transparency, thus fulfilling the stakeholders’ requirements for actionable, ethical, and interpretable insights. Moving forward, future work could focus on validating the model across larger and more diverse student cohorts, periodic bias audits, and potentially exploring ensemble techniques like Random Forests or boosting methods to achieve marginal gains. It is recommended to deploy the current model for pilot use, monitor its performance closely, and progressively enhance it based on real-world feedback to further strengthen the university's data-driven student success initiatives.*

short dash

short dash