**A Holistic Machine Learning Pipeline for Academic Outcome Prediction**

An exploratory study using EDA, Statistical Analysis, and Machine Learning

**Author:**

Selvi Ganesh Kumar

**Institution:**

CCT College

Dublin 2

Ireland

**Course:**

Higher Diploma in Science in Data Analytics

**Supervisor:**

Dr. Taufique Ahmed

**Date of Submission:**

18th of May 2025

**Version:**

Final Report

**Contact:**  
Email: [selvi.digitalsolutions@gmail.com]  
GitHub: github.com/research-selvi-datascience

*© 2025 Selvi Ganesh Kumar. All rights reserved.*

**Table of Contents**

|  |  |  |  |
| --- | --- | --- | --- |
| **Section** | **Title** | **Subsections** | **Page No:** |
| 1 | Introduction | * 1. Background | 5 |
|  |  | 1.2 Objective of the study | 5 |
|  |  | 1.3 Research questions | 5 |
|  |  | 1.4 Scope and Limitations | 6 |
| 2 | Methodology | 2.1 CRISP-DM Overview | 6 |
|  |  | 2.2 Data Preparation and Statistical analysis Pipeline | 6 |
|  |  | 2.3 Model Development and Experimentation | 7 |
| 3 | Data Description and EDA | 3.1 Dataset Overview | 9 |
|  |  | 3.2 Data Sources and Ethical considerations | 9 |
|  |  | 3.3 Descriptive Statistics | 9 |
|  |  | 3.4. EDA | 10 |
| 4 | Data Preprocessing | 4.1 Normality Analysis | 15 |
|  |  | 4.2 Dimensionality Reduction: PCA & LDA | 16 |
|  |  | 4.3 Class Imbalance & SMOTE | 16 |
|  |  | 4.4 Unsupervised Outlier Detection with “Isolation Forest” | 17 |
|  |  | 4.5 Multicollinearity and Feature engineering | 18 |
|  |  | 4.6 Feature Importances & Explainability tools – SHAP, LIME, RF Feature Importances | 19 |
|  |  | 4.7 Clustering Analysis | 21 |
| 5 | Statistical Tests Summary | 5.1 Parametric Tests | 23 |
|  |  | 5.2 Non- Parametric Tests | 23 |
| 6 | Model Selection and Training | 6.1 Model Categories and Algorithms | 24 |
|  |  | 6.2 Evaluation Metrics | 24 |
| 7 | Multiclass vs. Binary Classification strategy |  | 25 |
| 8 | Experiments and Model Performance Evaluation | 8.1 Experiment 1: Baseline Modeling (75:25 Split) | 25 |
|  |  | 8.2 Experiment 2: Baseline Modeling (90:10 Split) | 26 |
|  |  | 8.3 Experiment 3: Dimensionality Reduction (PCA & LDA) | 27 |
|  |  | 8.4 Experiment 4: Imbalanced Data handling (SMOTE, ADASYN, NearMiss) | 29 |
|  |  | 8.5 Experiment 5: Feature Selection Using Random Forest Importances | 31 |
|  |  | 8.6 Experiment 6: Boosting models + SMOTE | 31 |
|  |  | 8.7 Experiment 7: Multicollinearity + Feature engineering | 32 |
|  |  | 8.8 Experiment 8: Binary Classification | 33 |
|  |  | 8.9 Experiment 9: Binary Classification + SMOTE for Traditional + Boosting models | 34 |
| 9 | Cross-Experiment Summary | 9.1 Multiclass classification top results | 37 |
|  |  | 9.2 Binary Classification top results | 37 |
|  |  | 9.3 Multiclass (Non-SMOTE) | 38 |
|  |  | 9.4 Multiclass (SMOTE) | 38 |
|  |  | 9.5 ROC Curve & Precision-Recall curve | 38 |
|  |  | 9.6 Binary (Non-SMOTE) | 40 |
|  |  | 9.7 Binary (SMOTE) | 41 |
| 10 | Model Evaluation | 10.1 Best models across all experiments | 42 |
|  |  | 10.2 Recommendation | 43 |
|  |  | 10.3 Practical Recommendations | 43 |
|  |  | 10.4 Challenges | 43 |
|  |  | 10.5 Future Work | 44 |
|  |  | 10.6 Conclusion | 44 |
| 11 | References |  | 45 |

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Title** | **Page No.** |
| 1 | CRISP-DM Methodology | 6 |
| 2 | Curricular Units Enrolled, Evaluations, Approved – 1st and 2nd Semester | 10 |
| 3 | Curricular Units Grade – 1st and 2nd Semester | 11 |
| 4 | Courses vs Target Classes – Graduate, Dropout, Enrolled | 11 |
| 5 | Age vs Target Classes – Graduate, Dropout, Enrolled | 12 |
| 6 | Students' Financial Factors vs Target Classes – Graduate, Dropout, Enrolled | 12 |
| 7 | Students' Grades vs Target Classes – Graduate, Dropout, Enrolled | 13 |
| 8 | Outliers – Curricular Units: 1st and 2nd Semester | 13 |
| 9 | Isolation Forest Anomalies vs Normal Pattern in Curricular Units | 14 |
| 10 | Correlation Analysis – Features vs Target Classes | 15 |
| 11 | PCA Projection | 16 |
| 12 | LDA Projection | 16 |
| 13 | Anomaly Analysis – 3D Scatterplot | 18 |
| 14 | SHAP Results | 19 |
| 15 | SHAP Results – Graduate | 19 |
| 16 | SHAP Results – Dropout | 20 |
| 17 | SHAP Results – Enrolled | 20 |
| 18 | LIME Results | 20 |
| 19 | Visualization – KMeans Clustering | 22 |
| 20 | SVM – Evaluation Metrics | 38 |
| 21 | SVM – Evaluation Metrics | 38 |
| 22 | SVM – ROC Curve & PR Curve | 39 |
| 23 | LightGBM – ROC Curve & PR Curve | 40 |
| 24 | SVM – Evaluation Metrics | 40 |
| 25 | SVM – ROC & PR Curve | 41 |
| 26 | SVM – Evaluation Metrics | 42 |
| 27 | Top Models (Multiclass + Binary) Class-wise F1 Scores | 42 |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Title** | **Page No.** |
| 1 | Workflow: Data Preparation and Statistical Analysis | 7 |
| 2 | Workflow: Experimental Analysis – Multiclass Model | 8 |
| 3 | Experimental Analysis – Clustering & Binary Classification Workflow | 9 |
| 4 | ML Prediction Results – Prior Resampling | 17 |
| 5 | ML Prediction Results – Post Resampling | 17 |
| 6 | RF Feature Importances | 21 |
| 7 | Statistics – Parametric Tests | 23 |
| 8 | Statistics – Non-Parametric Tests | 23 |
| 9 | Statistics – Evaluation Metrics | 24 |
| 10 | Performance of the Models – Experiment 1 | 25 |
| 11 | Performance of the Models – Experiment 2 | 25 |
| 12 | Experiment 3: Results – LDA with Default Parameters | 27 |
| 13 | Experiment 3: Results – LDA with Hyperparameters | 28 |
| 14 | Experiment 3: Results – PCA with Default Parameters | 28 |
| 15 | Experiment 3: Results – PCA with Hyperparameters | 29 |
| 16 | Experiment 4 Results – RFC + Balancing Techniques | 29 |
| 17 | Experiment 4 Results – ML (Default) + SMOTE | 30 |
| 18 | Experiment 4 Results – ML (Hyperparameters) + SMOTE | 30 |
| 19 | Experiment 5 Results – RF Feature Importances | 31 |
| 20 | Experiment 6 Results – SGD + SMOTE | 31 |
| 21 | Experiment 6 Results – XGBoost + SMOTE | 31 |
| 22 | Experiment 6 Results – CatBoost + SMOTE | 32 |
| 23 | Experiment 6 Results – LightGBM + SMOTE | 32 |
| 24 | Experiment 7 Results – ML (Default) with Feature Engineering | 33 |
| 25 | Experiment 8 Results – ML (Default) for Binary Classification | 33 |
| 26 | Experiment 8 Results – ML (Hyperparameters) for Binary Classification | 34 |
| 27 | Experiment 9 Results – ML (Default) + SMOTE | 35 |
| 28 | Experiment 9 Results – ML (Hyperparameters) + SMOTE | 35 |
| 29 | Experiment 9 Results – Boosting + SMOTE | 36 |
| 30 | Top Results – Multiclass Classification | 37 |
| 31 | Top Results – Binary Classification | 37 |

**Abstract**  
This study explores various machine learning models for predicting student dropout and academic success using a real-world higher education dataset. Eleven experiments were conducted, each employing varied preprocessing techniques and modelling strategies. Dimensionality reduction methods such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) were explored. To address class imbalance, resampling techniques including SMOTE, ADASYN, and NearMiss were applied, with SMOTE demonstrating superior performance. Six traditional algorithms (Logistic Regression, Support Vector Machine, Random Forest, Neural Networks, Decision Trees, and K-Nearest Neighbors) were rigorously evaluated under both default and hyperparameter-tuned settings (using GridSearchCV) across SMOTE and non-SMOTE datasets. Boosting algorithms (Gradient Boosting, XGBoost, CatBoost, and LightGBM) were also tested, with hyperparameter tuning performed using Optuna.

Results showed that LightGBM and CatBoost models (all tuned) on SMOTE datasets, outperformed traditional models in multiclass classification. Outlier detection was conducted using Isolation Forest. Clustering techniques (KMeans, KMedoids) were applied after dimensionality reduction with PCA, t-SNE and Autoencoders. It revealed patterns based on academic and financial factors, significant overlap between clusters led to only moderate separation (silhouette score ≈ 0.41). So, binary classification experiments (Graduate vs. Dropout) was conducted, where traditional models on non-SMOTE datasets slightly outperformed boosting algorithms.

**1. Introduction:**

**1.1 Background**  
Student dropout rates are a significant concern for higher education institutions globally. Beyond individual consequences, dropout impacts economic growth, employment, and societal competitiveness, affecting students, families, institutions, and communities alike (Palomino & Ortega, 2023). The complexity of this issue—spanning academic, socio-economic, and personal factors—has made it a prominent research topic since the 1930s. Institutions strive to reduce dropout rates to minimize financial burdens, increase graduation rates, and ensure efficient educational investments (Manrique et al., 2019b).

**1.2 Objective of the study:**

Developing early warning systems based on predictive analytics can significantly aid in detecting students at risk of dropping out and in delivering targeted support.

**1.3 Research questions:**

* **RQ1:** Which machine learning algorithms are most effective in predicting student dropout?
* **RQ2:** How effective are different resampling techniques (SMOTE, ADASYN, NearMiss) in addressing class imbalance for dropout prediction?
* **RQ3:** How do boosting algorithms (Gradient Boosting, XGBoost, CatBoost, LightGBM) compare to traditional ML models in predicting dropout?
* **RQ4:** What key factors contribute to student success or failure, as identified by SHAP, LIME, and RF Feature Importance?
* **RQ5:** Can unsupervised ML techniques like clustering effectively reveal patterns or groupings within the dataset?

Through addressing these questions, this project aims to enhance the predictive modeling of student outcomes to support institutions.

**1.4 Scope and Limitations**

While constrained by a single institutional dataset, this work highlights effective modeling and preprocessing strategies in educational prediction tasks, offering valuable insights for future research and practical deployment. (Villar & Andrade, 2024)

**2. Methodology**

The approach is guided by the CRISP-DM framework (Cross-Industry Standard Process for Data Mining), a structured and iterative methodology for predictive analytics.

**2.1 CRISP-DM Overview**

The CRISP-DM process consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This project followed these phases systematically, enabling a robust foundation for model development and interpretation.



***Figure 1: CRISP-DM Methodology***

**2.2 Data Preparation and Statistical analysis Pipeline:**

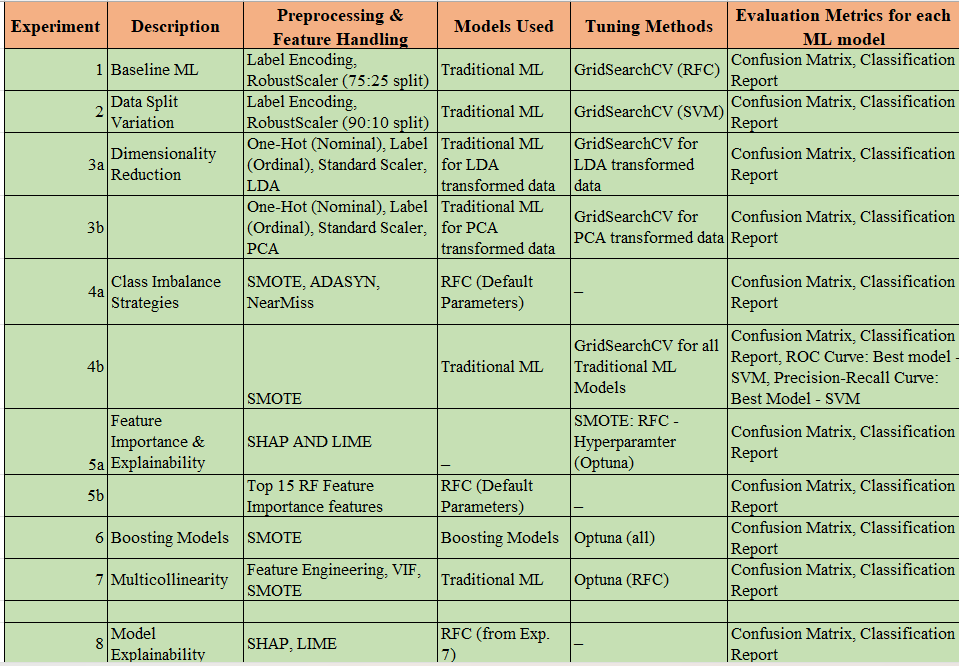
Table 1 summarizes the data preparation steps, which are further detailed in Section 4.

**Workflow: Data Preparation and Statistical Analysis:**

***(Table 1: Workflow: Data Preparation and Statistical Analysis)***

**2.3 Model Development and Experimentation : Summary of Experiments**

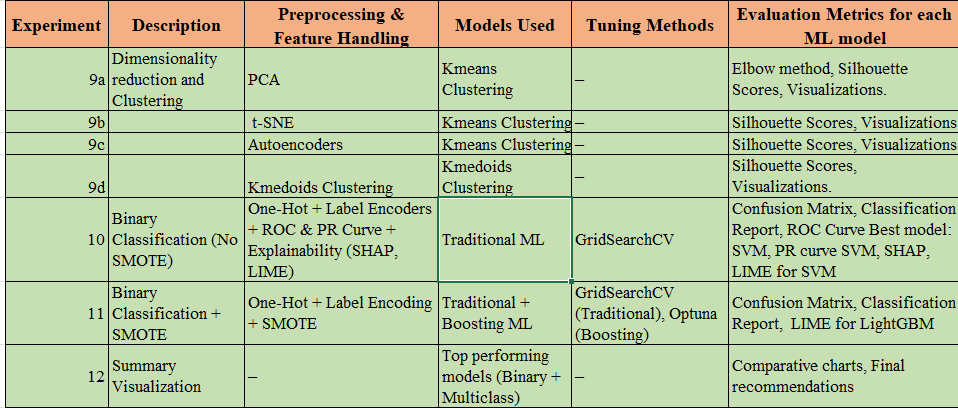
A categorized overview of the prediction experiments conducted is presented below and detailed in Section 4 & 8.

**Workflow: Experimental Analysis – Multi-Classification Model :**

***(Table 2: Workflow: Experimental analysis - Multiclass model)***

**Clustering Analysis and Binary Classification:**

Following clustering analysis, the classification task was restructured as binary to enhance interpretability and address performance trade-offs.

**Workflow: Experimental Analysis – Clustering & Binary classification**

***(Table 3: Experimental Analysis – Clustering & Binary classification Workflow)***

**3. Data description and Exploratory Data Analysis:**

**3.1 Dataset Overview:**

The dataset includes 4,424 records and 37 variables spanning demographic, academic, financial, socioeconomic, and macroeconomic indicators, used to predict student outcomes — categorized as “Dropout,” “Enrolled,” or “Graduate.” Details are provided in Subsections 3.3 and 3.4.

**3.2 Data Sources and Ethical considerations:**

The dataset for this Capstone project is sourced from the UC Irvine Machine Learning Repository, available at the following link: [Predict Students' Dropout and Academic Success - UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Student+Performance). It includes 4,424 records with 35 attributes related to undergraduate students' demographics, socioeconomic factors, and academic performance from 2008 to 2019. It was created by Valentim Realinho, Mónica Vieira Martins, Jorge Machado, and Luís Baptista from the Instituto Politécnico de Portalegre and is licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license (archive.ics.uci.edu, n.d.).

**Ethical Considerations**

This study prioritizes data privacy, ensuring that the dataset is anonymized and free of personally identifiable information (PII), in compliance with regulations like GDPR. The dataset adheres to FAIR principles (Findability, Accessibility, Interoperability, and Reusability), promoting transparency and ethical reuse. Furthermore, the project maintains best practices in data science, focusing on fairness, accountability, and avoiding bias in model predictions, ensuring the ethical application of the data.

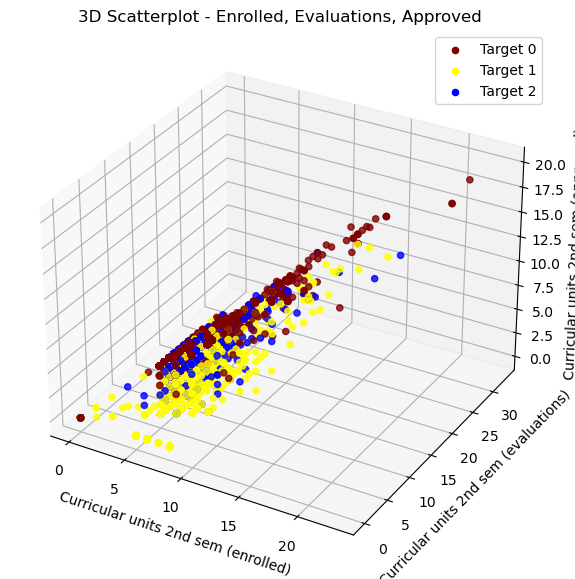
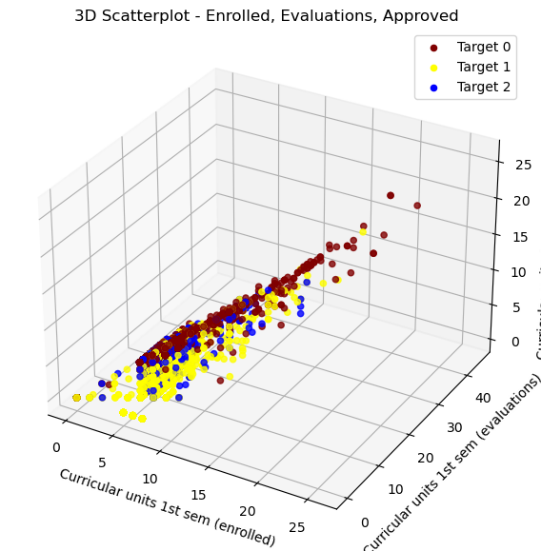
**3.3 Descriptive Statistics Summary**

* **Demographics:** Majority of students are Portuguese, single, and female.
* **Academic Background:** Most students have secondary education and an average admission grade of 126.98.
* **Course Enrollment:** Nursing is the most popular course. Most students attend daytime classes (3941).
* **Curricular Units:** Students enrol in many units (average 6 to 7 in the 1st & 2nd semester). Majority of the students complete their evaluations and get approved, while a few do less evaluations and struggle to get approved and graded. This shows the dropouts pattern, reflecting challenges in meeting standards.
* **Financial Indicators:** Most students are not in debt (4373), tuition is up to date, and few hold scholarships.
* **Economic Factors:** Unemployment, inflation, and GDP provide context for external pressures influencing dropout rates.
* **Target Variable:** "Graduate" is the most frequent class (2209), highlighting successful outcomes for most students.

**3.4. Exploratory Data Analysis:**

**1. Curricular Units for 1st & 2nd Semesters**:

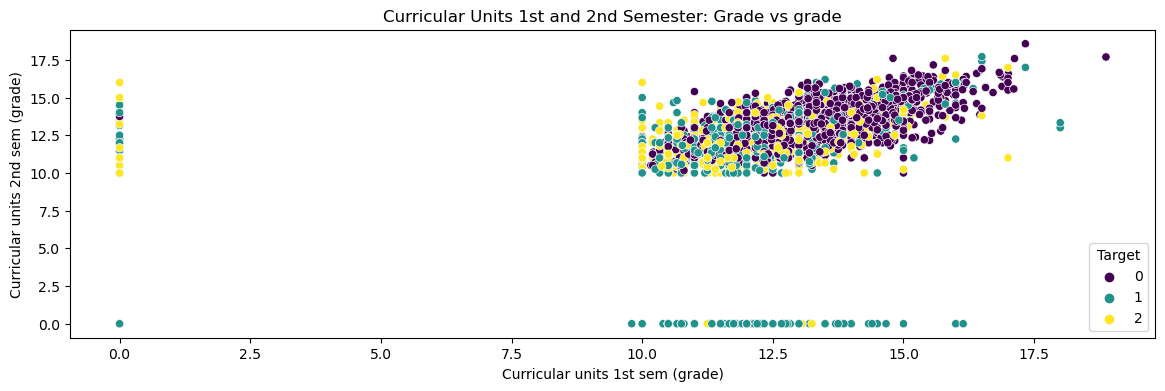
* **Graduates**: Smooth progression from enrolment to evaluation and approval.
* **Dropouts**: Minimal progression, with few evaluations advancing to approval.
* **Enrolled**: Progress from enrolment to evaluation but low approval rates.



***Figure 2: Curricular Units Enrolled, Evaluations, Approved – 1st and 2nd Semester***

**2. Grade for 1st and 2nd Semester:**

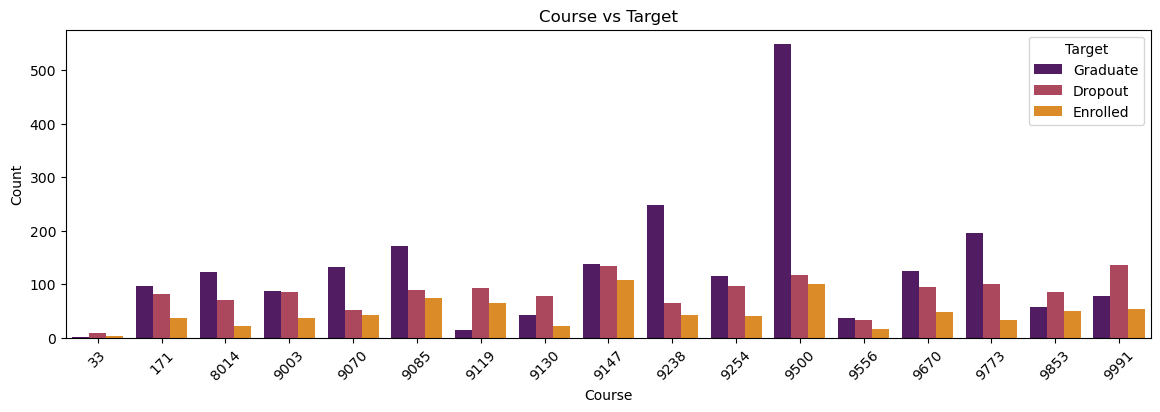
Graduates tend to have higher grades in both semesters. Some Dropouts exhibit good grades in the first semester but drop in the second semester. Enrolled students show mixed performance.



***Figure 3: Curricular Units Grade – 1st and 2nd Semester***

**3. Courses:**

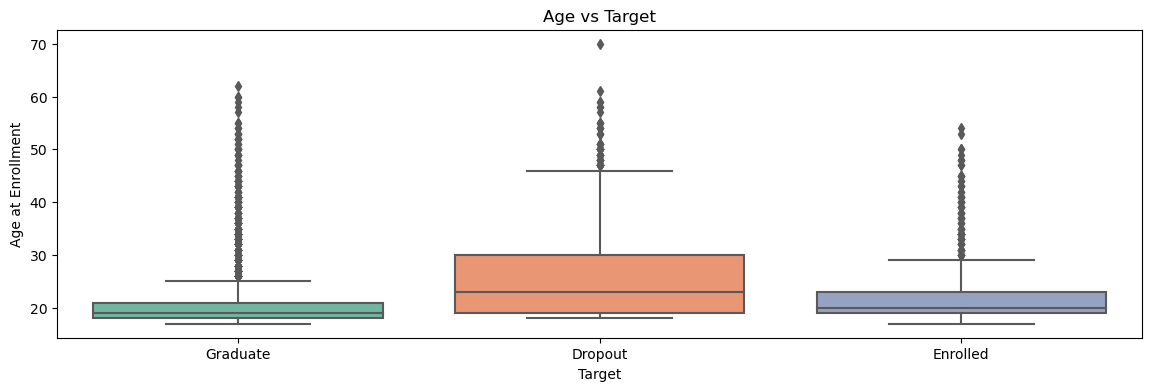
* **Nursing (9500):** Most graduates.
* **9238 (Social Service) , 9773 (Journalism and Communication):** High Graduates
* **9853 (Basic Education) , 9991 (Management – Evening attendance) :** High Dropouts and less graduates and enrolled



***Figure 4: Courses vs Target Classes – Graduate, Dropout, Enrolled***

**4. Age:**

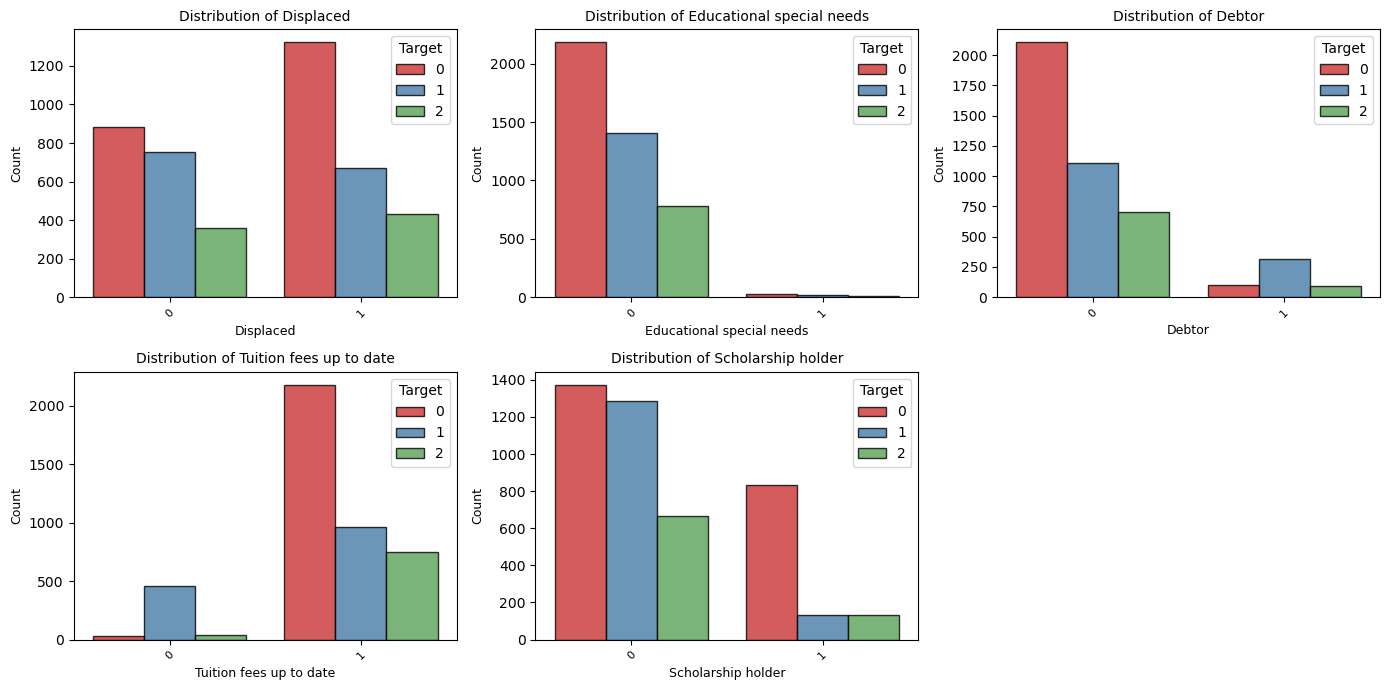
* **Graduates:** Predominantly aged 18-23, suggesting younger students are more likely to graduate.
* **Dropouts:** Older students (25-30+) may face external pressures, leading to higher dropout rates.



***Fig 5: Age vs Target classes – Graduate, Dropout, Enrolled***

**5. Financial & Other Factors:**

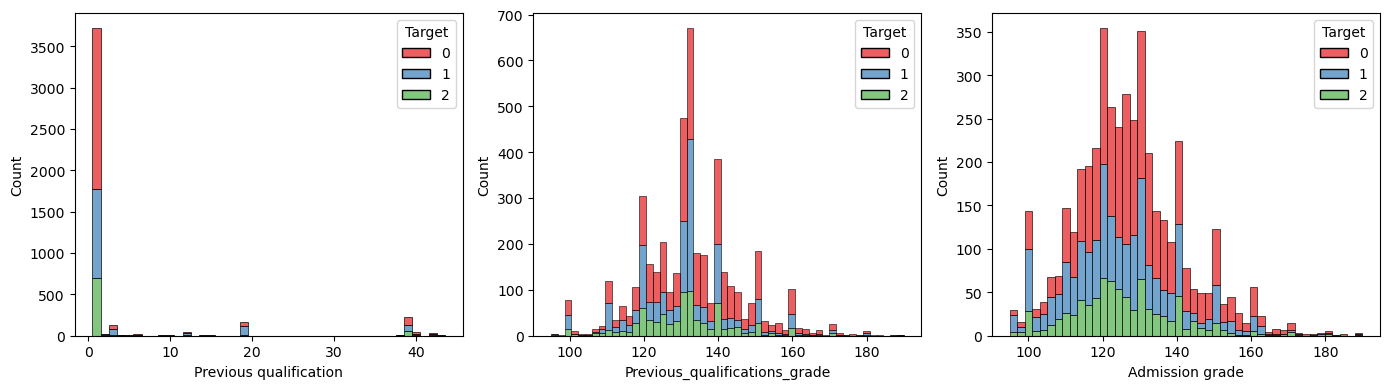
* **Displaced Students:** More graduates are displaced than non-displaced.
* **Debtor:** Higher debt proportion in dropouts compared to graduates.
* **Tuition Fees:** Most graduates have up-to-date tuition; some dropouts face financial issues.
* **Scholarships:** Graduates are more likely to hold scholarships.



***Figure 6: Students Financial Factors vs Target classes – Graduate, Dropout, Enrolled***

**6. Grades:**

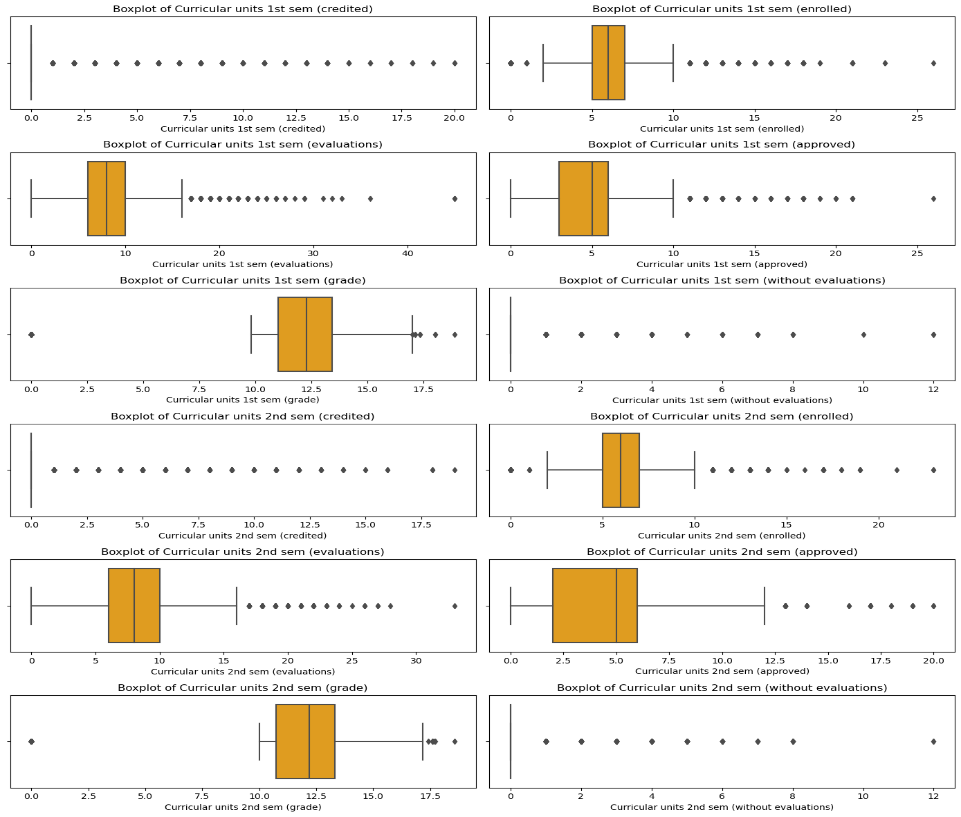
* **Previous Qualification:** Balanced grade distribution across all groups.
* **Admission Grade:** Students with lower admission grades (around 100) have higher dropout rates, indicating academic challenges.



***Figure 7: Students grades vs Target classes – Graduate, Dropouts, Enrolled***

**7. Outliers:**

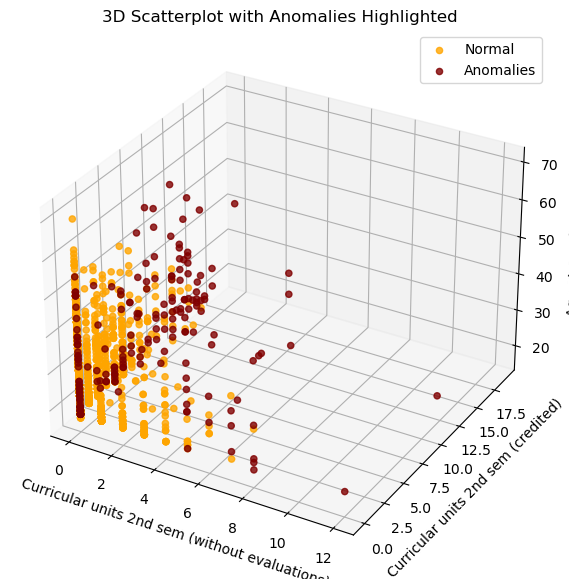
Outliers in evaluations, approvals, and grades reflect at-risk students or high achievers. Removing them could skew results, so they are retained for a comprehensive analysis.



***Figure 8: Outliers - Curricular Units: 1st and 2nd Semester***

**8. Anomalies:**

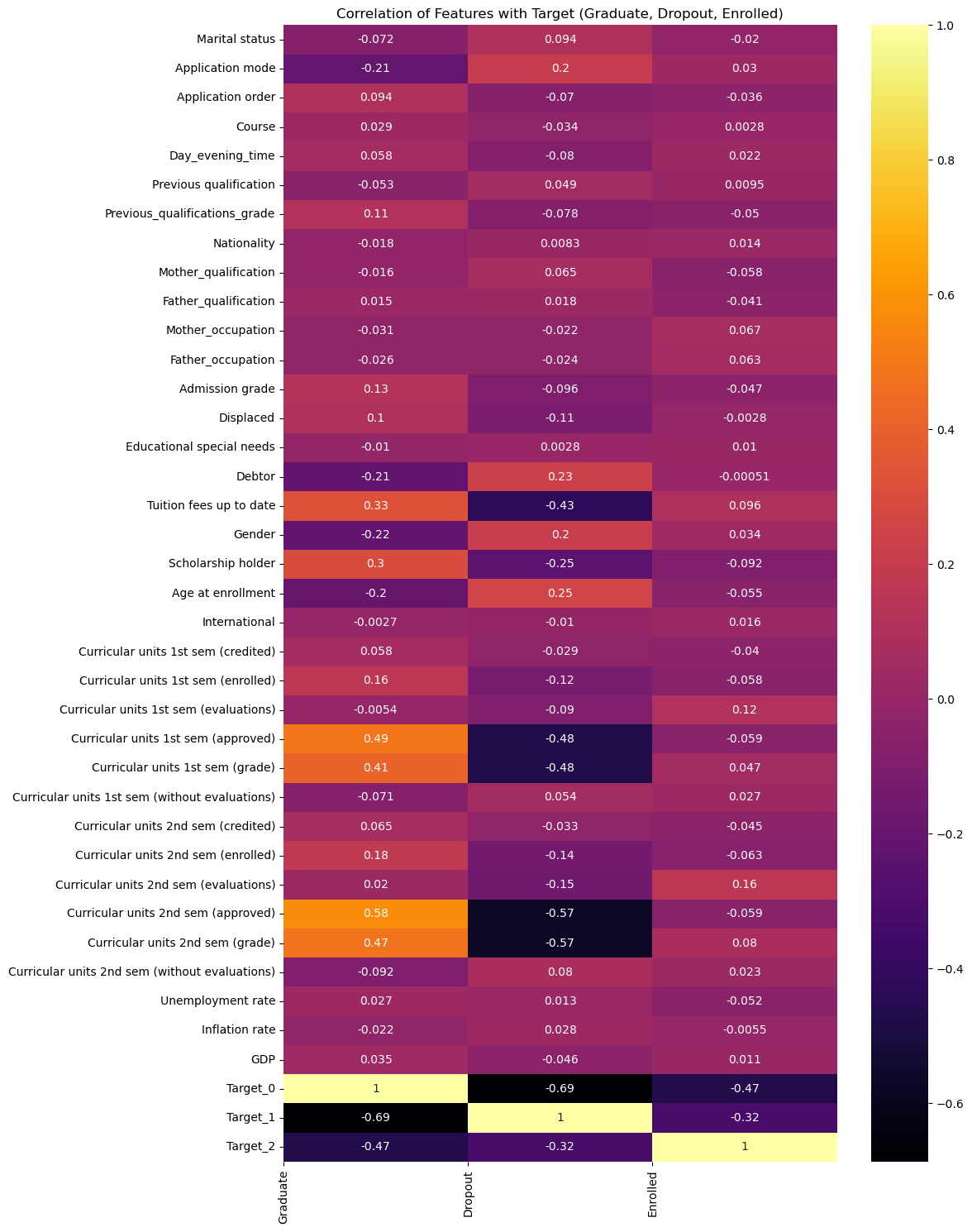
**IsolationForest** detected 195 anomalies (4.4%) due to extreme values in dropout and graduate classes, age, or nationality. These are kept to maintain data integrity.



***Figure 9: Isolation Forest Anomalies vs Normal pattern in Curricular units***

**9. Correlation Analysis: Features vs Target classes:**

Target\_0 (Graduate) is most strongly influenced by Curricular units 1st sem and 2nd Sem (approved) and Grade and the Financial factors. Target\_1 (Dropout) tends to have an inverse relationship with of Target\_0 (Graduate). Target\_2 has very weak correlations with most academic factors, but it is slightly positively related to Curricular units 1st and 2nd sem (evaluations). Weak correlations observed with other features.



***Figure 10: Correlation Analysis: Features vs Target classes***

**4. Data Preprocessing:**

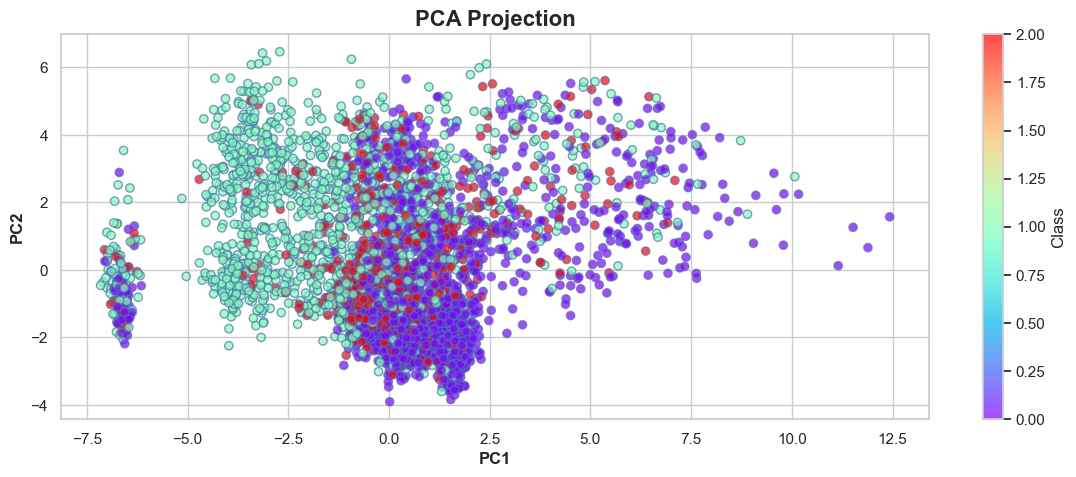
Key preprocessing techniques are elaborated in the following subsections.

**4.1 Normality Analysis: Histogram, Shapiro-Wilk Test, Box-Cox Transformation:**

Histograms & Shapiro-Wilk tests revealed all numerical features deviated from normality (p < 0.05). Box-Cox transformations were applied to reduce skewness, improving normality marginally but not fully correcting it. Q-Q plots post-transformation showed better alignment (e.g., Age, Grades), but curricular unit features remained skewed. Perfect normality is rare in real-world datasets; however, the Box-Cox transformation effectively reduced skewness and stabilized variance, aiding model assumptions.

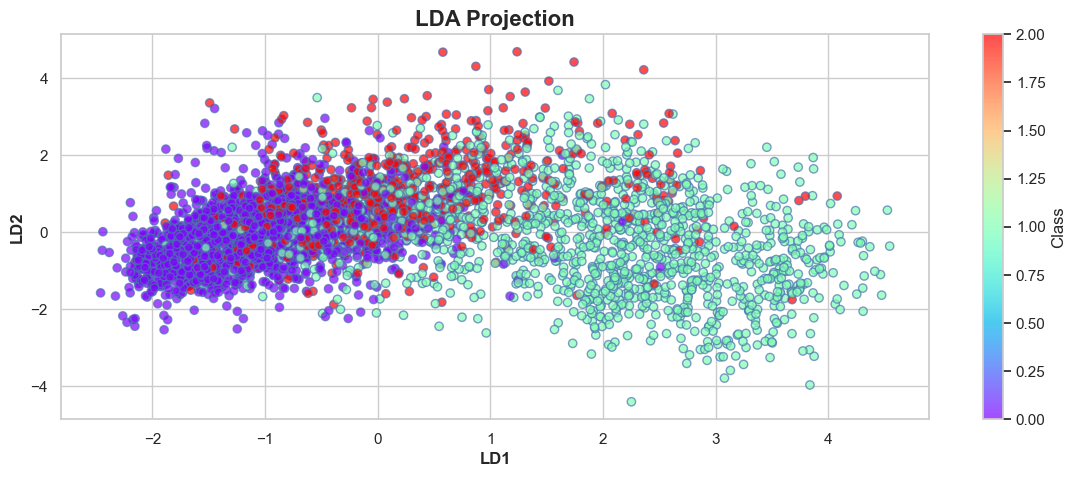
**4.2 Dimensionality Reduction: PCA & LDA:**

PCA and LDA were used to reduce the dataset's dimensionality and visualize class separability. PCA separated the Graduate and Dropout classes to certain degree but failed to clearly separate Enrolled class due to its overlapping distributions.



***Figure 11: PCA Projection***

**LDA** showed better separation between Graduate and Dropout classes, in comparison with PCA, but significant overlap remained with the Enrolled class, indicating that linear projections are insufficient for clear boundaries.



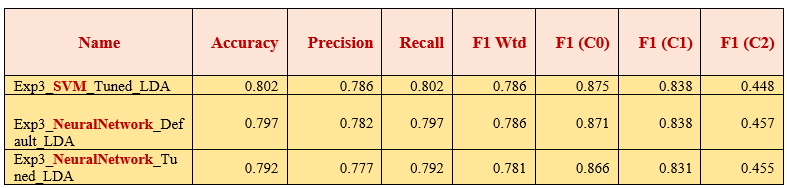
***Figure 12: LDA Projection***

When LDA-transformed data was used, machine learning models such as KNN, SVM, Logistic Regression, and Neural Networks performances improved significantly, achieving overall accuracy of ~80%, with the Enrolled class F1 score reaching ~45%, demonstrating the value of LDA in enhancing class separability for predictive modeling.

**4.3 Class Imbalance & SMOTE**

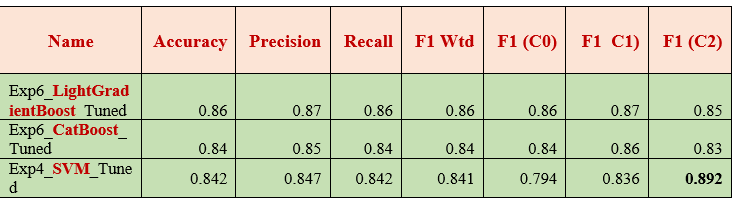
Class distribution: Graduate (2209 - 49.9%), Dropout (1421 - 32%), and Enrolled (794 - 17.9%). This imbalance negatively affected machine learning performance, especially for the minority Enrolled class, which consistently underperformed across all models with F1-scores in the 0.30s.

Traditional ML models performed well on the majority classes (Graduate & Dropout), with overall accuracy in the 0.70s, but failed to generalize to Class 2. Even after applying preprocessing techniques like PCA, LDA, and encoding, performance gains were marginal, and class imbalance persisted.



***Table 4: ML Prediction results – Prior resampling***

Three resampling strategies were explored: SMOTE, ADASYN, and NearMiss. Among them, SMOTE produced the most balanced results. SMOTE significantly improved F1 scores for the minority Enrolled class, with SVM achieving 0.89 and overall accuracy of 84%. These results underscore the importance of resampling in multiclass tasks. Boosting models achieved higher and more stable performance.



***Table 5: ML Prediction results – Post resampling***

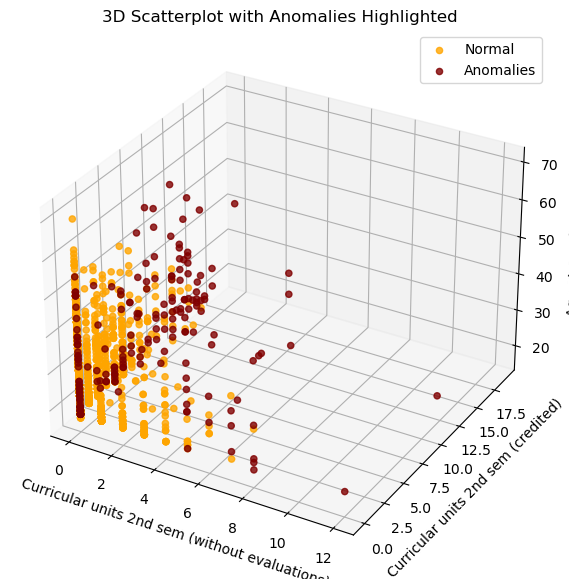
In the binary classification experiments, the class imbalance was comparatively less severe. The tuned SVM model (Non-SMOTE) achieved an accuracy of 95%, with strong and balanced class-wise performance. CatBoost and LightGBM (with SMOTE) reached accuracies of 93% and 92% respectively, again with consistent class-wise F1 scores. In summary, while SMOTE proved essential for multiclass classification, but less critical in this binary setting.

**4.4 Unsupervised Outlier Detection with “Isolation Forest”**

The Isolation Forest algorithm identified approximately 4.4% of the instances as anomalies.

**Class-wise anomalies:**

* Class 0 (Graduate): 86 anomalies
* Class 1 (Dropout): 84 anomalies
* Class 2 (Enrolled): 25 anomalies



***Figure 13: Anomaly analysis – 3D scatterplot***

**Anomaly Analysis:**

The anomalies were analyzed in terms of key features:

* Curricular Units without Evaluations: Subjects without evaluations are flagged as anomalies, as most subjects require evaluations for grades.
* Credited Units: Units without credit transfers may also be flagged as anomalies.

These anomalies are crucial for understanding student outcomes as they can differentiate between these groups. Removing these anomalies might lead to biased results. Rather than removing anomalies, transformations were applied to mitigate their skewing effects, maintaining their contribution to outcome variability.

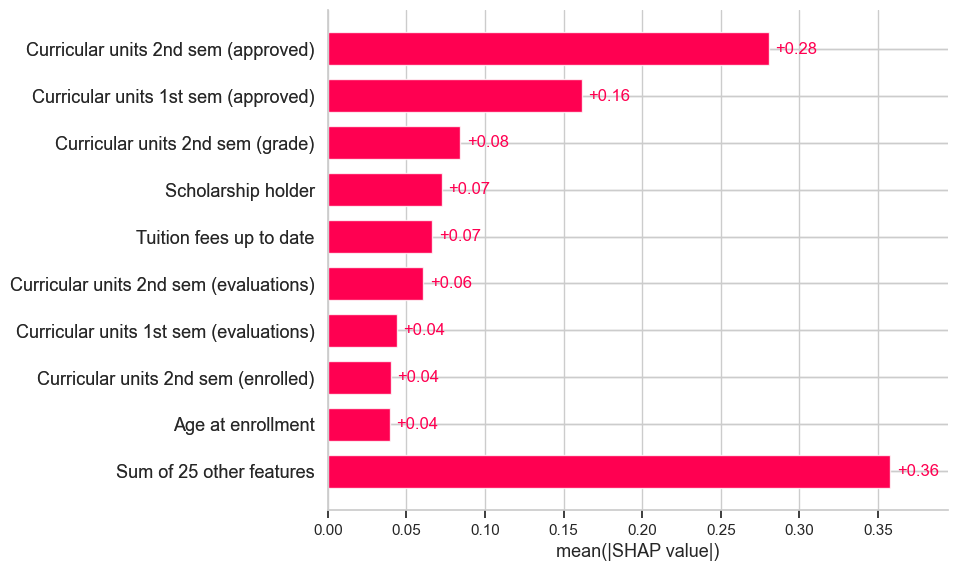
**4.5 Multicollinearity and Feature engineering:**

1. Variance Inflation Factor (VIF) exposed multicollinearity among semester-wise curricular unit features (VIF > 10).
2. Solution: Feature Engineering: Highly correlated semester features were combined into new features: Total Enrolled, Total Approved, Total Credited and the original features were dropped to minimize redundancy.
3. Post-Feature Engineering:Random Forest Classifier (Tuned) achieved 83% with balanced class F1 scores.

**4.6 Feature Importances & Explainability tools:**

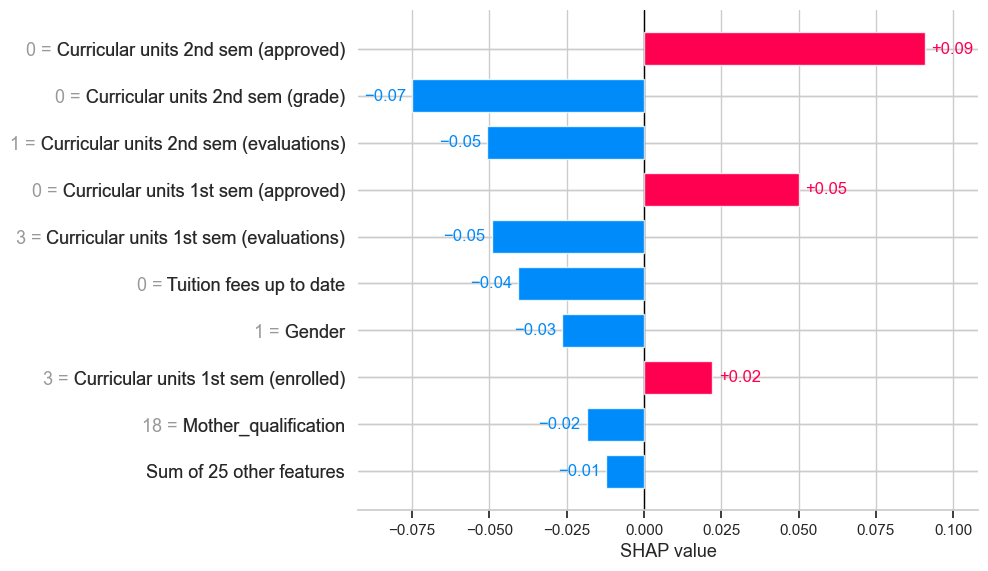
**SHAP (SHapley Additive exPlanations):**

SHAP was applied to the RFC experiment. The result showed the key drivers are the academic (+0.66), and financial factors (0.14) with lesser contribution of few demographic factors.



***Figure 14: SHAP Results***

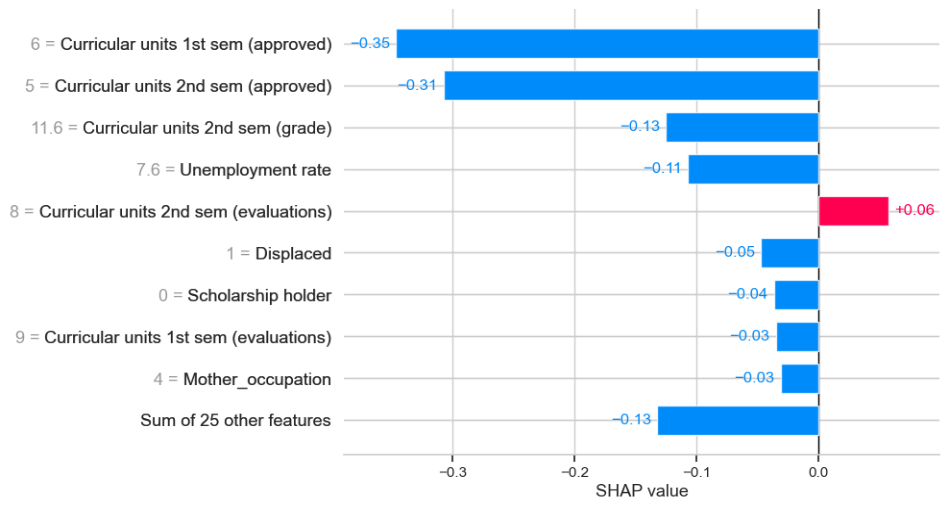
**SHAP - explanation for individual classes**: Overall, the academic performance and financial factors were the most decisive factor, with other variables contributing minimally.

**SHAP Explanation for class: Graduate**

***Figure 15: SHAP Results - Graduate***

**SHAP Explanation for class: Dropout**

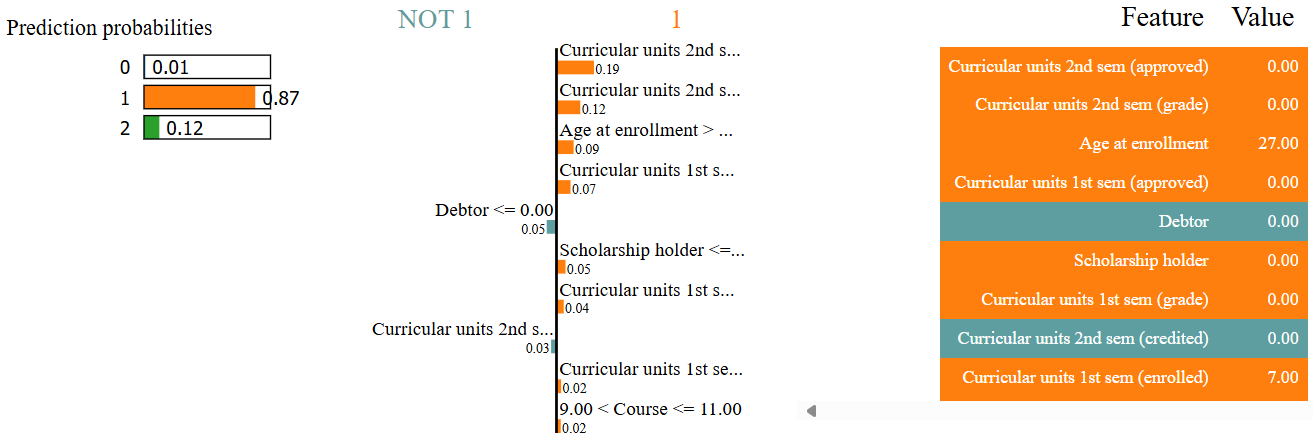
***Figure 16: SHAP Results - Dropout***

**SHAP Explanation for class: Enrolled**

***Figure 17: SHAP Results - Enrolled***

For the Enrolled class, SHAP values revealed that academic evaluations were the strongest positive contributors. However, overlapping features with Dropout and Graduate classes led to misclassifications, lowering the F1 score for Enrolled.

**LIME - LIME (Local Interpretable Model-agnostic Explanations):**

****

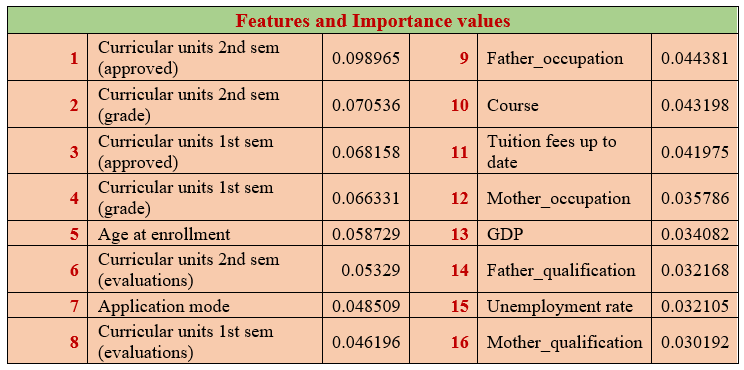
***Figure 18: LIME Results***

Random Forest Classifier predicted the student as Dropout with high confidence (87% probability). The key features were academic factors for both semesters, financial factors and minimal demographic factors contributed to the prediction.

The best-performing models (LightGBM, Random Forest) consistently highlighted the same key features.

**Random Forest Feature Importances and Performance**

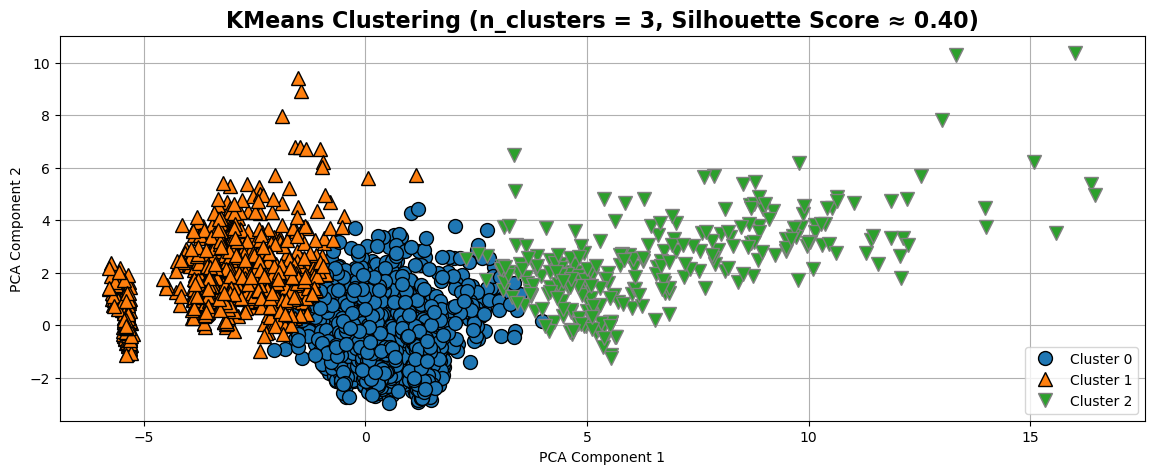
The top 15 features based on feature\_importances\_ closely align with the SHAP and LIME results across different models, strengthening the conclusion about dominant predictors.



***Table 6: RF Feature importances***

**4.7 Clustering Analysis: (Class Pattern Discovery)**

To explore inherent patterns in the data beyond supervised learning, unsupervised clustering (K-Means, K-Medoids) was applied following dimensionality reduction (PCA, t-SNE, Autoencoders)

K-Means and K-Medoids methods yielded moderate silhouette scores (~0.40), indicating weak clustering. 

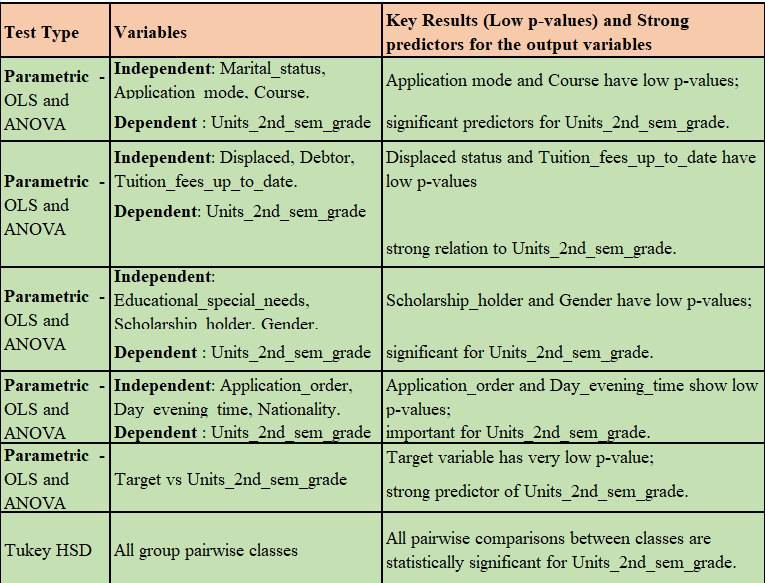
***Figure 19: Visualisation – KMeans Clustering***

Visualizations showed significant overlap between clusters and true classes due to shared academic and financial patterns across student types.

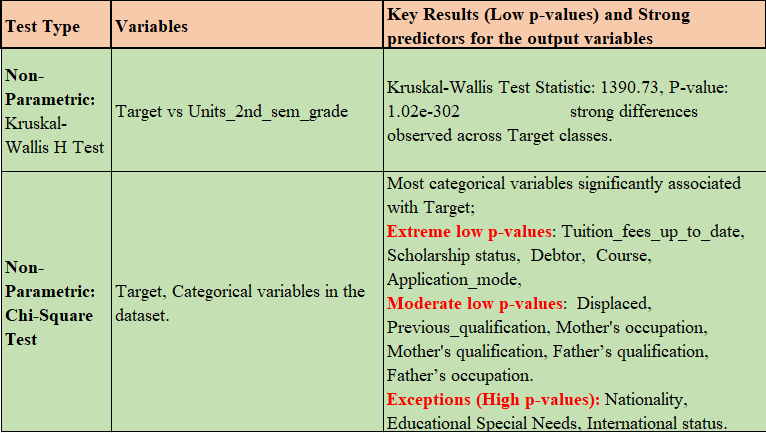
Clustering results suggested three loosely aligned student profiles: high performers, underperformers, and a mixed group. Due to the Enrolled class’s ambiguity, reframing the problem as binary classification improved both performance and interpretability

**5. Statistical Tests Summary:**

Parametric and Non-parametric statistical tests were conducted to evaluate the significance of each feature with respect to the target variable.

**5.1 Parametric Tests** 

***Table 7: Statistics – Parametric Tests***

**5.2 Non- Parametric Tests**

***Table 8: Statistics : Non-Parametric Tests***

The Curricular Units and Financial factors show a strong association with the Target variables. Some features, such as Course, Displaced status, and Mother's and Father's qualification and occupation, also display associations but are notably weaker. Variables like International status, Nationality, and Special Educational Needs show no significant impact on the Target.

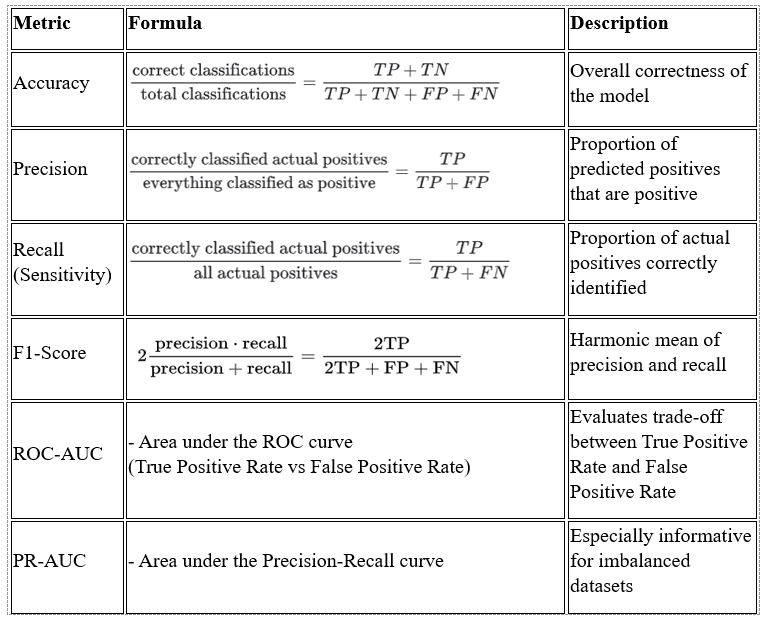
**6. Model Selection and Training**

**6.1 Model Categories and Algorithms:**

**Traditional Models:** Logistic Regression , Support Vector Machine (SVM), Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP Neural Network).

**Boosting Models:** Gradient Boosting, Extreme Gradient Boosting (XGBoost), CatBoost, LightGBM.

**6.2 Evaluation Metrics**

****

****

**7. Multiclass vs. Binary Classification strategy:**

Initially, machine learning models were applied to the multiclass Target (Graduate, Dropout, Enrolled). While the models performed very well for the Graduate and Dropout classes, the Enrolled class consistently achieved scores that were approximately half of the others, primarily due to class imbalance. Application of SMOTE improved the results, bringing all three class scores closer to the 0.80 range.

Statistical analysis such as correlation analysis showed:

* Graduate: Strongly associated with academic and financial factors.
* Dropout: Patterns opposite to the Graduate class.
* Enrolled: Correlated mainly with evaluation aspects of curricular units, but without strong, consistent patterns.

Further dimensionality reduction techniques like PCA and LDA struggled to clearly differentiate the Enrolled class. Clustering analysis was also performed to explore natural groupings within the data. Results confirmed that:

* Academic performance is the primary driver of separation between groups.
* Financial aspects also contribute notably.
* Demographic and macroeconomic factors have minimal predictive value.

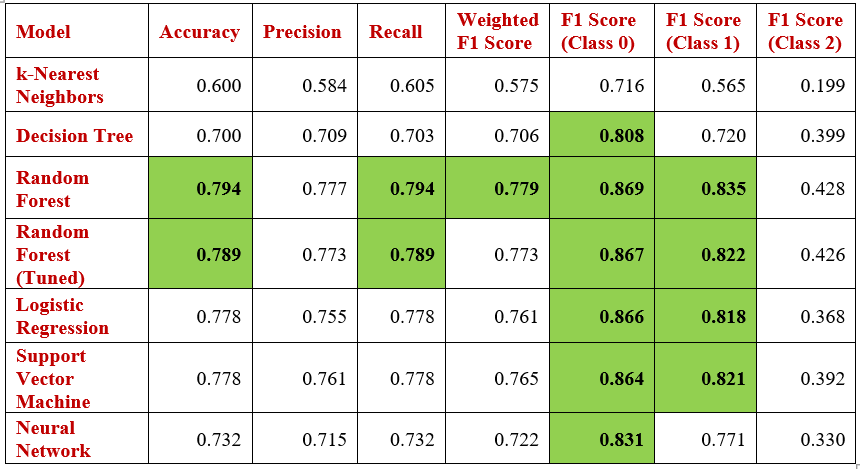
The Enrolled class (Target 2) lacks distinct feature patterns and overlaps significantly with both Graduates and Dropouts. This overlap introduces noise, which reduces precision, recall, and F1-scores for the Enrolled group, despite strong results for Graduates and Dropouts.

Given these findings, binary classification experiments were conducted by removing the Enrolled class and reframing the task to focus on Graduates vs. Dropouts. This simplification allowed models to focus on meaningful differences and improved overall prediction accuracy and stability.

**8. Experiments and Model Performance Evaluation**

This section presents the sequence of machine learning experiments.

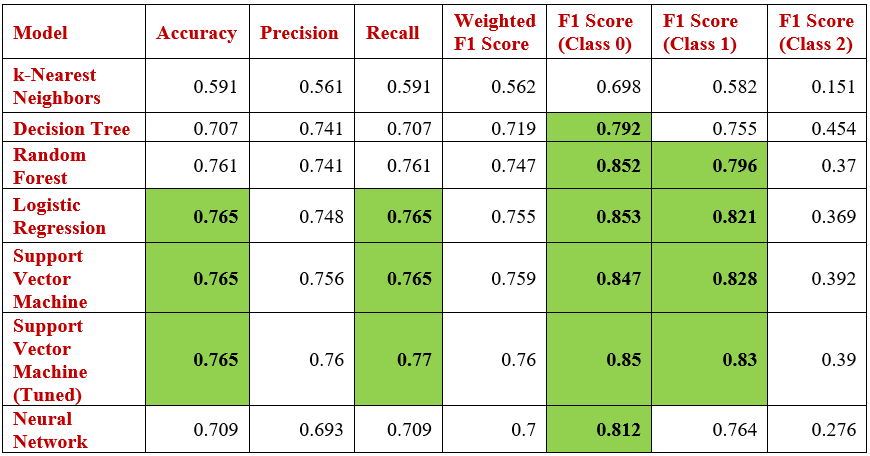
**8.1 Experiment 1: Baseline Modeling (75:25 Split)**



***Table 10: Performance of the models – Experiment 1***

Random Forest delivered the best performance (Accuracy: 0.795), with strong F1 scores for majority classes. SVM and Logistic Regression were close competitors, performing well on well-separated classes. Hyperparameter tuning improved Random Forest marginally, especially in model depth and splits. Class 2 (Enrolled) remained difficult to classify, with minimal F1 improvement—suggesting the need for rebalancing or more advanced techniques.

**8.2 Experiment 2: Baseline Modeling (90:10 Split)**



***Table 11: Performance of the models – Experiment 2***

* Default Accuracy: 76.52%,
* Tuned Accuracy: 76.52% (Train: 75.76%).
* Best Params: C=1, gamma='scale', kernel='linear'
* No performance gain from using more training data (90%) vs. 75% (which gave higher accuracy with RFC at 80%). Hyperparameter tuning confirmed linear kernel as optimal and demonstrated good generalization (no overfitting), despite no test accuracy improvement.

**8.3 Experiment 3: Dimensionality Reduction (PCA & LDA):**

**Model Performance on LDA-Transformed Data (Default Parameters)**

***Table 12: Experiment 3: Results: LDA with Default parameters***

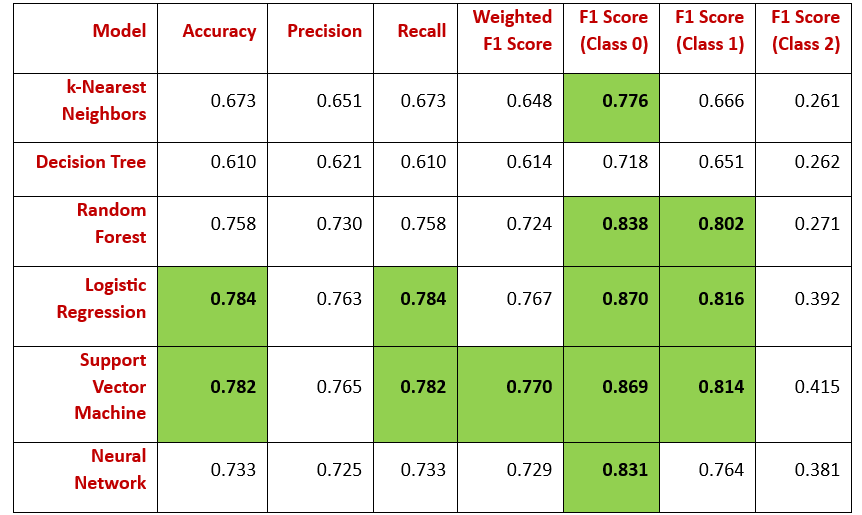
Neural Network achieved the highest accuracy (79.7%) and F1 scores across all classes, indicating it adapted best to the LDA-reduced feature space. Logistic Regression and SVM also performed strongly ( accuracy 79%), (linear classifiers benefit from LDA's linear separability).

**Model performance of LDA transformed data (Hyperparameters)**

***Table 13: Experiment 3: Results: LDA with Hyperparameters***

SVM achieved the highest overall performance, with an accuracy of ~80% and high F1 Score across all classes, making it the top-performing model on LDA-transformed data.

All the linear models — including Logistic Regression, Support Vector Machine, and Neural Network — showed improved performance on the LDA-transformed data. This aligns with the fact that Linear Discriminant Analysis projects the data into a space that maximizes class separability using linear combinations, which naturally benefits linear classifiers.

**Model Performance on PCA-Transformed Data (Default Parameters)**

***Table 14: Experiment 3: Results: PCA with Default parameters***

Logistic Regression has shown the best performance on PCA-transformed data, achieving an accuracy of approximately 78%, with a high recall rate.

Support Vector Machine (SVM) closely follows with 78% accuracy and comparable recall.

**Model Performance on the PCA transformed data (Hyperparameters)**

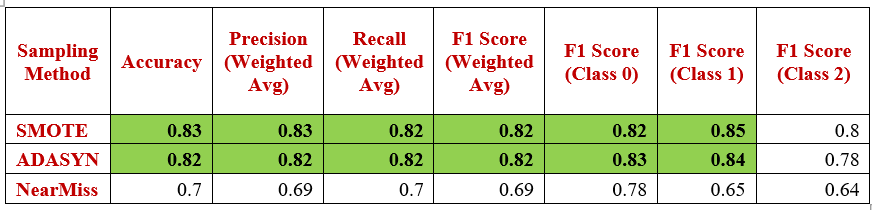
***Table 15: Experiment 3: Results: PCA with Hyperparameters***

SVM and Logistic Regression models performed well with higher accuracy of 79% and F1 Score of the two classes.

**8.4 Experiment 4: Imbalanced Data handling (SMOTE, ADASYN, NearMiss):**

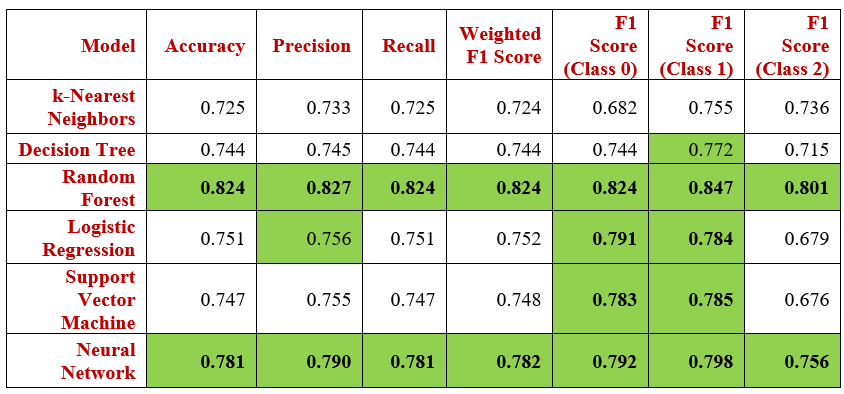
Random Forest Classifier was evaluated with three imbalance handling techniques: SMOTE, ADASYN (both oversampling), and NearMiss (undersampling).

**RandomForest Classifer (Default Paramters)**

****

***Table 16: Experiment 4 Results: RFC + Balancing techniques***

SMOTE achieved the highest accuracy at 83%, slightly outperforming ADASYN (82%), while NearMiss lagged at 70%. All techniques improved F1 scores across all classes, including the difficult "Enrolled" class. SMOTE was preferred for further analysis due to its balance and stability, whereas ADASYN introduced minor noise and NearMiss underperformed due to loss of majority class data.

**ML models + SMOTE (Default parameters)**

***Table 17: Experiment 4 Results: ML (Default) + SMOTE***

Random Forest led with 82.4% accuracy and strong F1-scores across all classes, notably 0.801 for class 2. Neural Network followed with 78.1% accuracy and balanced F1 performance. SVM and Logistic Regression had similar results, with better precision but weaker F1 for class 2. KNN underperformed, showing limited benefit from SMOTE due to sensitivity to feature overlap and scaling.

**ML models + SMOTE (Hyperparameters)**

***Table 18: Experiment 4 Results: ML (Hyperparamter) + SMOTE***

Support Vector Machine (SVM) was the top performer after tuning, achieving 84.2% accuracy and a standout F1 score of 0.892 for class 2. Random Forest continued its strong performance with a balanced profile across all classes.

**8.5 Experiment 5: Feature Selection Using Random Forest Importances**

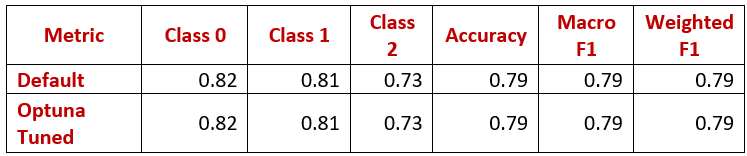
The top 15 features from the ‘Random Forest features importances’ was experimented with SMOTE at 80:20 split. Random Forest Classifier achieved 83% accuracy, slightly improving over previous results. Performance remained balanced across classes, especially for classes 0 and 2. This confirms that a reduced, focused feature set can maintain accuracy while improving model efficiency and interpretability.



***Table 19: Experiment 5 Results: RF Feature Importances***

**8.6 Experiment 6: Boosting models + SMOTE**

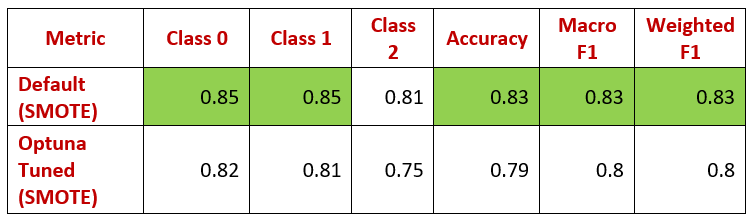
**Gradient Descent Classifier (SGDClassifier):**



***Table 20: Experiment 6 Results: SGD + SMOTE***

Performed consistently across default and Optuna-tuned versions with no improvement, indicating model stability. It handled classes 0 and 1 reasonably well but struggled with class 2.

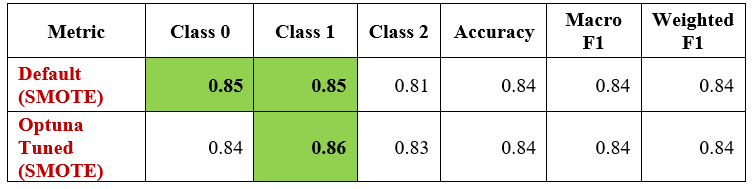
**Extreme Gradient Boosting (XGBoost)**

****

***Table 21: Experiment 6 Results: XGBoost + SMOTE***

The default settings outperformed Optuna-tuned version (83% vs. 79% accuracy). It struggled with class 2 (F1 ≈ 0.75), indicating limitations in class imbalance sensitivity.

**CatBoost Classifier**

****

***Table 22: Experiment 6 Results: CATBOOST + SMOTE***

Both configurations perform equally well overall with 84% accuracy. The Optuna-tuned version slightly improves performance for Class 2 (Dropouts), which saw an F1 increase from 0.81 to 0.83. CatBoost with default settings is already robust, but Optuna tuning fine-tunes class balance, especially for underperforming classes.

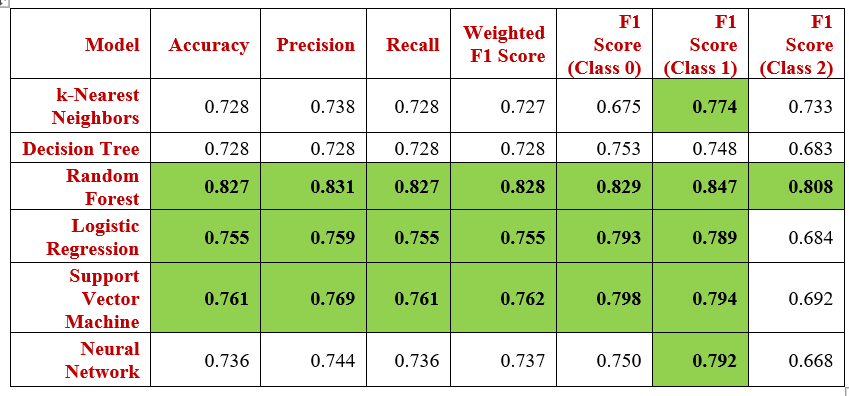
**Light Gradient Boost Classifier: Best performance**

***Table 23: Experiment 6 Results: LightGBM + SMOTE***

Optuna tuning improves the overall accuracy from 84% to 86%. F1 score boost for Class 2 (Dropouts): from 0.81 to 0.85, showing better sensitivity to the minority class, with minor improvements in other classes. This demonstrates the model’s scalability with careful hyperparameter optimization.

**8.7 Experiment 7: Multicollinearity + Feature engineering**

Multicollinearity using VIF, was identified among semester-based academic features (VIF > 10). To address this, these were consolidated into combined features: Total\_credited, Total\_enrolled, and Total\_approved. This reduced redundancy while preserving key predictive signals.

**ML models with feature engineering (Default parameters):**

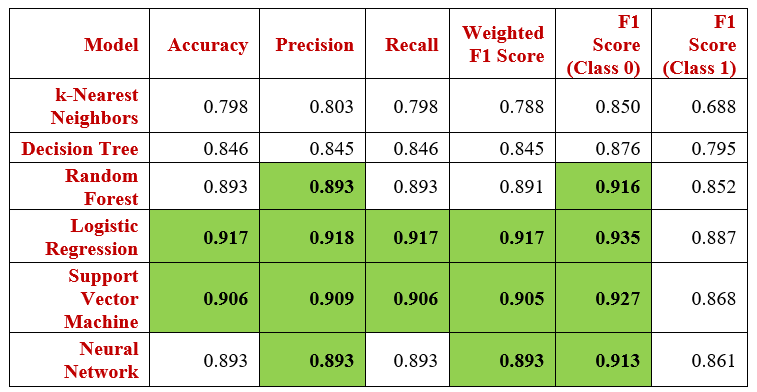
***Table 24: Experiment 7 Results: ML (Default) with Feature engineering***

Random Forest performed well with high accuracy and balanced performances of all classes to all the other models. Random Forest achieved similar results with hyperparameters as well.

Combining semester-based features preserved predictive power and Random Forest remained the strongest performer post-engineering.

**8.8 Experiment 8: Binary Classification**

Models trained on two student outcomes: Graduate (Class 0) and Dropout (Class 1) with dropping the ambiguous Class 2 (Enrolled) to create a binary classification problem

**Model Performance (Default Parameters)**

***Table 25: Experiment 8 Results: ML (Default) for Binary Classification***

* Logistic Regression: 92% accuracy — best overall; simple and robust. SVM: 91% accuracy, strong generalization. Random Forest & Neural Network: 89% accuracy; captured complex patterns well

**ML models (Hyperparameters):**

***Table 26: Experiment 8 Results: ML (Hyperparameters) for Binary Classification***

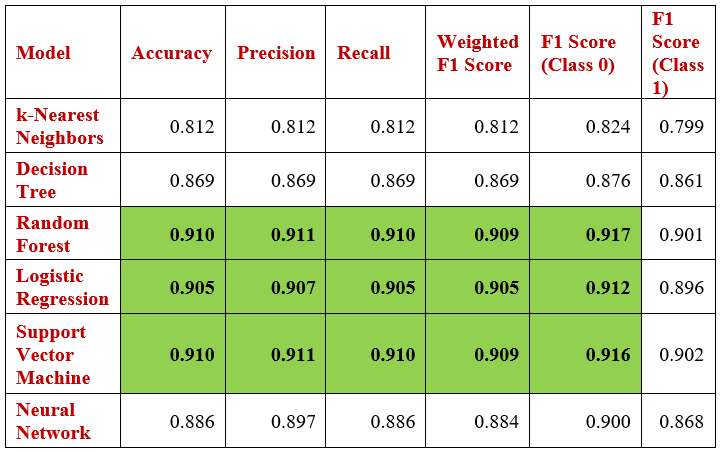
With hyperparameter tuning, SVM delivered the best results with 95% accuracy and balanced precision, recall, and F1 score (~95%) across both classes. Logistic Regression closely followed with 94% accuracy, maintaining strong, stable performance.

**Performance Comparison: Before vs After Removing 'Enrolled' Class**

Removing the ambiguous Enrolled class — characterized by overlapping features and weak correlations, eliminated noise that had reduced model performance. Originally, models averaged ~80% accuracy with poor F1 (~40%) for Enrolled. After shifting to binary classification (Graduate vs Dropout), class separability improved significantly, especially through academic metrics. This led to a substantial accuracy boost (up to 95%) and F1 scores nearing 95%, particularly with tuned SVM.

**8.9 Experiment 9: Binary Classification + SMOTE for Traditional + Boosting models:**

To address class imbalance between Graduate (2209) and Dropout (1421), SMOTE was applied with traditional and Boosting models.

**ML Models (SMOTE + Default Parameters)**

***Table 27: Experiment 9 Results: ML (Default) + SMOTE***

Random Forest, Logistic Regression, and SVM (with default settings) each achieved ~91% accuracy and balanced F1 scores for both classes, demonstrating SMOTE’s effectiveness in improving class parity.

However, these SMOTE-based models did not surpass the performance of the hyperparameter-tuned SVM (95% accuracy) trained on Non-SMOTE data. This highlights that while SMOTE improves baseline consistency, hyperparameter tuning and a clean binary setup have a greater impact on performance.

**ML Models (SMOTE + Hyperparameters):**

***Table 28: Experiment 9 Results: ML (Hyperparameters) + SMOTE***

Random Forest, SVM, and Logistic Regression achieved ~91% accuracy with well-balanced F1 scores (90–92%) after applying SMOTE, confirming effective class balance. However, these models still underperformed compared to their non-SMOTE counterparts with hyperparameter tuning—SVM (95%) and Logistic Regression (94%).

**Boosting Algorithms (with SMOTE): Default + Tuned (Optuna)**

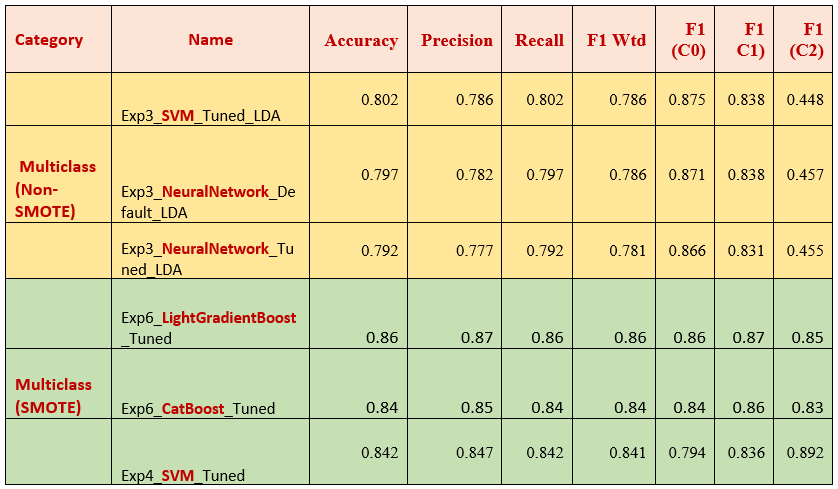
****

***Table 29: Experiment 9 Results: Boosting + SMOTE***

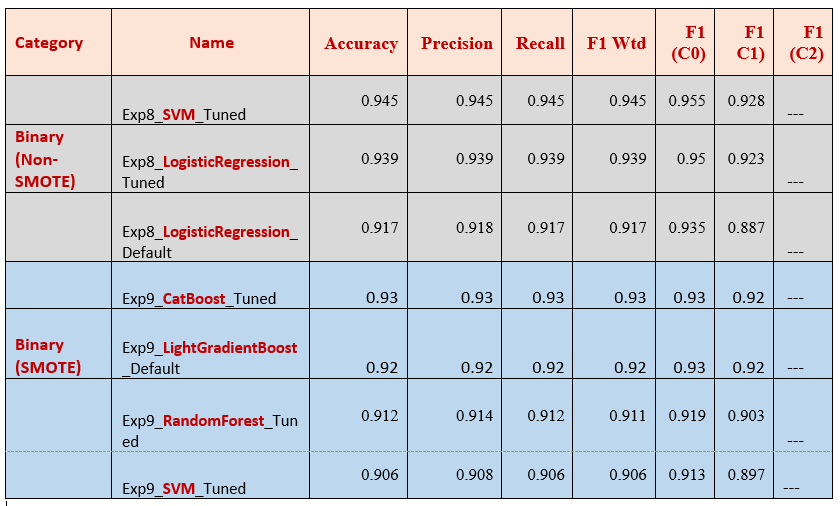
CatBoost (Optuna tuned) achieved the highest performance overall, reaching 93% accuracy and F1 scores across both classes. XGBoost and LightGBM (default) also performed exceptionally well, reaching 92% accuracy, showing that their default settings are already highly effective. Hyperparameter tuning via Optuna brought marginal to no gains for most models except CatBoost, suggesting limited room for improvement or that defaults were close to optimal.

**9. Cross-Experiment Summary: Multiclass & Binary Classification**

The performed ML models across various experiments are divided into two main categories – Multiclass and Binary classification, further subdivided into SMOTE and Non-SMOTE conditions to gain deeper insights. Below are the top-performing models across each category based on accuracy, precision, recall, class-level F1 scores and ROC/PR curves.

**9.1 Multiclass classification top results**

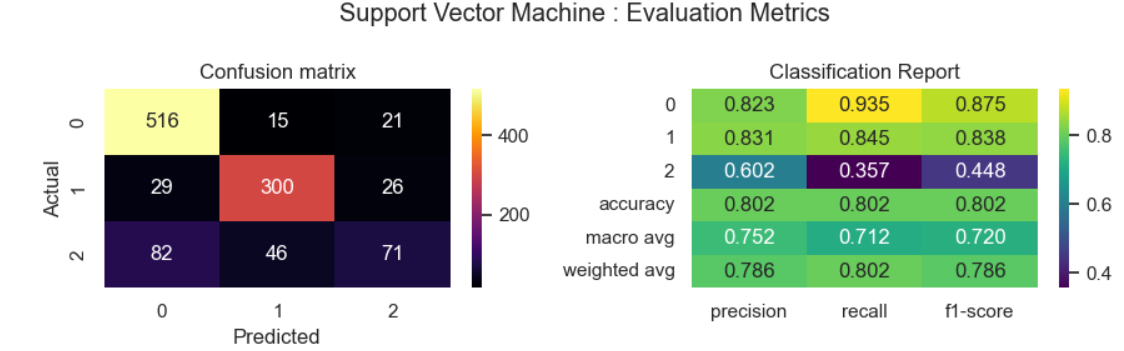
***Table 30: Top Results: Multiclass classification***

**9.2 Binary Classification top results**

***Table 31: Top Results: Binary classification***

**9.3 Multiclass (Non-SMOTE)**

**Exp3\_SVM\_Tuned\_LDA**: Highest overall accuracy (80.2%) in non-SMOTE multiclass models. Balanced performance across all classes, especially good F1 for Class 0 and 1. High True Positives for Class 0 and Class 1, while all models struggled with Class 2 in non-SMOTE strategy due to high class imbalance. The F1 score of Class 2 is 44.8% in comparison with the other classes, but improved score compared to other models.



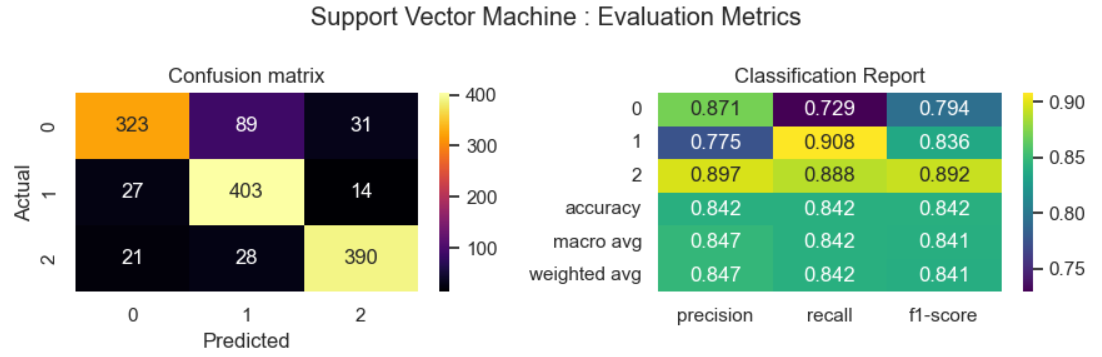
***Figure 20: SVM : Evaluation Metrics***

**Exp3\_NeuralNetwork\_Default\_LDA**: Best F1 for all classes especially Class 2. Slightly lower accuracy but better F1 for Class 2 (0.457), making it more inclusive of minority class.

**Exp3\_NeuralNetwork\_Tuned\_LDA**: Tuned version still strong on Class 2 and well balanced, solid overall.

**9.4 Multiclass (SMOTE)**

**Exp4\_SVM\_Tuned**: Top F1 score for minority Class 2 (0.892), showcasing excellent balance after SMOTE and hyperparameter tuning. The True Positives are high in all the three classes with less miscalculations. The F1 Score of Dropout is improved to 83.6%, which is essential in prediction of the dropouts.



***Figure 21: SVM - Evaluation Metrics***

**9.5 ROC Curve Evaluation:**

* Class 0: AUC = 0.93
* Class 1: AUC = 0.96
* Class 2: AUC = 0.96

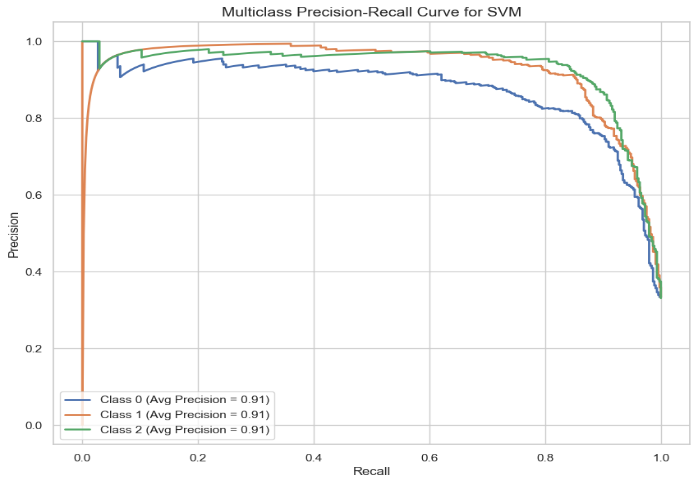
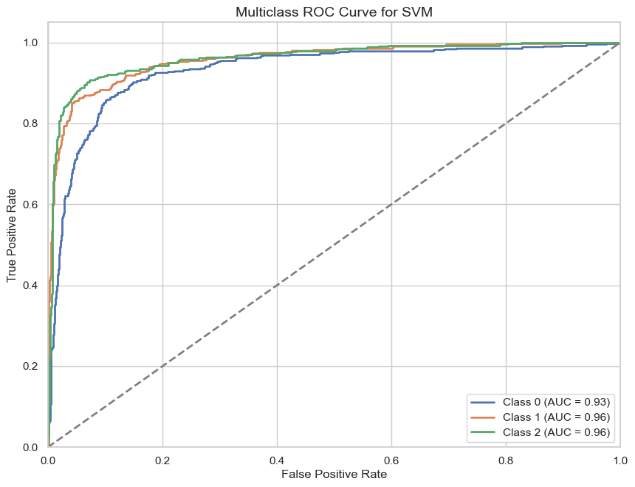
The SVM model, demonstrates very strong discriminative performance across all classes. The AUC values above 0.90 indicate the model can reliably distinguish each class from the others. Slightly lower AUC for Class 0 (0.93) might indicate marginally more overlap with other classes, but still within high-performance range.

**Precision-Recall curve – Average Precision (AP):**

* Class 0: 0.91
* Class 1: 0.91
* Class 2: 0.91

The PR curves further reinforce the model’s strong class-wise precision and recall balance, particularly important for imbalanced or multiclass datasets.

**Multiclass ROC Curve & PR Curve – SVM**



***Figure 22: SVM - ROC Curve & PR Curve***

* **Exp6\_CatBoost\_Tuned**: Hyperparameters. Consistently high scores across all classes, with a notable F1 score of 0.83 for Class 2.
* **Exp6\_LightGradientBoost\_Tuned**: Hyperparameters. Best overall accuracy (0.86) with robust class-wise performance, especially Class 2 (0.85).

**ROC Curve (Receiver Operating Characteristic):**

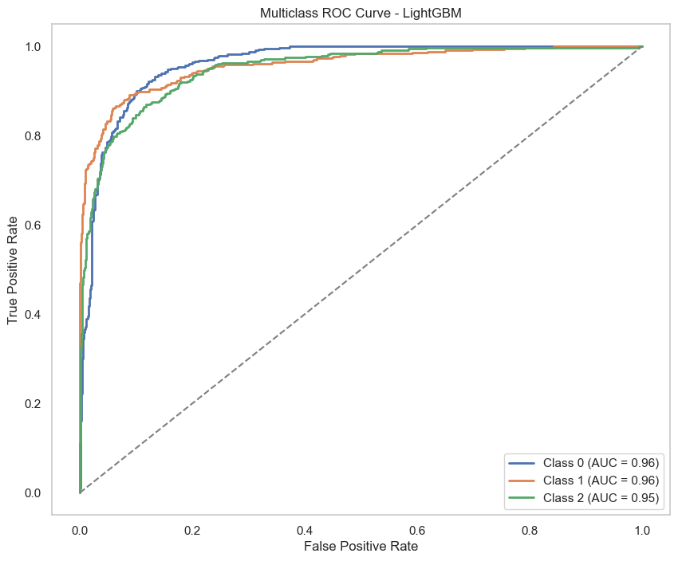
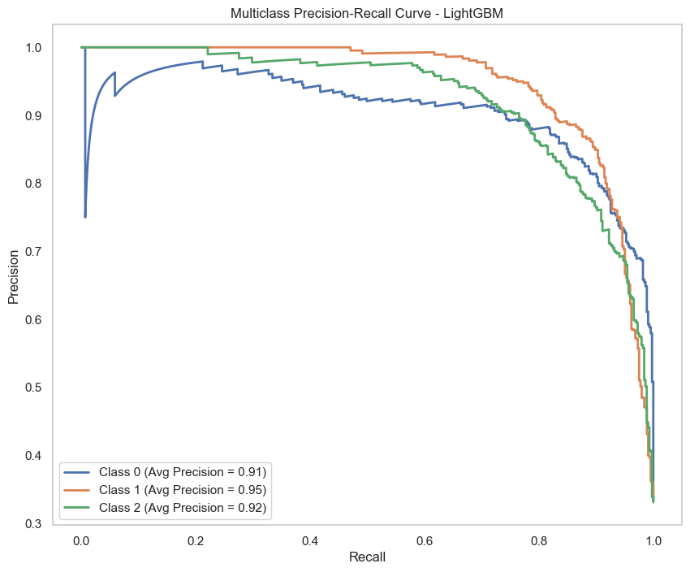
* Class 0: AUC = 0.96
* Class 1: AUC = 0.96
* Class 2: AUC = 0.95

These high AUC values indicate that the LightGBM model, when combined with SMOTE and fine-tuned via Optuna, performs exceptionally well in distinguishing between all three classes. Values above 0.90 shows indicate strong classifier performance.

**Precision-Recall Curve:**

* Class 0: Avg Precision = 0.91
* Class 1: Avg Precision = 0.95
* Class 2: Avg Precision = 0.92

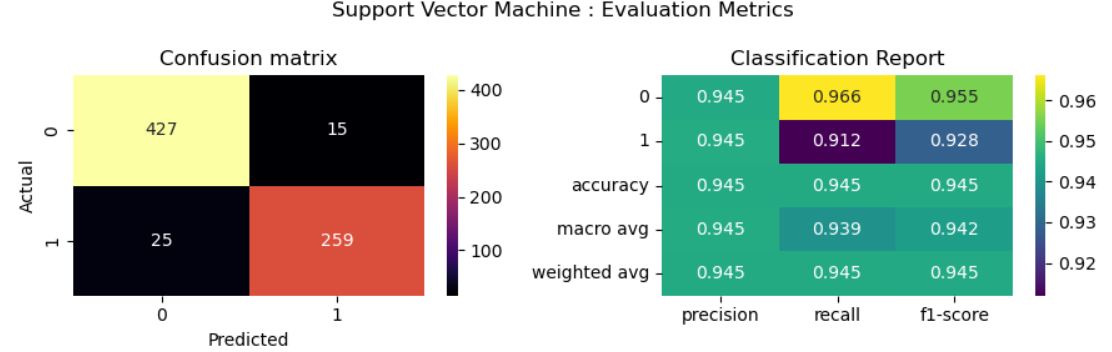
The average precision values (area under the PR curve) reflect the model’s ability to maintain high precision across various recall thresholds. This is particularly useful in imbalanced datasets (which is often why SMOTE is used), and these scores show that the model balances precision and recall well across all classes.

**Multiclass ROC Curve & PR Curve - LightGBM** 

***Figure 23: LightGBM - ROC Curve & PR Curve***

**9.6 Binary (Non-SMOTE)**

**Exp8\_SVM\_Tuned**: Best performing binary model overall. Highest accuracy 95% and excellent precision-recall balance. In this approach, the True positives and True negatives are very higher with lesser miscalculations. The F1 Score of class 0 and class 1 is high at 96% and 93%.

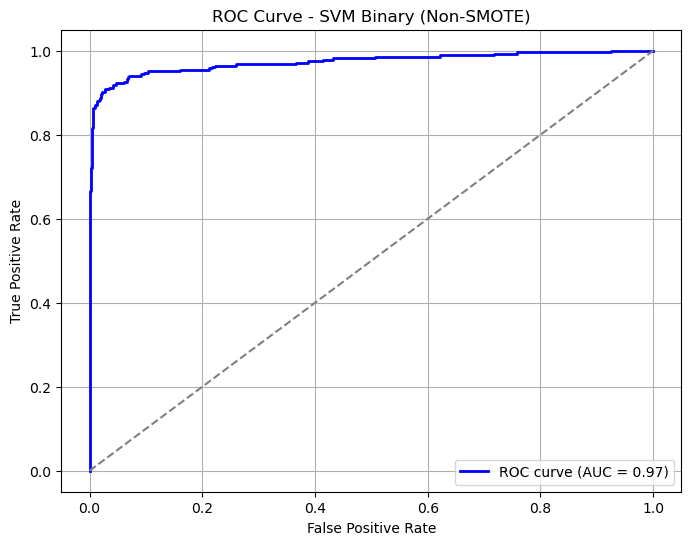
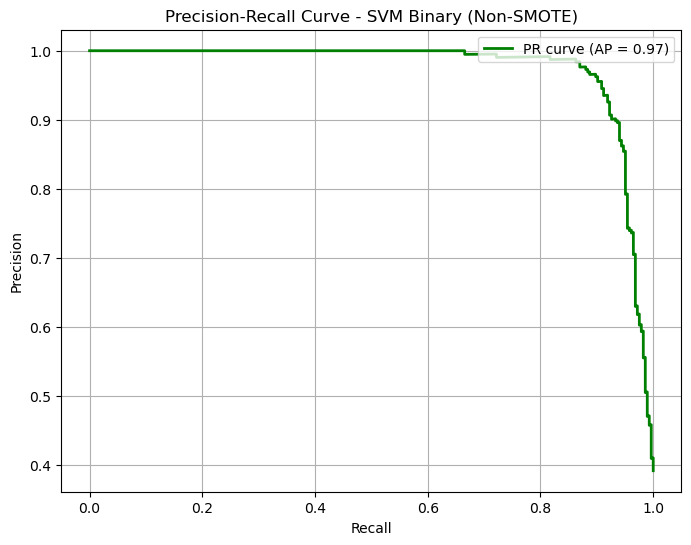


***Figure 24: SVM : Evaluation metrics***

**ROC Curve:** ROC AUC = 0.97 shows that there is an excellent discriminative power. The closer to 1.0, the better the classifier is at separating the two classes – Graduate and Dropout.

**Precision-Recall Curve:** AP (Average Precision) = 0.97 This also excellent. It means that across different thresholds, precision and recall remain very high.

The ROC and Precision-Recall curves show strong separability and precision consistency. The AUC and average precision scores both exceed 0.95, confirming the robustness of the classifier in distinguishing between the two classes without SMOTE augmentation.

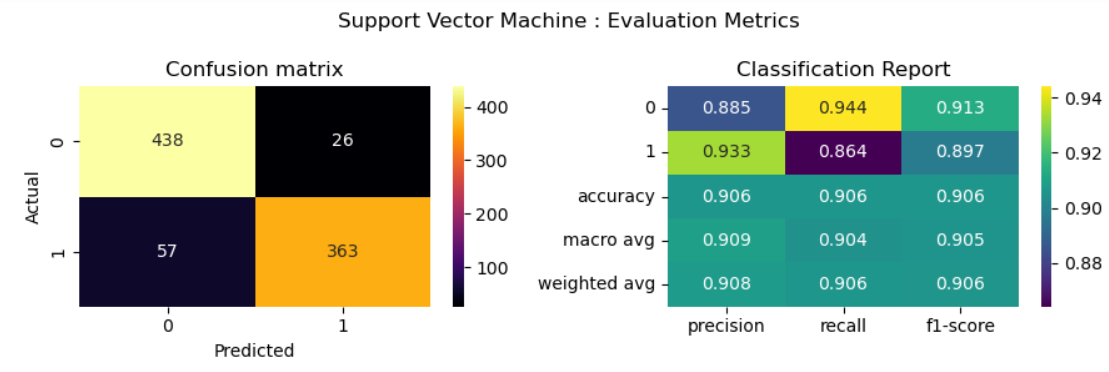
**Binary ROC Curve & PR Curve – SVM (Non-SMOTE)** 

***Figure 25: SVM – ROC & PR Curve***

* **Exp8\_LogisticRegression\_Tuned**: Close performance to SVM, with better F1 for Class 0 (0.95).
* **Exp8\_LogisticRegression\_Default**: Strong baseline model, performs surprisingly well even without tuning.

**9.7 Binary (SMOTE)**

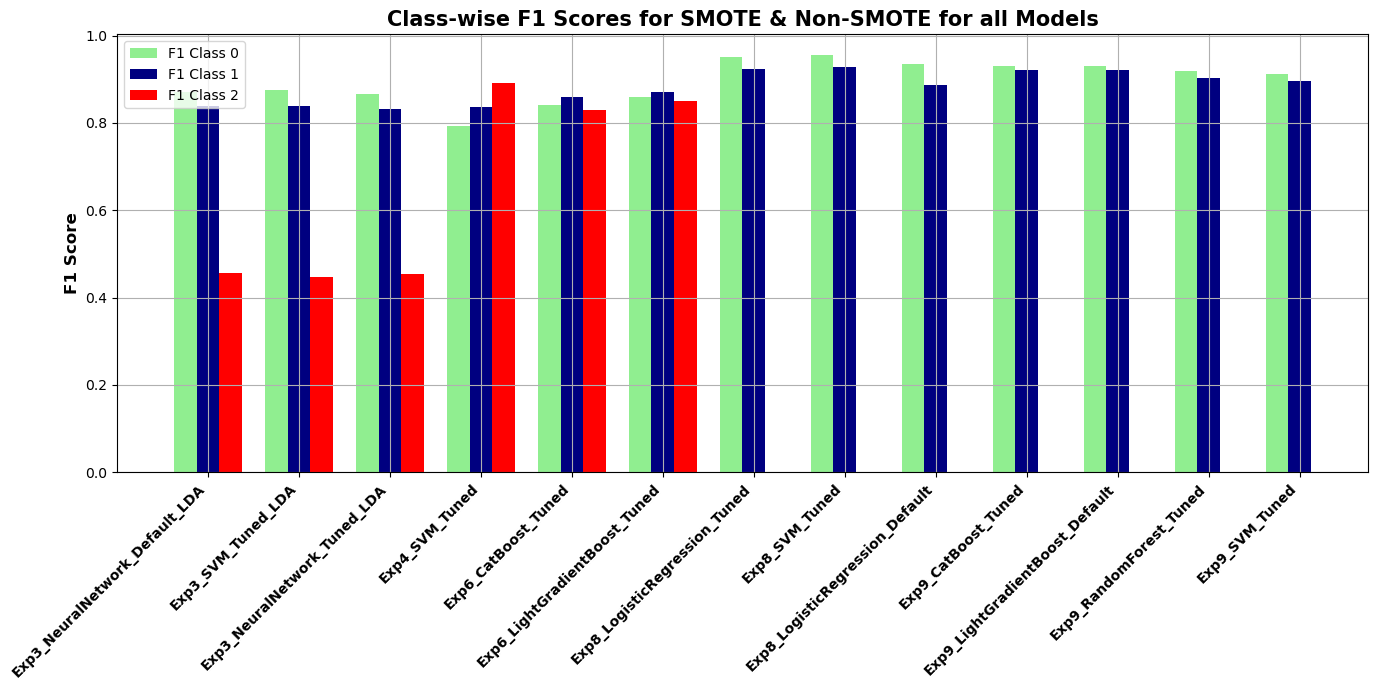
* **Exp9\_CatBoost\_Tuned**: Exceptional performance, especially with F1 of 0.93 for both classes — benefiting strongly from SMOTE.
* **Exp9\_LightGradientBoost\_Default**: Solid performance, especially F1 score of 0.93 for Class 0 — good generalizability.
* **Exp9\_RandomForest\_Tuned**: Consistent performance across metrics, and strong in Class 0 (0.919).
* **Exp9\_SVM\_Tuned**: It performed better and was close to the Boosting models. The True Positives and the True Negatives are high with few misclassifications but the False Negatives are high. The F1 score of 91% and 90% shows balance across classes.



***Figure 26: SVM – Evaluation metrics***

**10. Model Evaluation:**

The following are the top 13 results across Multiclass and Binary classification, using both SMOTE and Non-SMOTE methods.

**Class-wise F1 Scores – Top models (Multiclass + Binary)**

***Figure 27: Top Models (Multiclass + Binary) Class-wise F1 scores***

**10.1 Best models across all experiments:**

**1. SVM Tuned:**

The Support Vector Machine with hyperparameter tuning consistently achieved the highest scores across all strategies — both in Multiclass and Binary classifications, with and without SMOTE.

Under Non-SMOTE, it was the top performer across all metrics. With SMOTE, it closely matched the performance of top Boosting models (LightGBM and CatBoost). ROC and Precision-Recall curves consistently demonstrated AUC and Average Precision scores above 0.90 for different experiments, reinforcing its reliability. Due to its robustness, balance, and computational efficiency, SVM (Tuned) emerged as a consistently top-performing model.

**2. Boosting Models:**

Boosting models, applied within the SMOTE framework, showed significant improvement in class-wise F1 balance and overall performance.

CatBoost and LightGBM, in particular, performed exceptionally well across all classes. Their ROC and PR curves revealed AUC and Average Precision values exceeding 0.90, indicating strong discriminative ability and precision, especially under class imbalance.  
Although slightly more computationally intensive, they present strong alternatives to SVM.

**3. Traditional Models:**

While SVM Tuned led overall, other traditional classifiers — including Random Forest, Neural Networks and Logistic Regression (both default and tuned) — also demonstrated competitive performance, especially when paired with SMOTE and feature importances.  
Their consistency across both Multiclass and Binary settings makes them viable candidates, particularly when simplicity or faster training is a priority.

**10.2 Recommendation:**

**Multiclass Classification:**

SVM Tuned consistently provides balanced performance and class stability, with a slight edge in accuracy and F1 scores. It remains the more computationally efficient option, especially for practical applications with limited resources

Boosting Models (CatBoost, LightGBM) with SMOTE slightly outperform SVM Tuned in accuracy and class-wise F1 scores, making them great choices when class imbalance is critical.

**Binary Classification:**

SVM Tuned again emerges as the top performer, providing balanced performance across both classes.

Boosting Models with SMOTE show strong results, especially in balancing class F1 scores, but with higher computational costs.

**10.3 Practical Recommendations**

For real-time or resource-constrained systems, SVM (Tuned) is the preferred model due to its efficiency and reliability. For maximum accuracy and class balance, Boosting models with SMOTE are excellent—especially in multiclass settings. Institutions should prioritize academic performance tracking and financial support monitoring to proactively identify at-risk students.

**10.4 Challenges:**

**1. Data Limitations**

The study utilized a single dataset from one institution, limiting the generalizability of findings across diverse educational environments. Key student attributes—such as year of admission, attendance records, extracurricular involvement, and mental health indicators—were not available, potentially constraining the model's predictive power.

**2. Model and Computational Constraints**

Resource-intensive models, particularly boosting algorithms and neural networks with hyperparameter optimization, demanded substantial computational power and memory, necessitating advanced hardware. Tuning processes using GridSearchCV and Optuna significantly increased training duration—requiring hours posing limitations on rapid experimentation and scalability.

**10.5 Future Work**

This research applied a broad range of machine learning models and preprocessing techniques to predict student outcomes, including dimensionality reduction, class balancing, feature engineering, and hyperparameter optimization. The findings and methodologies serve as a strong foundation for future researchers seeking to improve or extend this work.

**10.6 Conclusion**

Across all experiments, Support Vector Machine (SVM) with hyperparameter tuning emerged as the most balanced and computationally efficient model for both multiclass and binary classification. Boosting models like CatBoost and LightGBM also delivered strong results, particularly in imbalanced scenarios when combined with SMOTE, achieving up to 95% accuracy and high F1 scores. ROC and Precision-Recall curves consistently showed AUC and Average Precision scores above 0.90, confirming their robustness under class imbalance.

SMOTE notably enhanced minority class performance, especially for the ‘Enrolled’ category in multiclass settings, raising F1 scores from the 0.40s to 0.89 with SVM. In binary classification, SMOTE contributed to class balance but offered limited advantage over tuned models on clean data. ADASYN and NearMiss were evaluated but not used in final models due to SMOTE's superior early performance.

SVM remained favourable over boosting models in terms of computational efficiency, offering similar or better accuracy with significantly lower resource demands—making it a practical option for real-world applications.

Clustering analysis revealed patterns primarily based on academic and financial variables, with limited contribution from demographic features. However, the lack of behavioural and psychological data restricted the depth of unsupervised insights, highlighting the need for richer datasets.

Feature importance tools (SHAP, LIME, Random Forest Importances) and statistical tests consistently identified academic performance, financial status and Course as the most influential factors. Demographic factors had minimal impact, especially features such as nationality and special needs showed negligible relevance.

Overall, consistent findings across statistical analysis, feature attribution tools, and clustering validated the robustness of the models and confirmed the key drivers behind student dropout and success.

* 1. **References**

1. AlAssafi, H.S., 2021. *CRISP — DM Methodology (Cross-Industry Standard Process for Data Mining)*. [online] Medium. Available at: <https://hamadalassafi.medium.com/crisp-dm-methodology-cross-industry-standard-process-for-data-mining-51b9f93758ae> [Accessed 30 Apr. 2025].
2. Alban, M. and Mauricio, D., 2019. *Predicting University Dropout through Data Mining: A Systematic Literature*. *Indian Journal of Science and Technology*, 12(4), pp.1–12. <https://doi.org/10.17485/ijst/2019/v12i4/139729>.
3. Andrade-Girón, D., Sandivar-Rosas, J., Marín-Rodriguez, W., Susanibar-Ramirez, E., Toro-Dextre, E., Ausejo-Sanchez, J., Villarreal-Torres, H. and Angeles-Morales, J., 2023. *Predicting Student Dropout based on Machine Learning and Deep Learning: A Systematic Review*. *EAI Endorsed Transactions on Scalable Information Systems*. <https://doi.org/10.4108/eetsis.3586>.
4. archive.ics.uci.edu, n.d. *UCI Machine Learning Repository*. [online] Available at: <https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success> [Accessed 4 May 2025].
5. Bouchrika, I., 2022. *College Dropout Rates: 2022 Statistics by Race, Gender & Income*. [online] Research.com. Available at: <https://research.com/universities-colleges/college-dropout-rates> [Accessed 4 May 2025].
6. Del Savio, A.A., Galantini, K. and Pachas, A., 2022. *Exploring the relationship between mental health-related problems and undergraduate student dropout: A case study within a civil engineering program*. *Heliyon*, 8(5), p.e09504. <https://doi.org/10.1016/j.heliyon.2022.e09504>.
7. IBM, 2021. *CRISP-DM Help Overview*. [online] Available at: <https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=dm-crisp-help-overview> [Accessed 4 May 2025].
8. Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G. and Loumos, V., 2009. *Dropout prediction in e-learning courses through the combination of machine learning techniques*. *Computers & Education*, 53(3), pp.950–965. <https://doi.org/10.1016/j.compedu.2009.05.010>.
9. Manrique, R., Nunes, B.P., Marino, O., Casanova, M.A. and Nurmikko-Fuller, T., 2019. *An Analysis of Student Representation, Representative Features and Classification Algorithms to Predict Degree Dropout*. In: *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. <https://doi.org/10.1145/3303772.3303800>.
10. Nagy, M. and Molontay, R., 2018. *Predicting Dropout in Higher Education Based on Secondary School Performance*. [online] IEEE Xplore. <https://doi.org/10.1109/INES.2018.8523888>.
11. Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E. and Nshimyumukiza, P.C., 2022. *Predicting student’s dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization*. *Computers and Education: Artificial Intelligence*, 3, p.100066. <https://doi.org/10.1016/j.caeai.2022.100066>.
12. Palomino, J.C.V. and Ortega, A.M., 2023. *Dropout Intentions in Higher Education: Systematic Literature Review*. *Journal on Efficiency and Responsibility in Education and Science*, 16(2), pp.149–158. <https://doi.org/10.7160/eriesj.2023.160206>.
13. Piscitello, J., Kim, Y.K., Orooji, M. and Robison, S., 2022. *Sociodemographic risk, school engagement, and community characteristics: A mediated approach to understanding high school dropout*. *Children and Youth Services Review*, 133, p.106347. <https://doi.org/10.1016/j.childyouth.2021.106347>.
14. Realinho, V., Machado, J., Baptista, L. and Martins, M.V., 2022. *Predicting Student Dropout and Academic Success*. *Data*, 7(11), p.146. <https://doi.org/10.3390/data7110146>.
15. Sletten, M.A., Tøge, A.G. and Malmberg-Heimonen, I., 2022. *Effects of an early warning system on student absence and completion in Norwegian upper secondary schools: a cluster-randomised study*. *Scandinavian Journal of Educational Research*, pp.1–15. <https://doi.org/10.1080/00313831.2022.2116481>.
16. Villar, A. and Andrade, C.R.V. de, 2024. *Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study*. *Discover Artificial Intelligence*, 4(1), p.NA. <https://doi.org/10.1007/s44163-023-00079-z>.
17. Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R. and Gonzalez-Beltran, A., 2016. *The FAIR Guiding Principles for Scientific Data Management and Stewardship*. *Scientific Data*, 3(1). Available at: <https://www.nature.com/articles/sdata201618> [Accessed 4 May 2025].
18. Ytimg.com, 2024. *[image]*. Available at: <https://i.ytimg.com/vi/QrU5VHLZ1iw/maxresdefault.jpg?sqp=-oaymwEmCIAKENAF8quKqQMa8AEB-AH-CYAC0AWKAgwIABABGGUgZShlMA8=&rs=AOn4CLAfKxzSbKoOTXVHNRjZLCJ9QKphCQ> [Accessed 22 Oct. 2024].

|  |  |
| --- | --- |
| **GITHUB URL:** | <https://github.com/CCT-Dublin/ca1-capstone-project-proposal-Analytics-Selvi> |