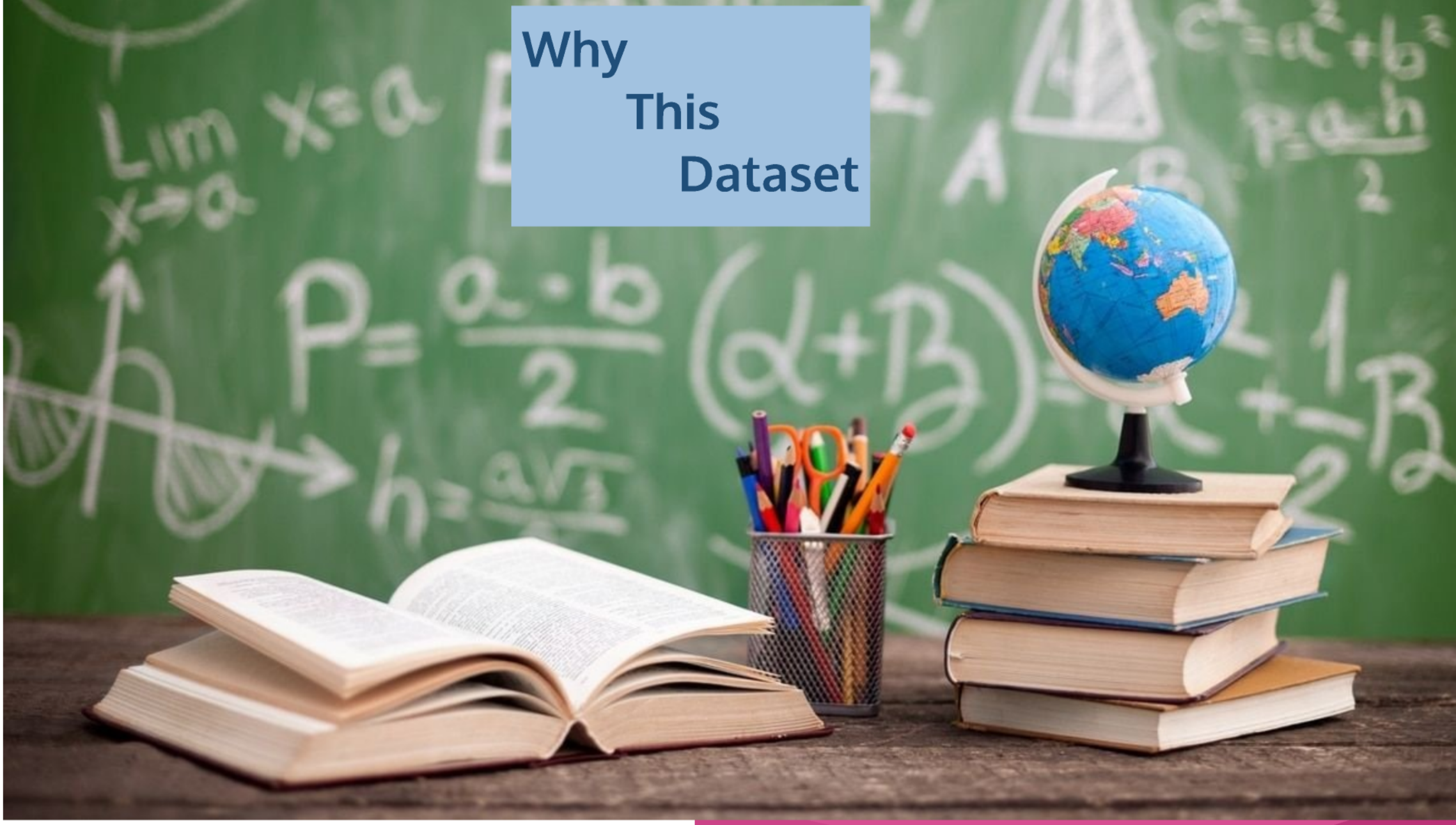
Academic Success, or College Dropout

Why This Dataset



What Are The Features

A diverse pool of information was selected from this student body. This ranged from their

- Nationality
- Parents' level of education
- Their occupation
- Inflation rate
- Tuition owed
- Class Times AM/PM
- Gender
- Educational Special Needs

Goals of the Project

Assess the reliability of:

- Logistic Regression model
- Random Forest Classifier

Predict whether students will:

- Graduate
- Drop out of college

Use insights to:

- Identify "at risk" students
- Enhance educational outcomes
- Develop proactive strategies

Null Hypothesis (H0)

Student data features such as (e.g., marital status, academic performance, economic indicators) do not meaningfully predict dropout, enrollment, or graduation outcomes.

Alternative Hypothesis (H1):

Student data features do help predict dropout, enrollment, or graduation, and models like Random Forest and Logistic Regression can classify these outcomes accurately.

Data Collection Methods

Data was collected from the Instituto Politécnico de Portalegre in Portugal. It comprises data from students enrolled in various undergraduate programs. The information was gathered from multiple disjoint databases within this higher education institution.

Data Cleanup

This chapter outlines the essential methods employed for data collection and cleanup, ensuring the dataset's integrity for accurate dropout predictions.

Data Cleanup Processes

Dataset cleanup included:

- Removal of a value in the 'Target' column to ensure true binary classification
- Verification of no duplicate or missing values
- Detection of grammatical inconsistencies

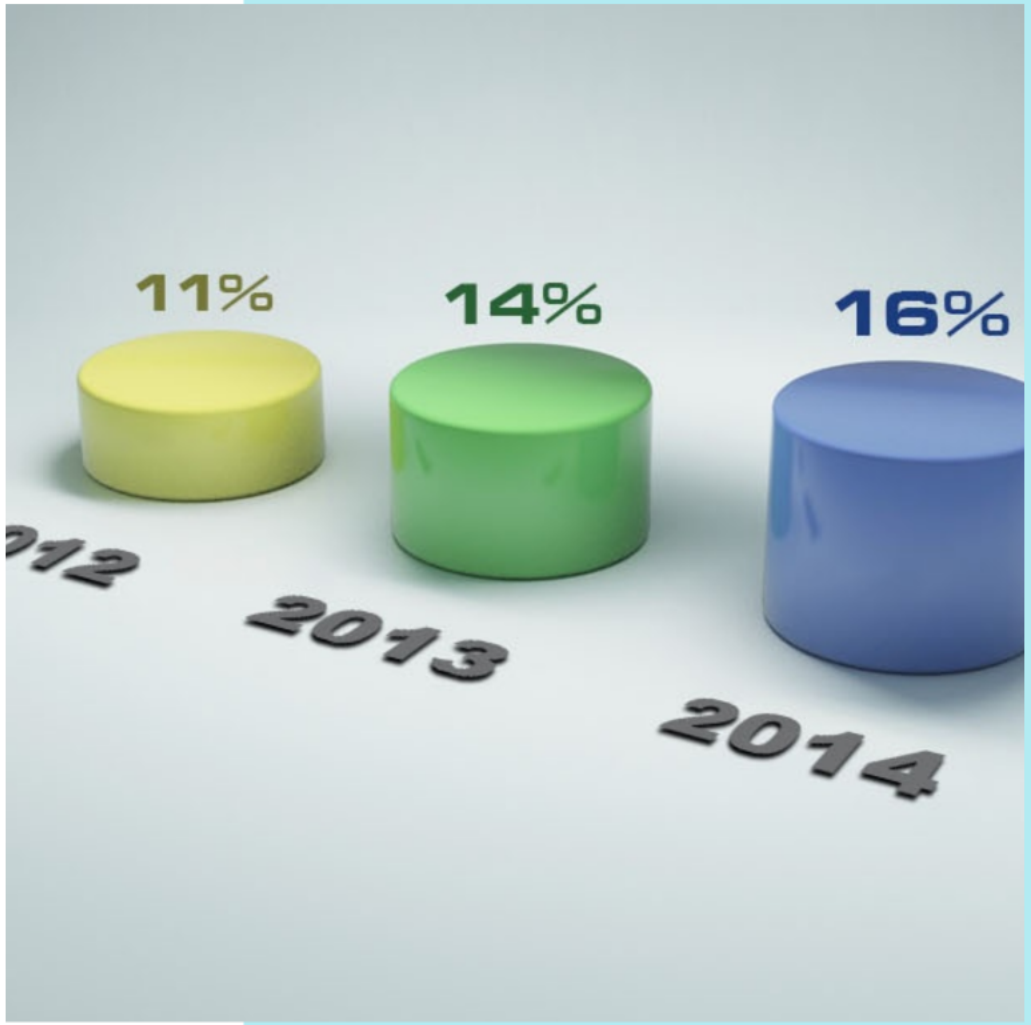
Cleanup was crucial for:

- Maintaining data quality
- Enhancing the model’s predictive accuracy

Model Training

Model Training Process

- Preprocessed and scaled the training data
- Trained models using Logistic Regression and Random Forest Classifiers
- Tuned hyperparameters for optimal performance



Methodology

- Systematically selected relevant features
- Processed data for training
- Emphasized rigorous evaluation using accuracy metrics
- Ensured reliable prediction of dropout rates based on the training dataset

Metrics for Accuracy Assessment

Model Evaluation Metrics

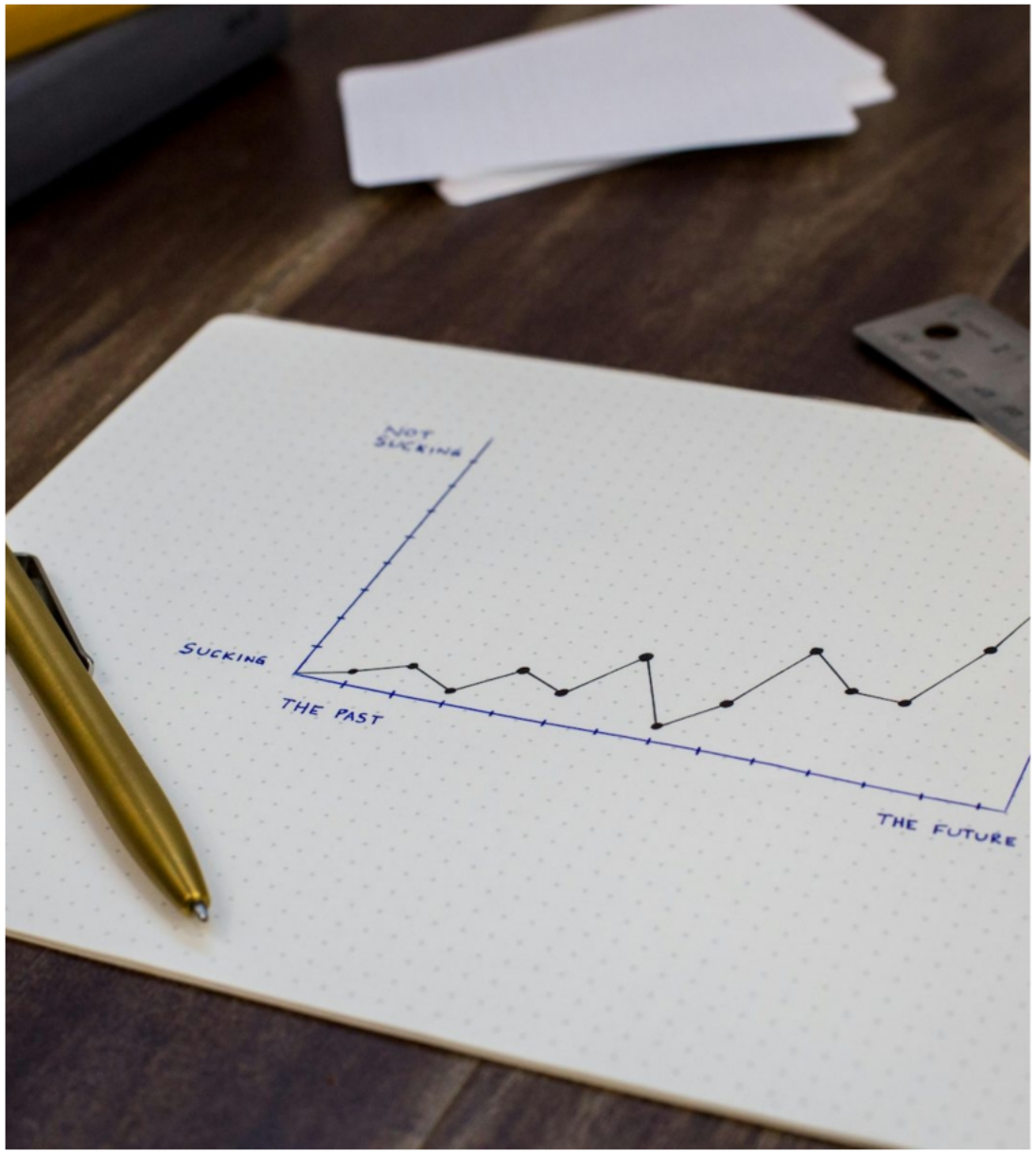
Measured accuracy using:

- Accuracy Score
- Classification Report
- Confusion Matrix
- Provided a comprehensive view of model performance
- Helped assess how well the model predicts dropout vs. graduation

Performance Before Tuning

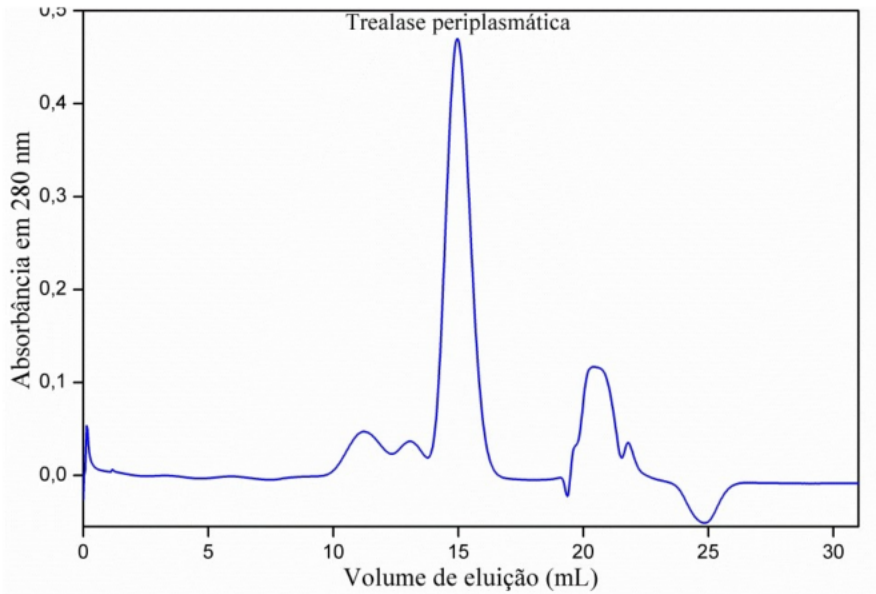
The Logistic Regression model slightly outperformed the Random Forest Classifier before any optimization took place.

	LOG	RF
• Accuracy:	0.886,	0.877
• ROC AUC:	0.933,	0.927

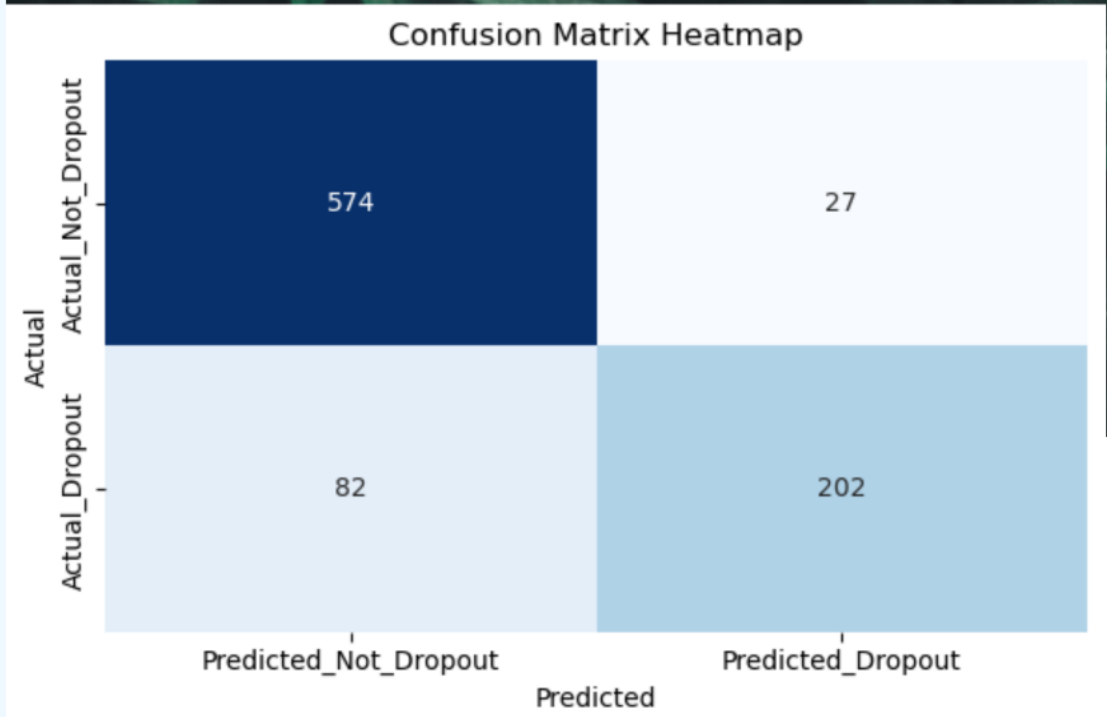


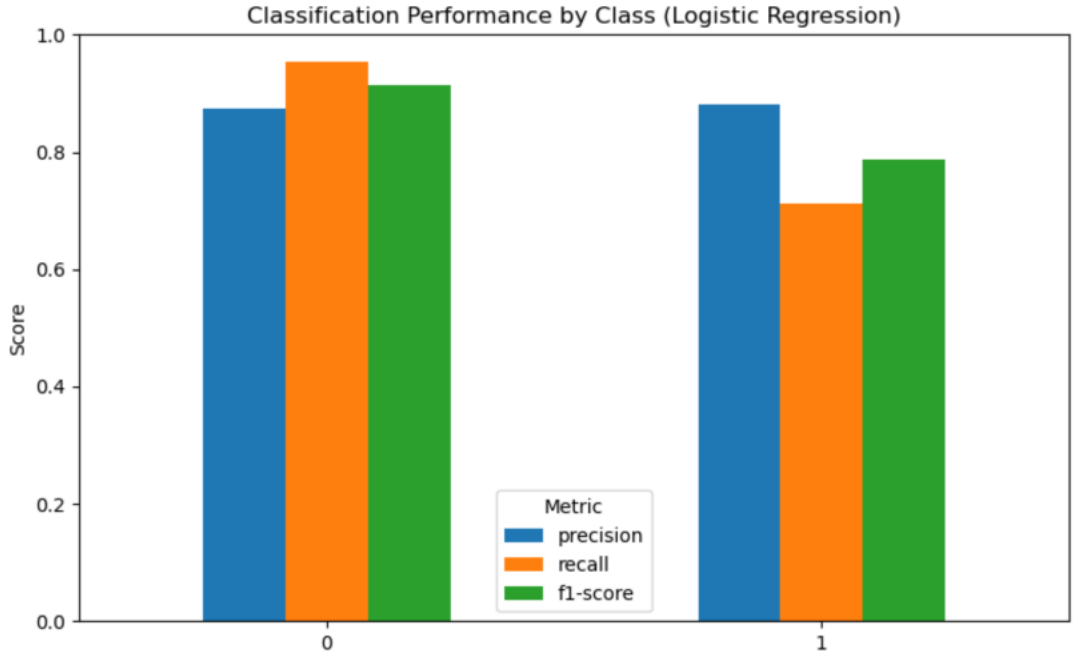
Data Exploration Techniques

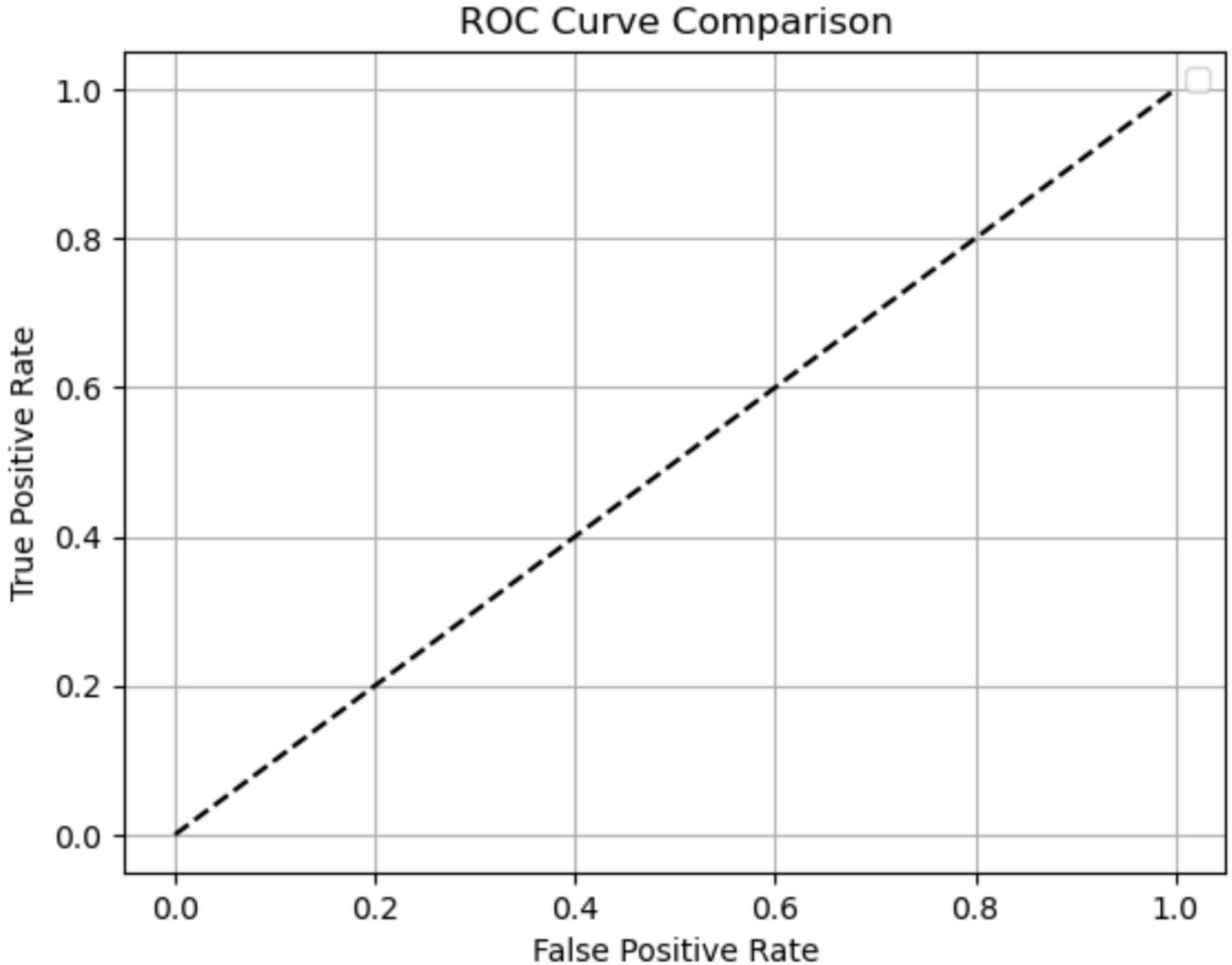
Exploratory data analysis (EDA) was performed using a Heatmap, and visualization of a bar graph to identify trends, patterns, and correlations. This analysis informed feature selection and contributed to our collective ability to evaluate the data. .



Data Visualizations






ROC Curve Comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Model Optimization & Evaluation

- Optimization: Define Hyperparameters for Tuning
- Identify the key hyperparameters to tune. Common ones include:
- n_estimators: The number of decision trees in the forest.
- max_depth: The maximum depth of each decision tree.
- max_features: The maximum number of features to consider at each split.
- min_samples_split: The minimum number of samples required to split an internal node.
- min_samples_leaf: The minimum number of samples required to be at a leaf node.
- Use RandomizedSearchCV for efficient tuning. It randomly samples from the defined hyperparameter space.
- After tuning, evaluate the best model found by RandomizedSearchCV.

Results & Key Insights

- Logistic Regression slightly outperformed Random Forest:

- For feature selection: extract the feature importances.
- Rank Features: Rank the features based on their importance scores in descending order.
- Set a Threshold or Number of Features to Keep: Decide how many features to keep or set a threshold for the importance score.
- Threshold: Keep only the features with an importance score above a certain threshold. This threshold could be based on your understanding of the data, experimentation, or a specific business need.
- Number of Features: Keep the top N features based on their importance scores.
- Retrain the Model: Retrain the model using only the selected features and evaluate its performance on the test set.
- Compare Performance: Compare the performance of the model trained on the selected features with the performance of the model trained on the full feature set.
- Iterate: Experiment with different thresholds or numbers of features to keep, and evaluate the model's performance to find the optimal feature set.

Key Findings

- Top Features from Random Forest:

-Admission Grade – Higher grades = better outcomes
-Curricular Unit Performance – Strong predictor of success
-Scholarship Holder – Linked to higher success rates
-Economic Indicators – Moderate impact (e.g., unemployment rate, GDP)

Conclusions

- Expand scholarship programs to support students from economically disadvantaged backgrounds.

- Implement systems to track student performance in real-time and flag those who may need additional resources.

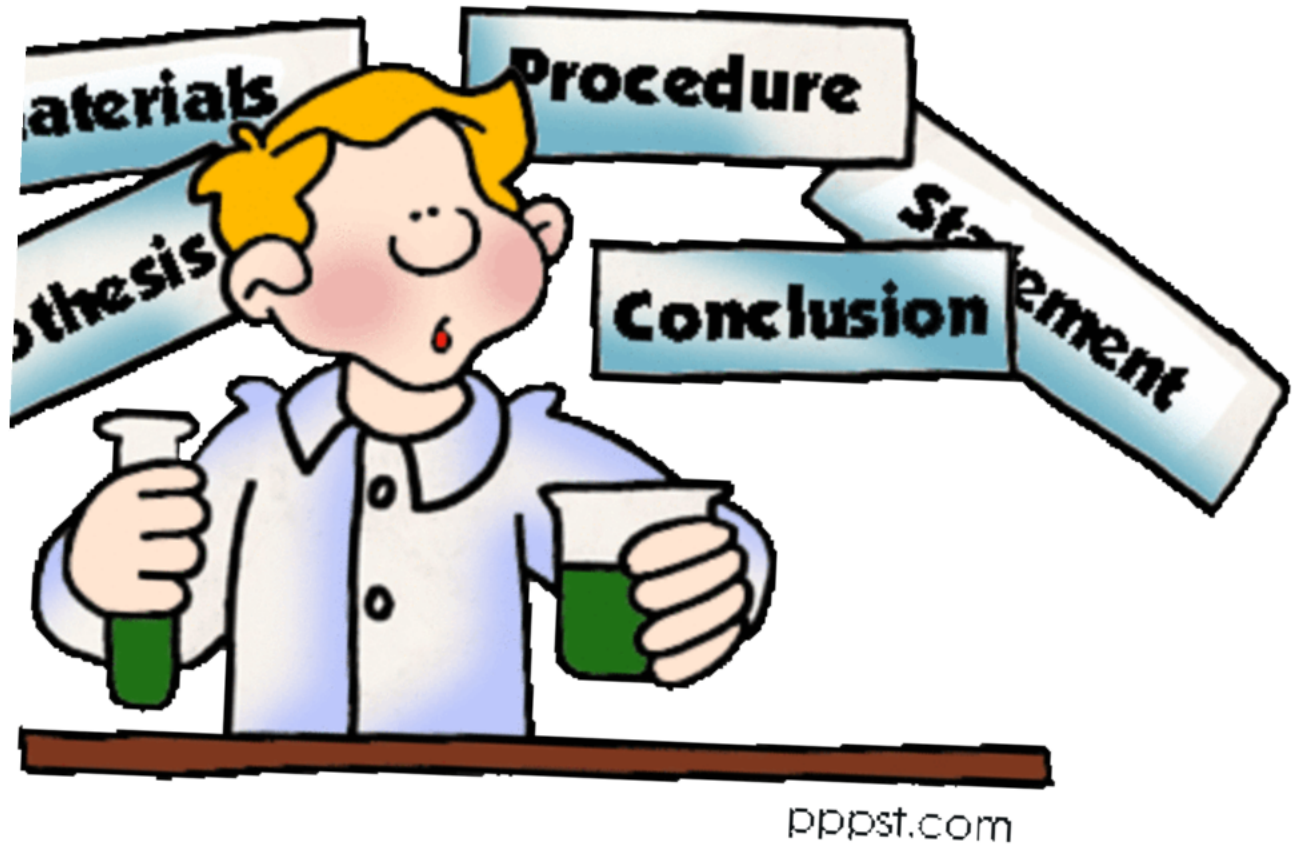
- Pair students with mentors who can provide guidance and encouragement throughout their academic journey.

- Advocate for policies that improve economic conditions, as external factors like unemployment and inflation significantly impact student outcomes.

Final Hypothesis

- Dropout rates are linked to academics, financial support, and economic factors.

- Students with low grades, poor course performance, and no scholarships are more likely to drop out.



Collaboration and Workflow

The project leveraged a collaborative team effort, utilizing version control tools and concise methodologies for consistent communication. Regular meetings ensured alignment on objectives and allowed for efficient workflow adjustments as needed throughout the project.

Challenges Faced

Throughout the project, we encountered challenges such as the dataset reflecting a multi-class classification, and feature selection dilemmas. These obstacles required iterative solutions and adaptive strategies, ultimately enhancing our problem-solving capabilities.

Future Research Directions

Additional research could expand the dataset to include longitudinal studies to capture dropout trends over time. Exploring machine learning techniques may improve prediction accuracy further and provide deeper insights into the factors influencing student retention.