DSAI (SC1015) Project

FCE3

Student Dropout Prediction

By:
Denzel Elden Wijaya
Federrico Hansen Budianto
Reswara Anargya Dzakirullah

Introduction and Overview

Education is the cornerstone of personal growth and societal progress. Nowadays, dropout cases often occur, with 32.9% of undergraduates not completing their studies.

To address this pressing issue, we conduct a thorough examination of student dropout rates in school education, focusing particularly on undergraduate students.

Introduction and Overview

UCI Machine Learning Repository

Predict Students' Dropout and Academic Success

Donated on 12/12/2021

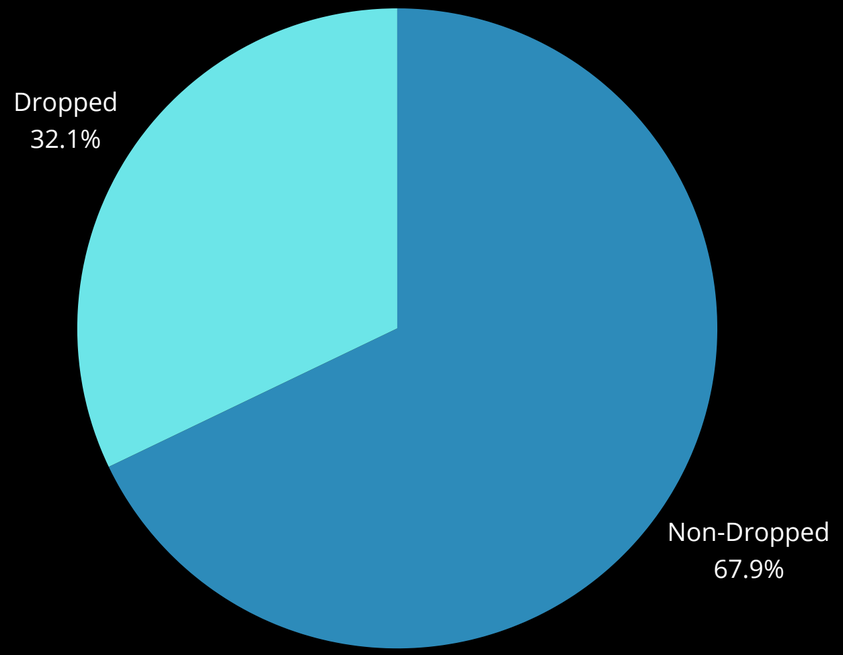
https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

Demographic Analysis

Economic Factor

Academic Performance

Social and Special Needs



Sample COLLECTION

Practical MOTIVATION

Introduction and Overview

UCI Machine Learning Repository

Predict Students' Dropout and Academic Success

Donated on 12/12/2021

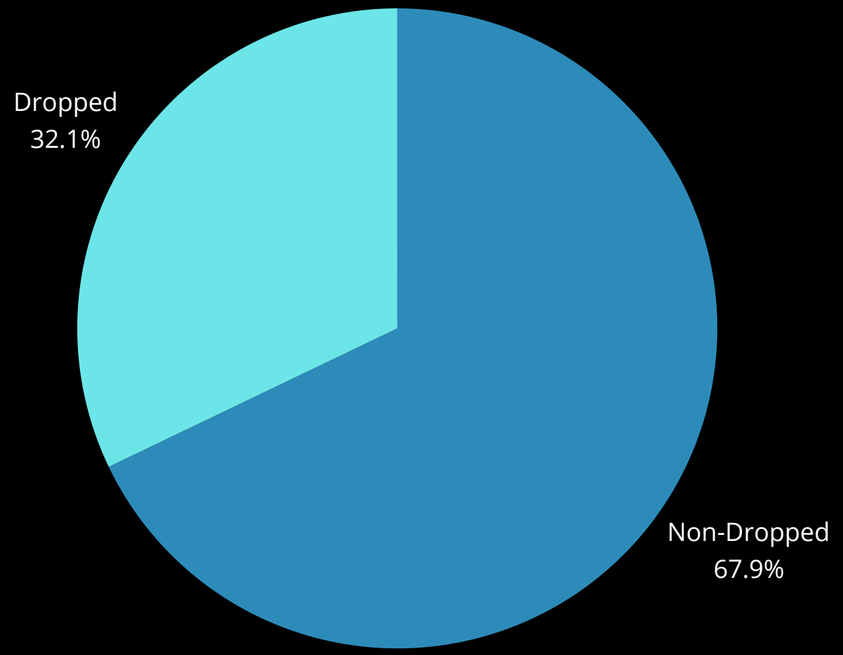
https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

1. **Demographic Analysis:** investigate how demographic factors like gender, age at enrollment, marital status, and nationality correlate with dropout rates.

2. **Economic Factors**: explore the impact of economic factors such as parental occupation, tuition fee payment status, and eligibility for scholarships on dropout rates. Furthermore, we also analyze some other indicators such as unemployment rate, inflation rate, and GDP growth and dropout rates, considering their indirect effects on education outcomes.

3. **Academic Performance**: analyze how students' academic performance, including curricular units and evaluations, influences their likelihood of dropping out.

4. **Social and Special Needs**: examine whether students with educational special needs or facing unique challenges like displacement or debt are more prone to dropping out.



Sample COLLECTION

Practical MOTIVATION

Introduction and Overview

Column Name	Description
Marital status	The marital status of the student. (Categorical)
Application mode	The method of application used by the student. (Categorical)
Application order	The order in which the student applied. (Numerical)
Course	The course taken by the student. (Categorical)
Daytime/evening attendance	Whether the student attends classes during the day or in the evening. (Categorical)
Previous qualification	The qualification obtained by the student before enrolling in higher education. (Categorical)
Previous qualification (grade)	The grade of previous qualification between 0 and 200 (Numerical)
Nacionality	The nationality of the student. (Categorical)
Mother's qualification	The qualification of the student's mother. (Categorical)
Father's qualification	The qualification of the student's father. (Categorical)
Mother's occupation	The occupation of the student's mother. (Categorical)
Father's occupation	The occupation of the student's father. (Categorical)
Admission Grade	Admission grade between 0 and 200 (Numerical)
Displaced	Whether the student is a displaced person. (Categorical)
Educational special needs	Whether the student has any special educational needs. (Categorical)
Debtor	Whether the student is a debtor. (Categorical)
Tuition fees up to date	Whether the student's tuition fees are up to date. (Categorical)
Gender	The gender of the student. (Categorical)
Scholarship holder	Whether the student is a scholarship holder. (Categorical)
Age at enrollment	The age of the student at the time of enrollment. (Numerical)
International	Whether the student is an international student. (Categorical)

Some of the columns that we are going to use in our Notebook

Purpose & Questions

What is our target?

What are the biggest factors in the students' drop out rate?

Does parental background heavily impact the students' drop out rate?

Create a model to analyze student drop out rate

Data Preparation

Make the data suitable for further processing and analysis

Check values

NULL

```
Number of null values per column:
 Marital status                                  0
Application mode                                 0
Application order                                0
Course                                           0
Daytime/evening attendance\t                     0
Previous qualification                           0
Previous qualification (grade)                   0
Nationality                                      0
Mother's qualification                           0
Father's qualification                           0
Mother's occupation                              0
Father's occupation                              0
Admission grade                                  0
Displaced                                        0
Educational special needs                        0
Debtor                                           0
Tuition fees up to date                          0
Gender                                           0
Scholarship holder                               0
Age at enrollment                                0
International                                    0
Curricular units 1st sem (credited)              0
Curricular units 1st sem (enrolled)              0
Curricular units 1st sem (evaluations)           0
Curricular units 1st sem (approved)              0
Curricular units 1st sem (grade)                 0
Curricular units 1st sem (without evaluations)   0
Curricular units 2nd sem (credited)              0
Curricular units 2nd sem (enrolled)              0
Curricular units 2nd sem (evaluations)           0
Curricular units 2nd sem (approved)              0
Curricular units 2nd sem (grade)                 0
Curricular units 2nd sem (without evaluations)   0
Unemployment rate                                0
Inflation rate                                   0
GDP                                              0
Student Status                                   0
```

No null value

NA

No missing value

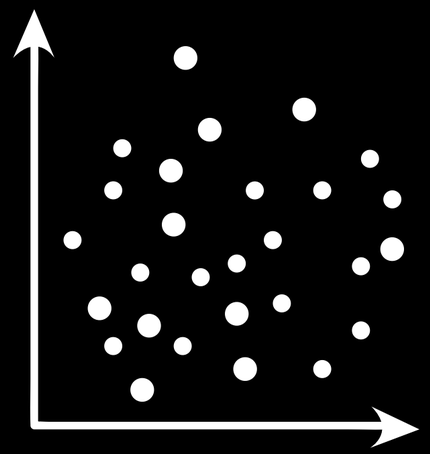
```
Number of missing values per column:
 Marital status                                  0
Application mode                                 0
Application order                                0
Course                                           0
Daytime/evening attendance\t                     0
Previous qualification                           0
Previous qualification (grade)                   0
Nationality                                      0
Mother's qualification                           0
Father's qualification                           0
Mother's occupation                              0
Father's occupation                              0
Admission grade                                  0
Displaced                                        0
Educational special needs                        0
Debtor                                           0
Tuition fees up to date                          0
Gender                                           0
Scholarship holder                               0
Age at enrollment                                0
International                                    0
Curricular units 1st sem (credited)              0
Curricular units 1st sem (enrolled)              0
Curricular units 1st sem (evaluations)           0
Curricular units 1st sem (approved)              0
Curricular units 1st sem (grade)                 0
Curricular units 1st sem (without evaluations)   0
Curricular units 2nd sem (credited)              0
Curricular units 2nd sem (enrolled)              0
Curricular units 2nd sem (evaluations)           0
Curricular units 2nd sem (approved)              0
Curricular units 2nd sem (grade)                 0
Curricular units 2nd sem (without evaluations)   0
Unemployment rate                                0
Inflation rate                                   0
GDP                                              0
Student Status                                   0
```

DUPLICATE

```
Number of duplicate values:  0
```

No duplicate value

Uncorrelated Variables With Dropout Status

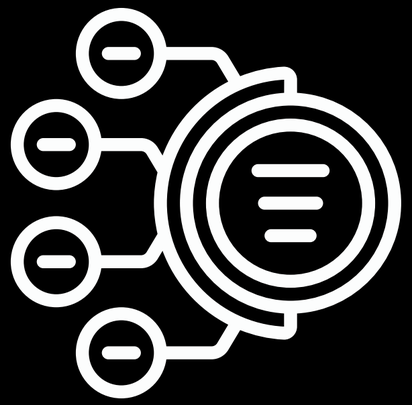


Using both Pearson correlation and Spearman’s correlation, find irrelevant variable and remove it

Spearman’s Rank Correlation approximately between 0.05 and -0.05

Curricular units 1st sem (without evaluations)	0.058391
Mother's qualification	0.039720
Inflation rate	0.023168
Unemployment rate	0.009534
Educational special needs	0.002806
Course	−0.002213
Nationality	−0.010045
International	−0.010360
Father's qualification	−0.010686
Curricular units 1st sem (credited)	−0.012024
Curricular units 2nd sem (credited)	−0.021780
Mother's occupation	−0.032522
Father's occupation	−0.035742
Curricular units 1st sem (evaluations)	−0.054826
GDP	−0.056601

Classifying Categories in Some Variable



Simplify and classify the categories into new sub-variables, from initial qualification and occupation variable to different sectors

- Mother's and Father's Qualification:

1. Below Elementary School
2. Above Elementary School below High School
3. Above High school below Degree
4. Above Degree

- Mother's and Father's Occupation:

1. Health Industry
2. Politic
3. Economic Industry
4. Office Workers including engineering
5. Education Industry
6. Technology Industry
7. Others

Cramer's V Statistic on Updated Variable

```
father occupations: 0.1363514919187367 2
mother occupations: 0.146768616063831 8
father qualifications: 0.11258781756277 3
mother qualifications: 0.135917051015240 24
```

Keep it or remove it?

Slightly increase but will cause an overfitting

Clean the Dataset

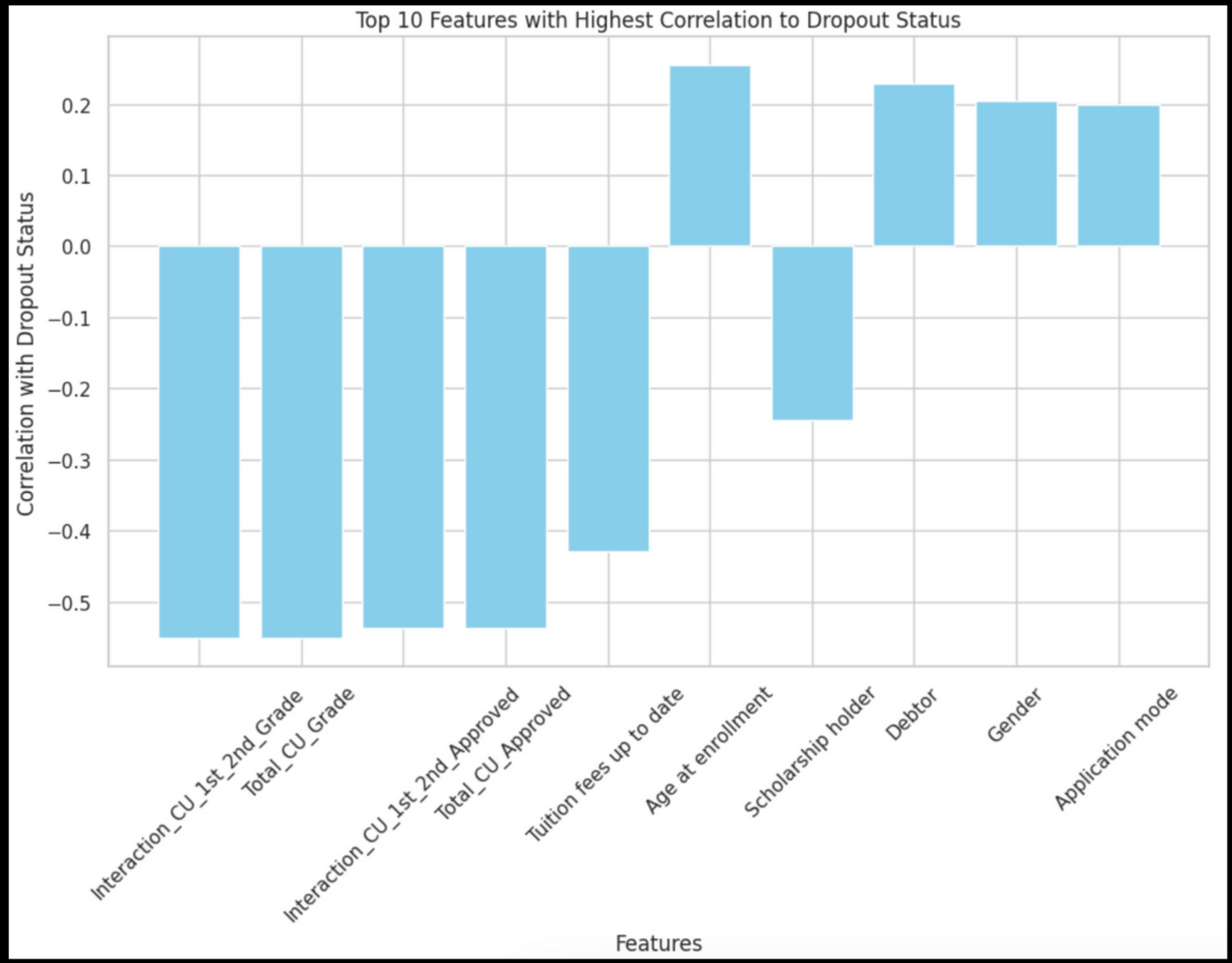
Keep	Remove
Age at enrollment Debtor Gender Application mode Previous qualification Marital status Curricular units 2nd sem (without evaluations)	Curricular units 1st sem (without evaluations) Mother's qualification Inflation rate Unemployment rate Educational special needs
Previous qualification (grade) Daytime/evening attendance\t Application order Admission grade Displaced Curricular units 2nd sem (evaluations) Curricular units 1st sem (enrolled)	Course Nationality International Father's qualification Curricular units 1st sem (credited)
Curricular units 2nd sem (enrolled) Scholarship holder Tuition fees up to date Curricular units 1st sem (grade) Curricular units 2nd sem (grade) Curricular units 1st sem (approved)	Curricular units 2nd sem (credited) Mother's occupation Father's occupation Curricular units 1st sem (evaluations) GDP

Exploratory Data Analysis

Go through the data more...

Identifying Key Predictors for Dropout Status

Presents a statistical analysis aimed at determining the top 10 features that are most strongly correlated with Dropout Status to obtain clear ranking

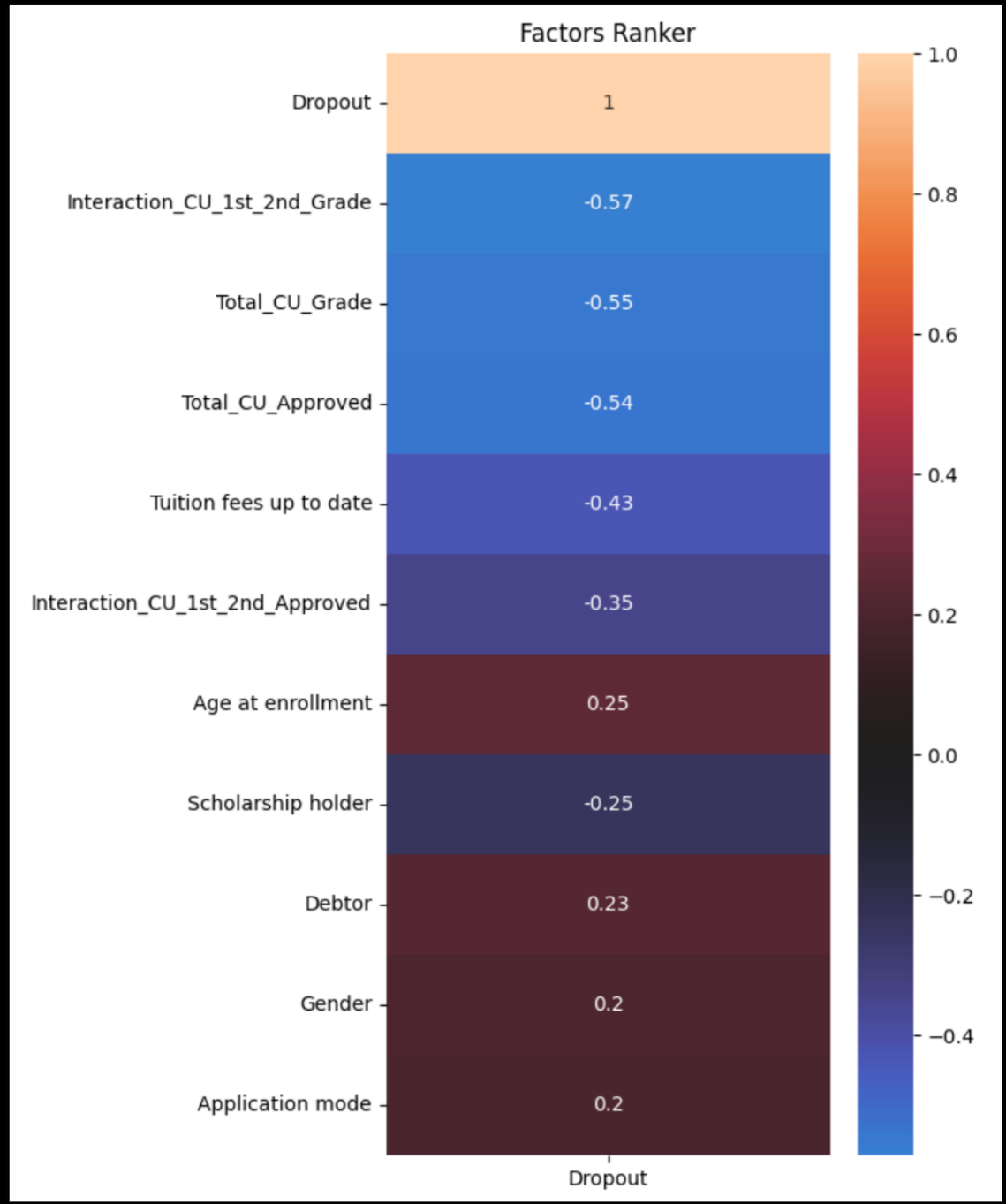


Exploratory ANALYSIS

Statistical DESCRIPTION

Factors Ranker Visualization

Offers a visually impactful and ordered representation of the correlation between Dropout Status and the top 10 features.



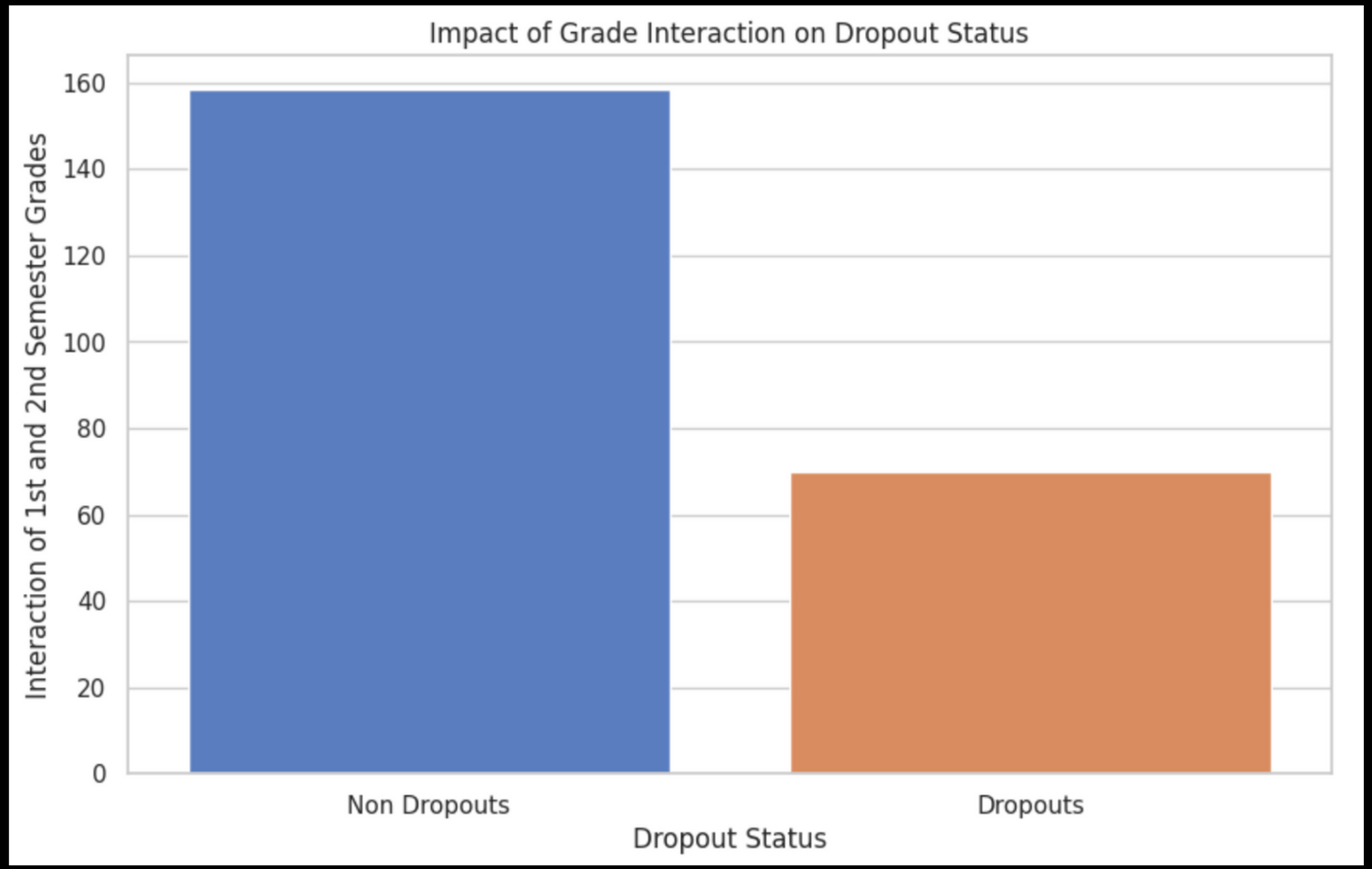
Exploratory ANALYSIS

Statistical DESCRIPTION

Sectors Impacting Dropout Status

1. Academic Performance
2. Financial Status
3. Age and Gender

Academic Performance and Its Impact on Dropout Status



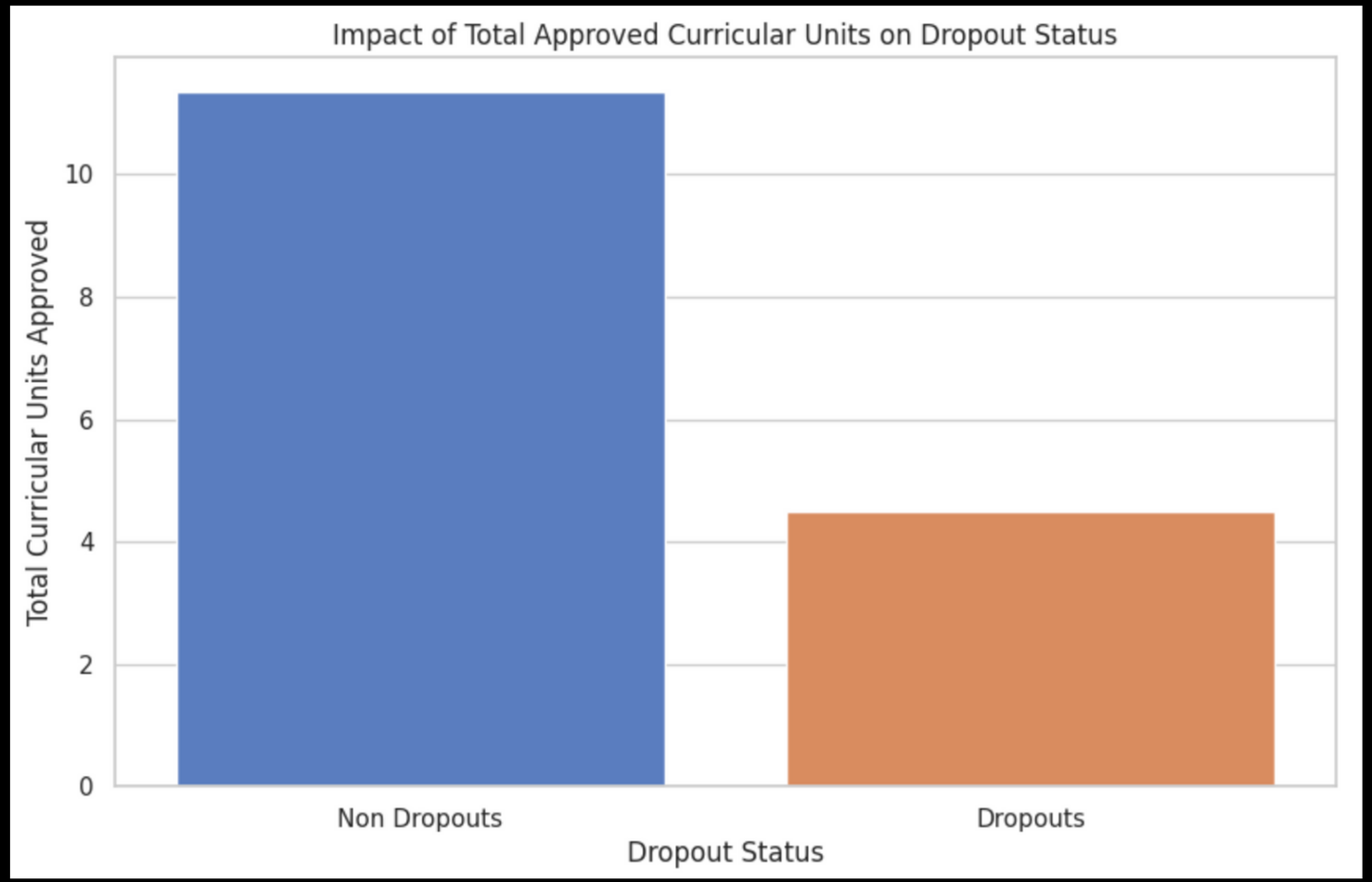
Dropout students have significantly lower combined grades than Non Dropouts

Combined performance over two semesters is a strong indicator of Dropout Status.

Analytic VISUALIZATION

Pattern RECOGNITION

Academic Performance and Its Impact on Dropout Status

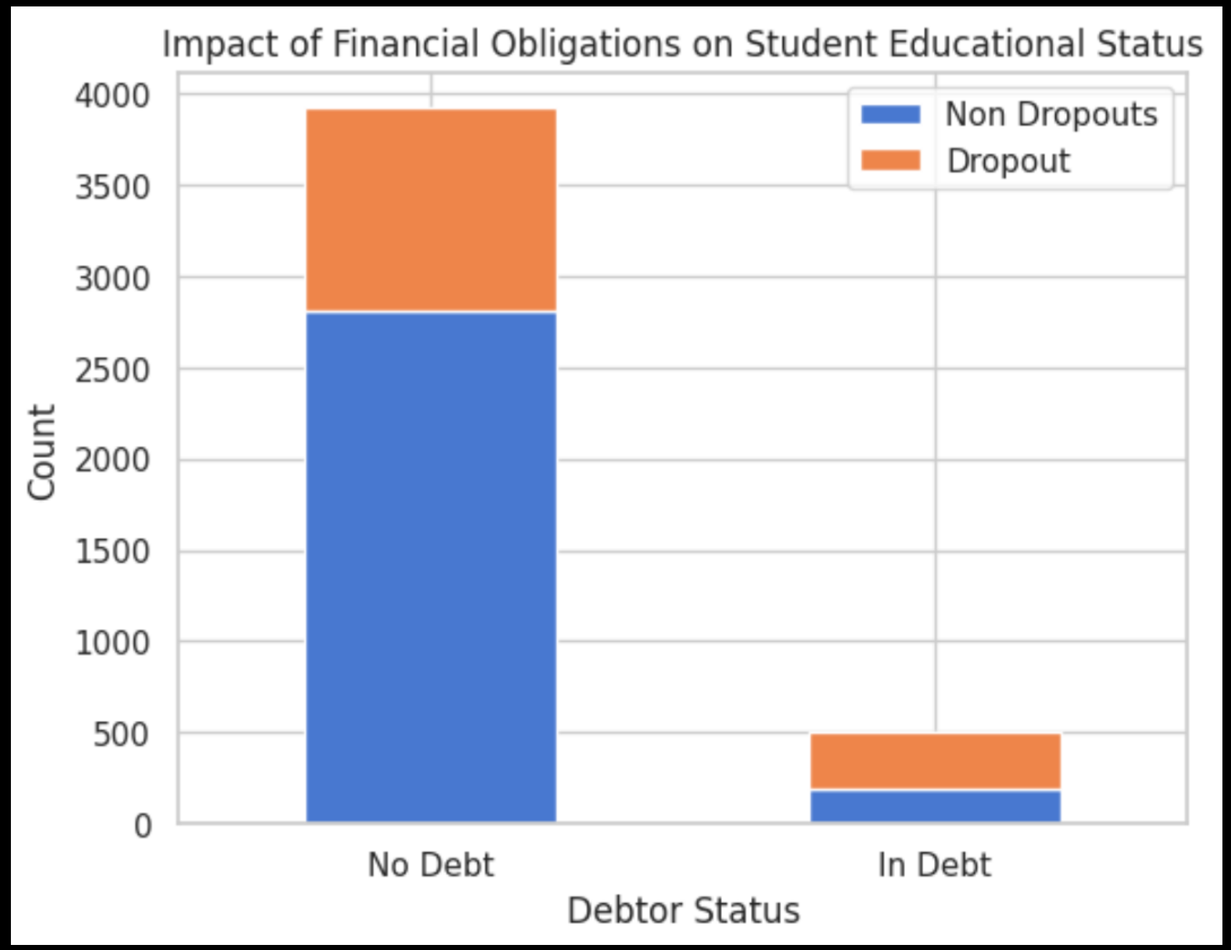


The dropout student's Total CU approved are less than half of the non dropouts, indicating this feature has significant impact on dropout status

Analytic VISUALIZATION

Pattern RECOGNITION

The Role of Financial Status in Educational Progress

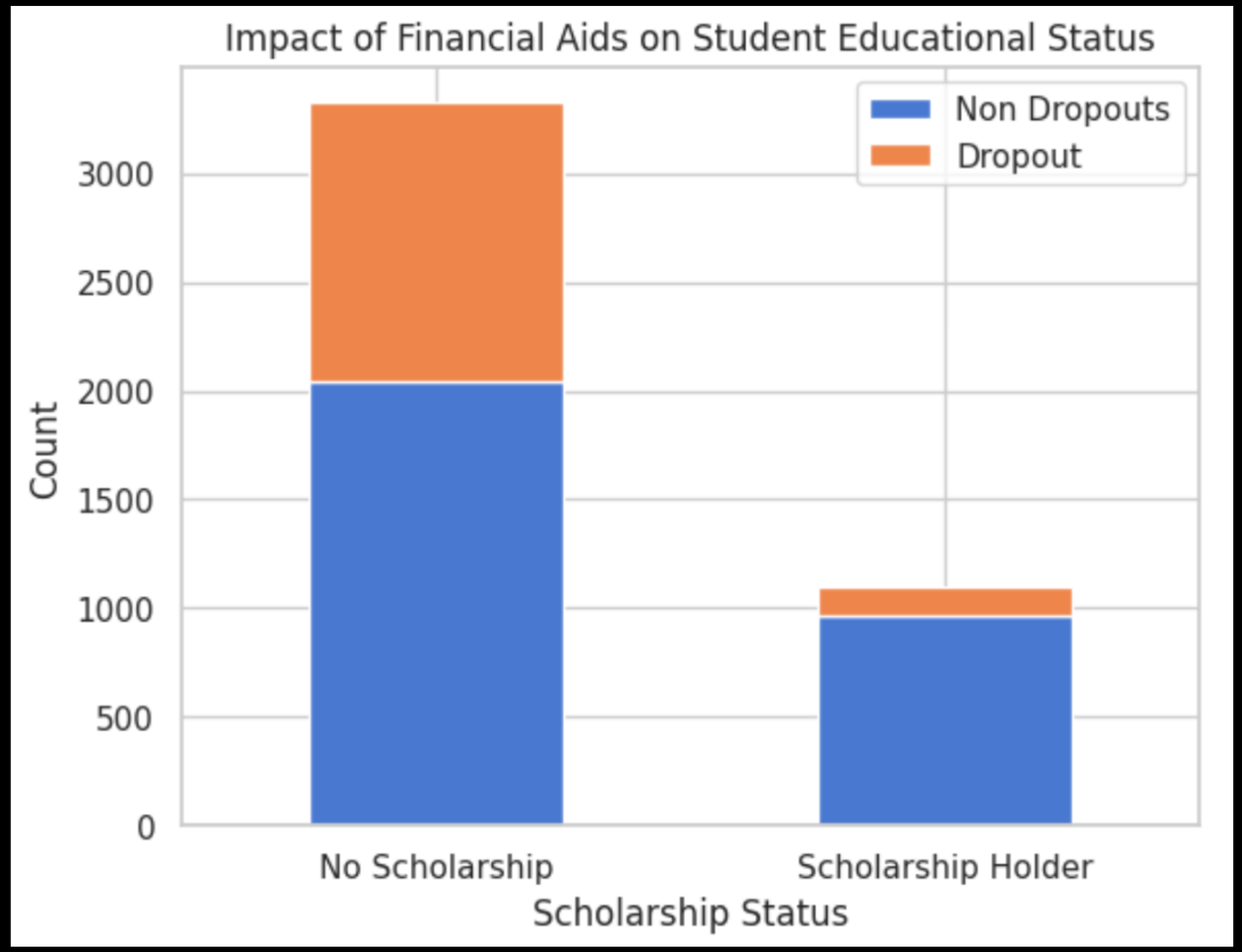


The proportion of students who dropout is noticeably larger within the indebted group, suggesting a detrimental effect of financial obligations on educational continuity.

Analytic VISUALIZATION

Pattern RECOGNITION

The Role of Financial Status in Educational Progress

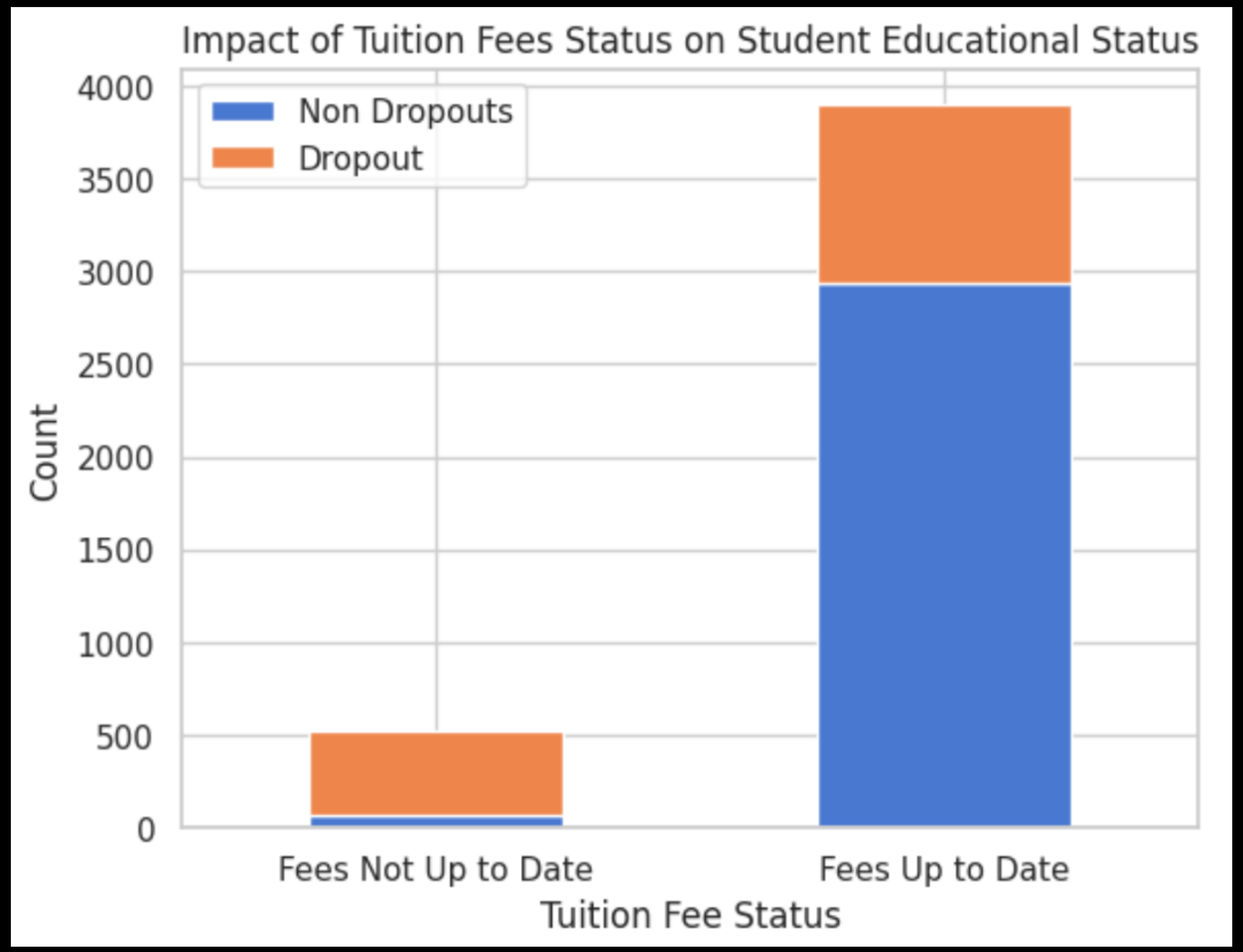


A majority of dropout students comes from group with no scholarships, which could indicate that financial support is a significant factor in helping students complete their education without dropped out.

Analytic VISUALIZATION

Pattern RECOGNITION

The Role of Financial Status in Educational Progress

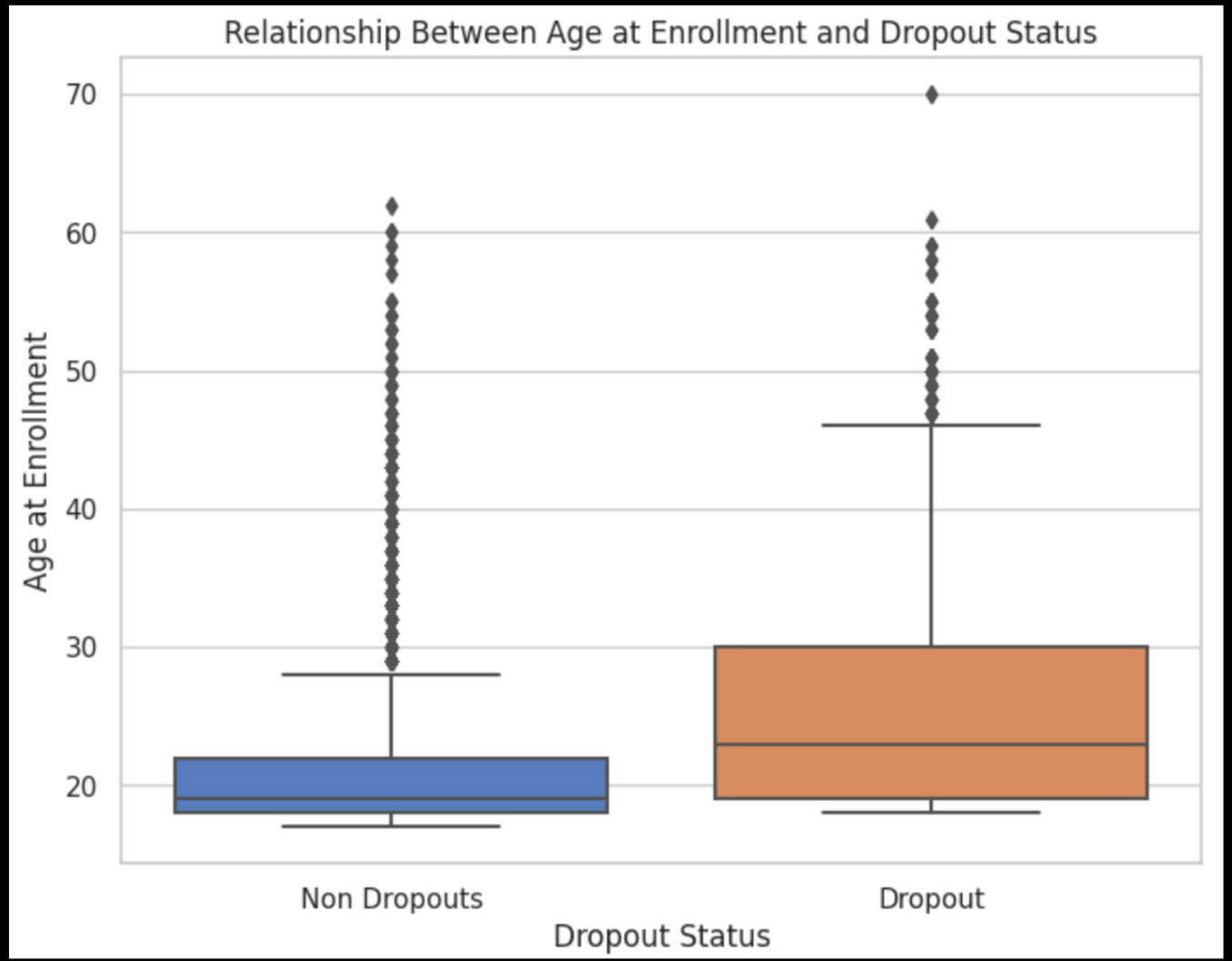


Students with outstanding tuition fees have a higher likelihood of dropping out, making this factor a strong predictor of educational discontinuation.

Analytic VISUALIZATION

Pattern RECOGNITION

The Interplay of Age and Gender in Academic Progression

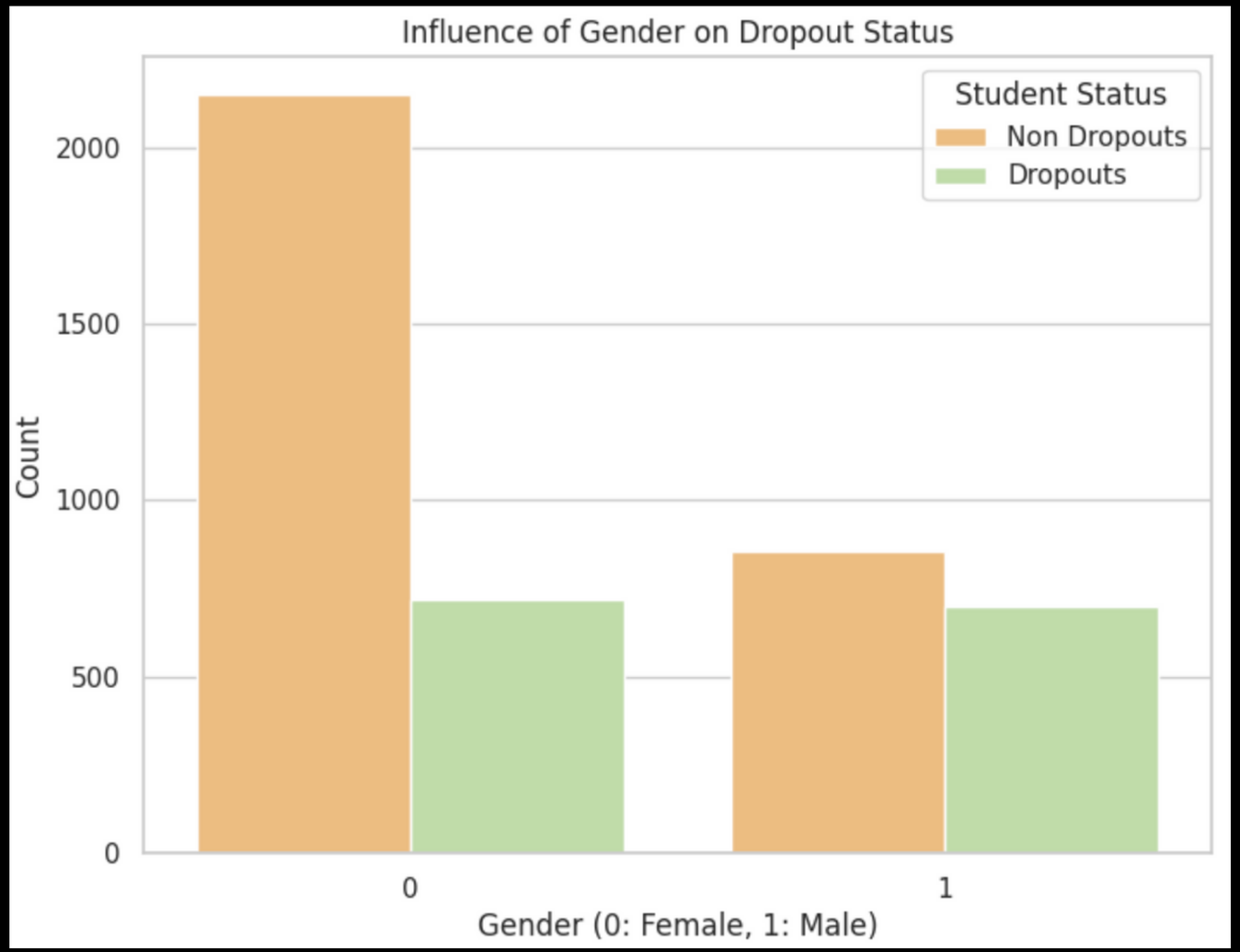


Non-dropped students tend to be younger, with a narrower age range suggesting a more uniform path to completion. In contrast, dropouts have a wider age range, indicating that dropping out affects a broader demographic.

Analytic VISUALIZATION

Pattern RECOGNITION

The Interplay of Age and Gender in Academic Progression



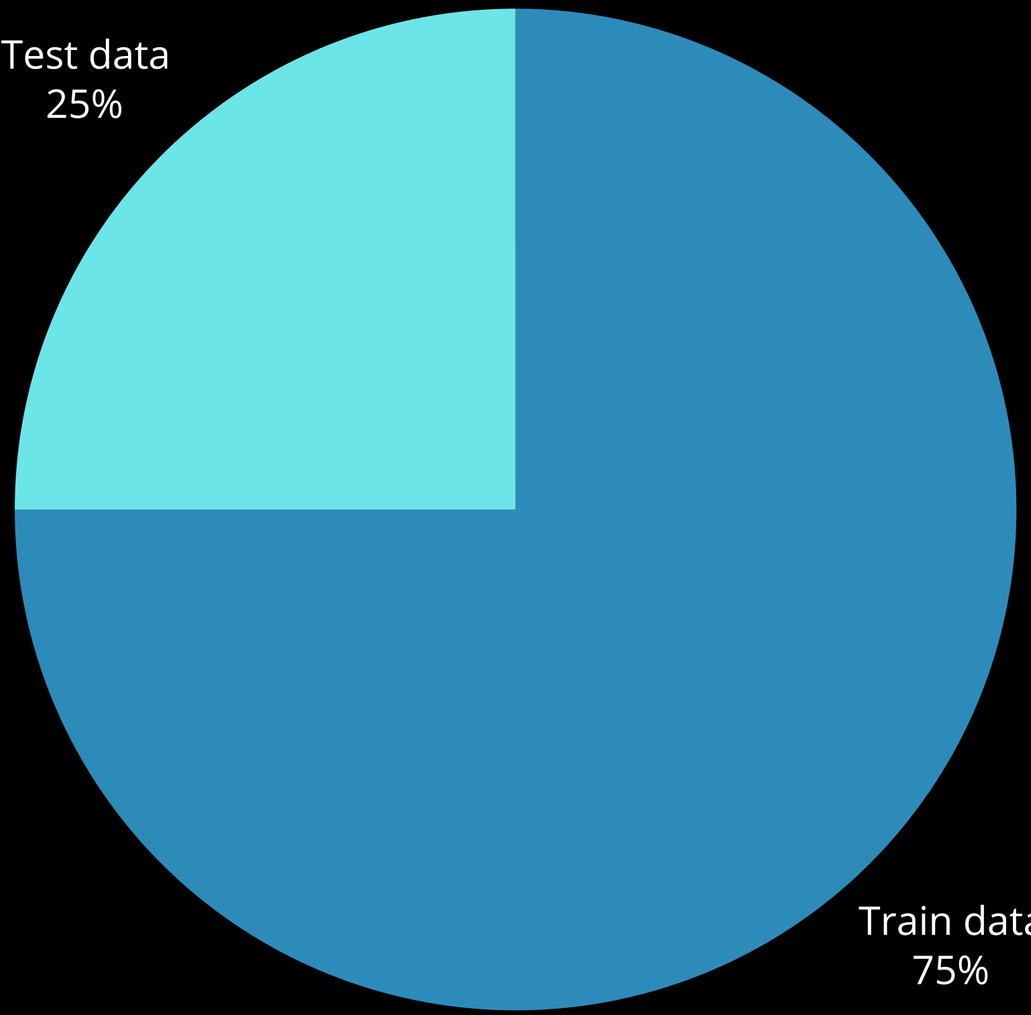
Non dropouts students are majority female. This visual suggests potential gender-based disparities in academic persistence or success that warrant further exploration.

Analytic VISUALIZATION

Pattern RECOGNITION

Machine Learning

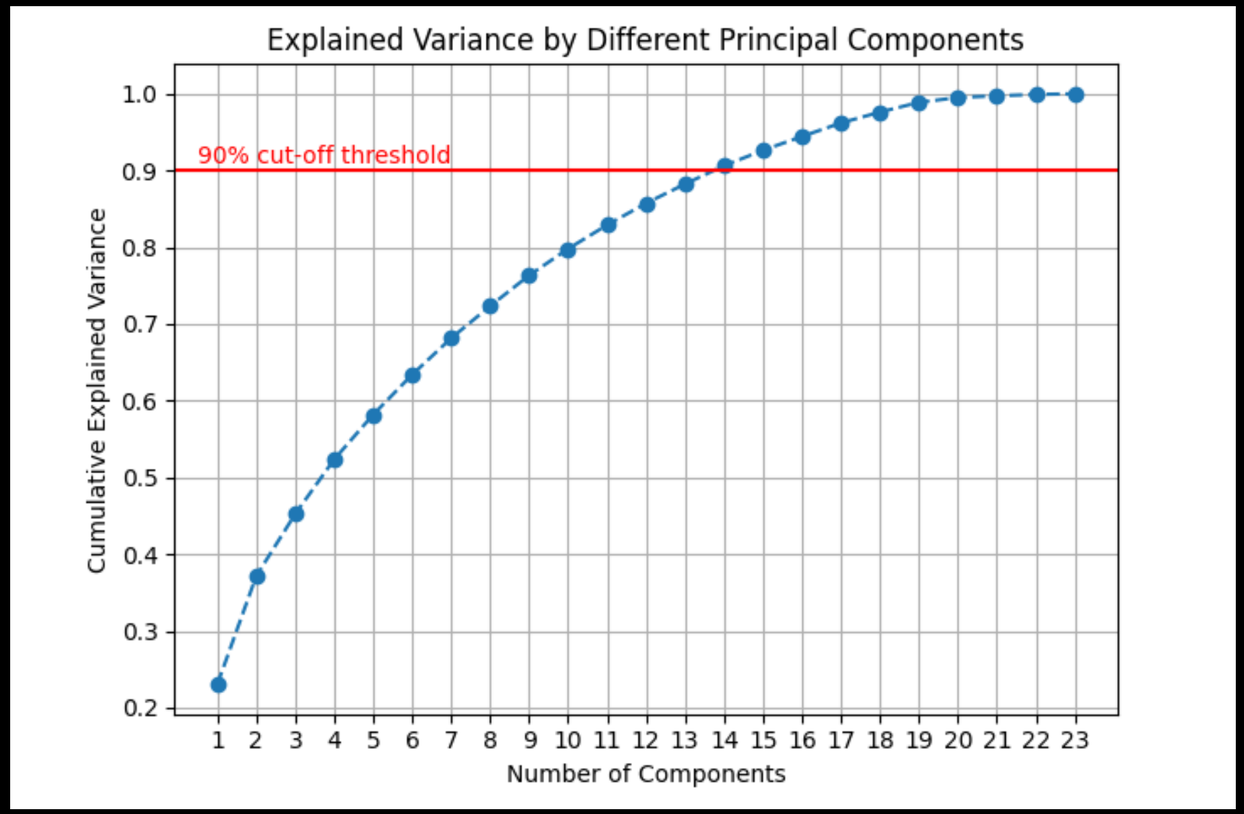
Let the machine decide...



Algorithmic
OPTIMIZATION

Machine
LEARNING

PCA



Reduce dimension

Algorithmic OPTIMIZATION

Machine LEARNING

Machine learning model

1. SVM
2. Logistic Regression
3. KNN
4. Random Forest

Algorithmic OPTIMIZATION

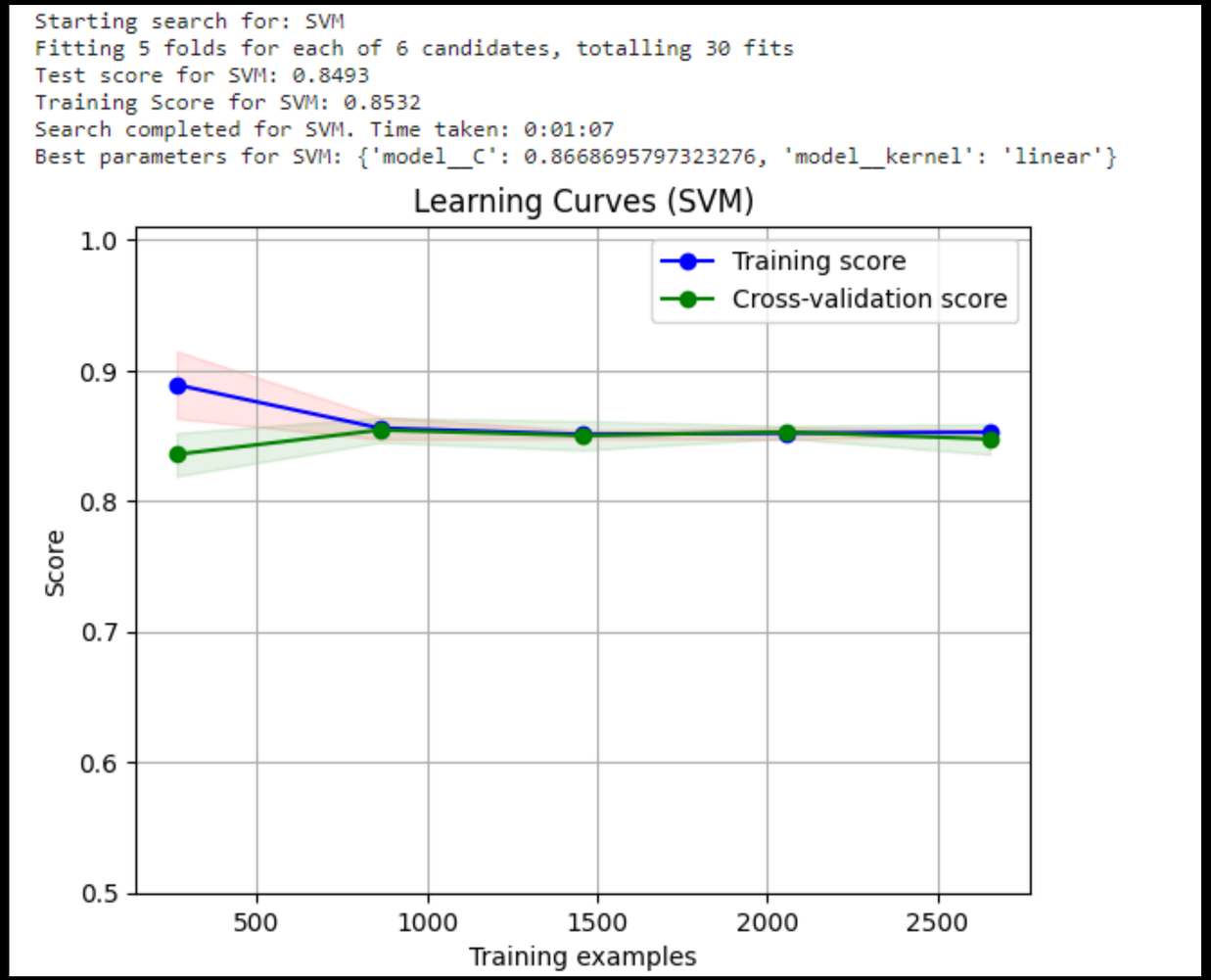
Machine LEARNING

Machine learning model

1

SVM

```
Starting search for: SVM
Fitting 5 folds for each of 6 candidates, totalling 30 fits
Test score for SVM: 0.8493
Training Score for SVM: 0.8532
Search completed for SVM. Time taken: 0:01:07
Best parameters for SVM: {'model__C': 0.8668695797323276, 'model__kernel': 'linear'}
```



Algorithmic OPTIMIZATION

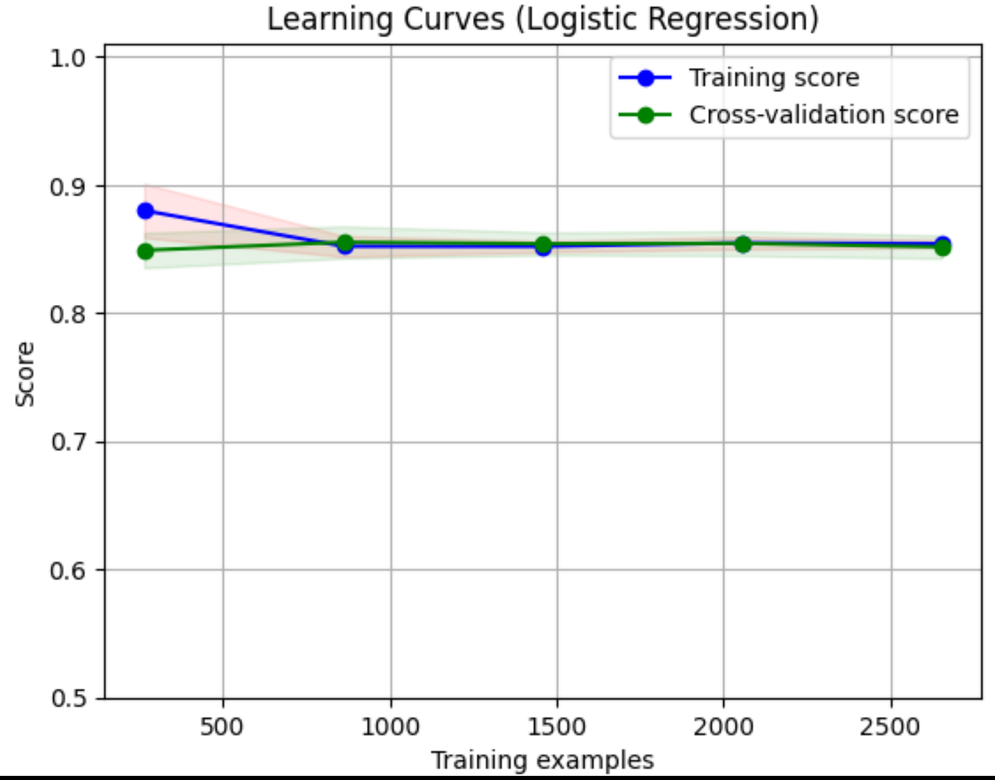
Machine LEARNING

Machine learning model

2

Logistic Regression

```
Starting search for: Logistic Regression
Fitting 5 folds for each of 6 candidates, totalling 30 fits
Test score for Logistic Regression: 0.8514
Training Score for Logistic Regression: 0.8526
Search completed for Logistic Regression. Time taken: 0:00:05
Best parameters for Logistic Regression: {'model__solver': 'liblinear', 'model__penalty': 'l2', 'model__
C': 0.005}
```



Algorithmic OPTIMIZATION

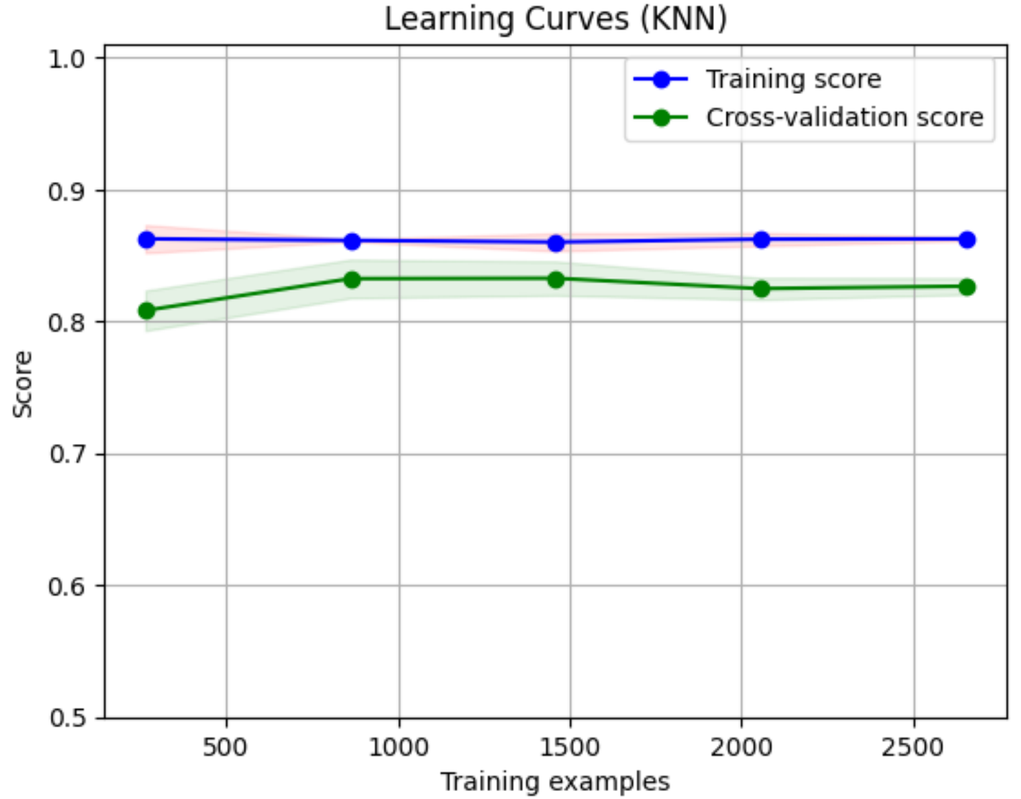
Machine LEARNING

Machine learning model

3

KNN

```
Starting search for: KNN
Fitting 5 folds for each of 6 candidates, totalling 30 fits
Test score for KNN: 0.8234
Training Score for KNN: 0.8659
Search completed for KNN. Time taken: 0:00:05
Best parameters for KNN: {'model__n_neighbors': 12, 'model__weights': 'uniform'}
```




```
Total time for all model searches: 0:02:14
```

Algorithmic OPTIMIZATION

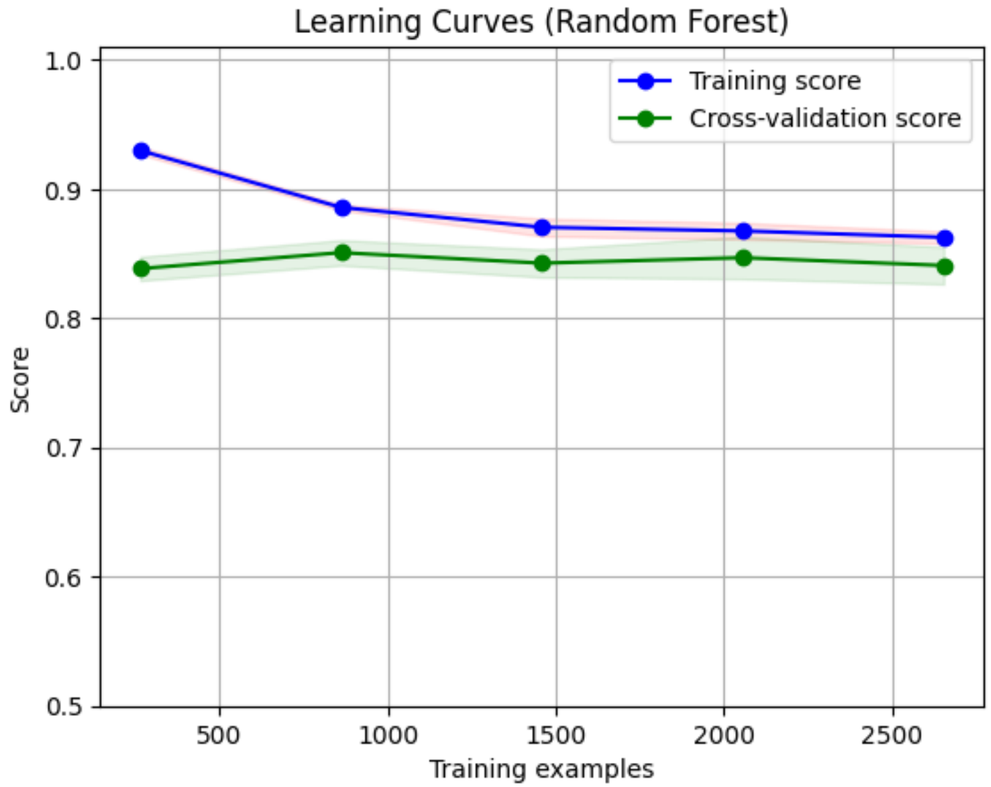
Machine LEARNING

Machine learning model

4

Random Forest

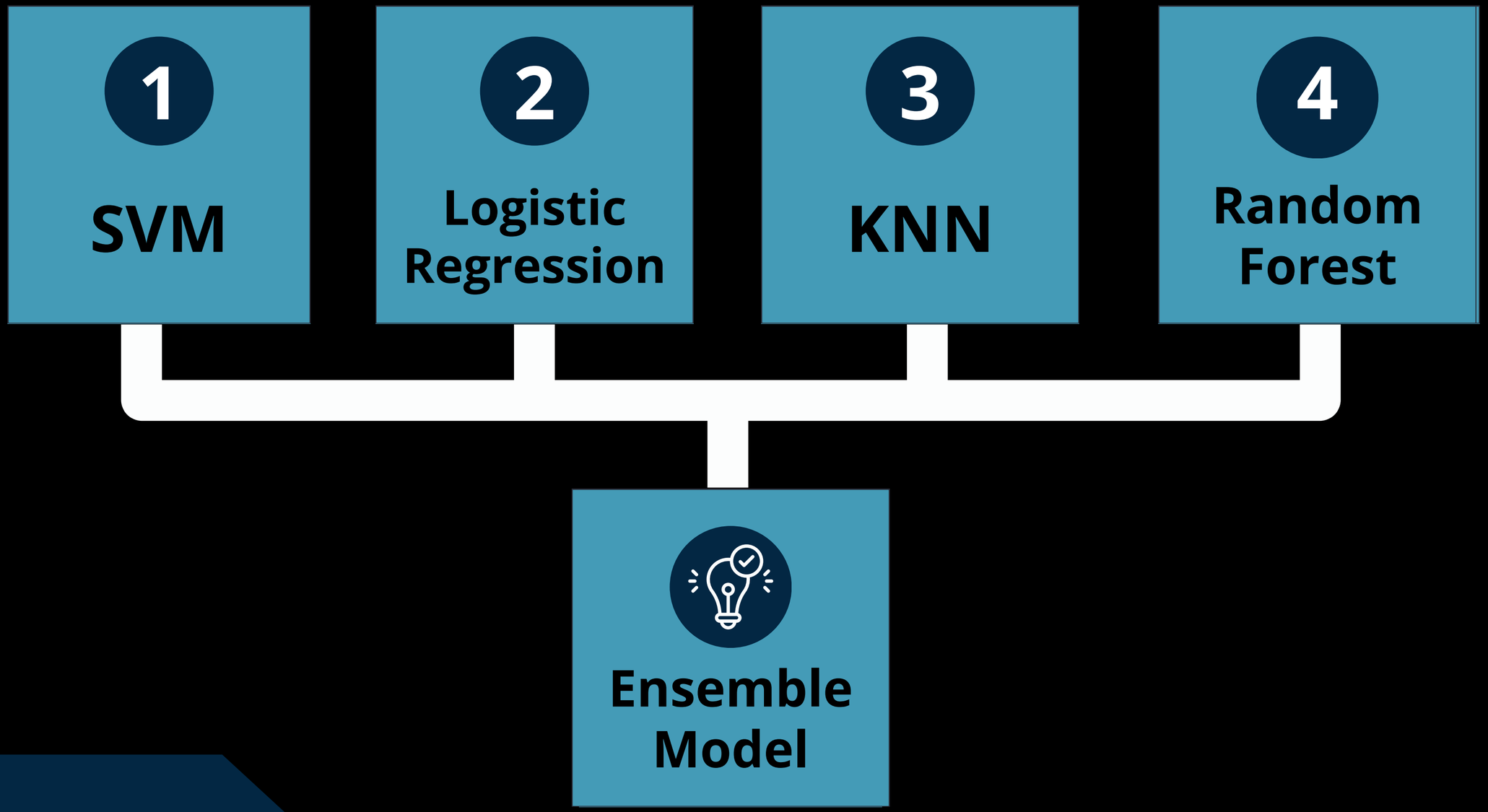
```
Starting search for: Random Forest
Fitting 5 folds for each of 6 candidates, totalling 30 fits
Test score for Random Forest: 0.8427
Training Score for Random Forest: 0.8626
Search completed for Random Forest. Time taken: 0:00:30
Best parameters for Random Forest: {'model__n_estimators': 100, 'model__min_samples_split': 6, 'model__m
in_samples_leaf': 3, 'model__max_features': 'log2', 'model__max_depth': 4}
```



Algorithmic OPTIMIZATION

Machine LEARNING

Machine learning model



Algorithmic OPTIMIZATION

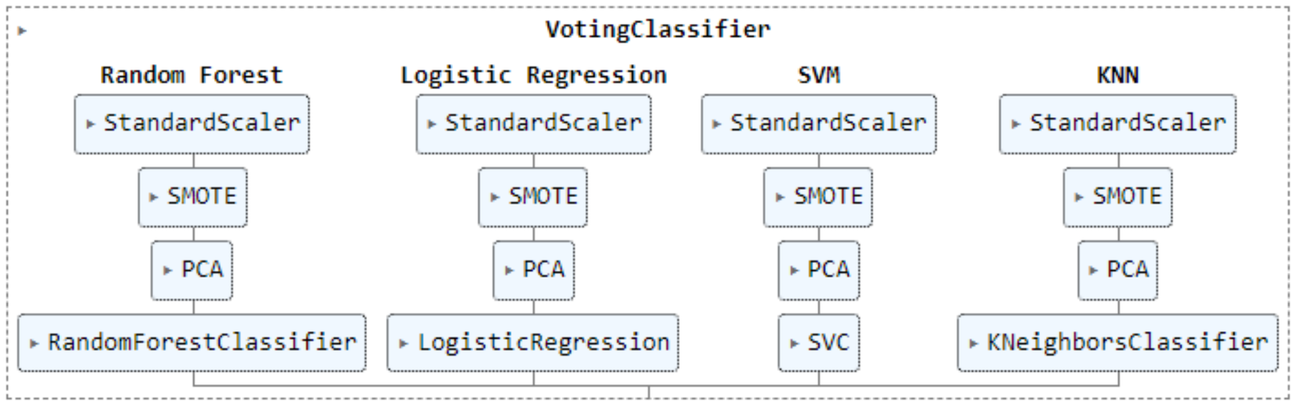
Machine LEARNING

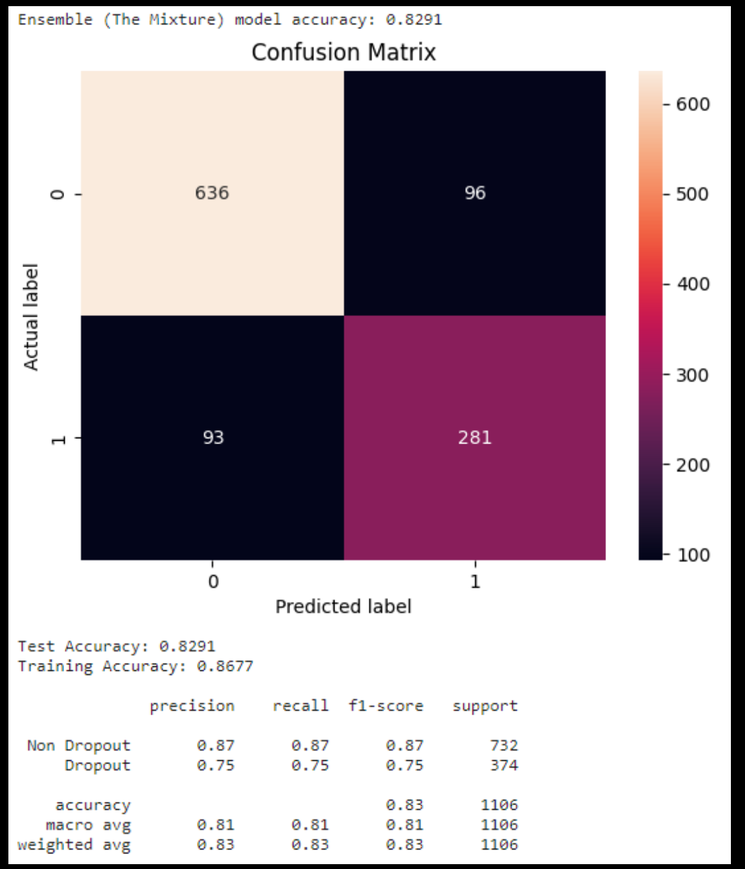
Machine learning model

```python
from sklearn.ensemble import VotingClassifier

mix = VotingClassifier(estimators = [(name, model)
                                     for name, model in best_models.items()], voting = '

mix.fit(X_train, y_train)
```





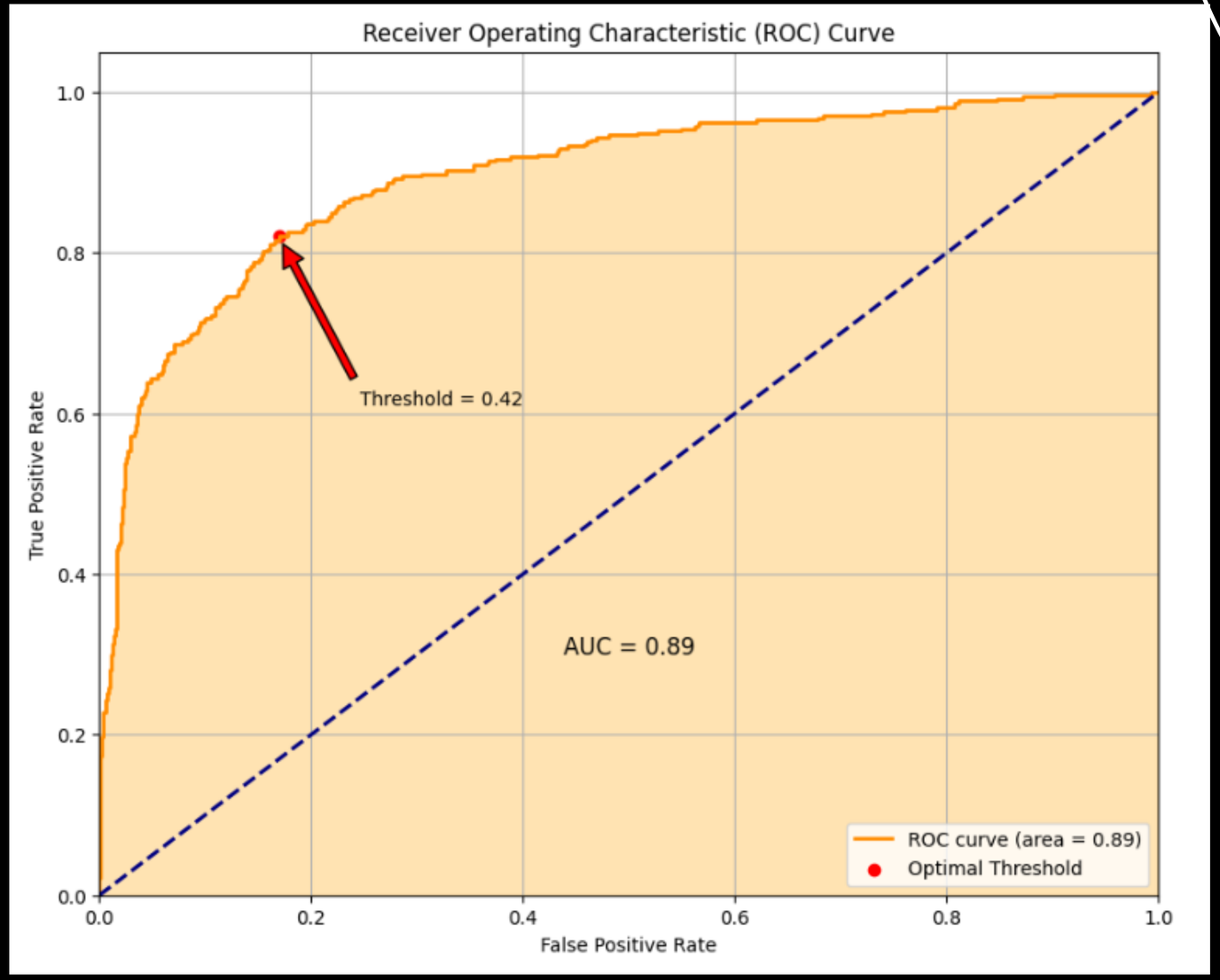
Test Accuracy: 0.8291
Training Accuracy: 0.8677

	precision	recall	f1-score	support
Non Dropout	0.87	0.87	0.87	732
Dropout	0.75	0.75	0.75	374
accuracy			0.83	1106
macro avg	0.81	0.81	0.81	1106
weighted avg	0.83	0.83	0.83	1106

Algorithmic OPTIMIZATION

Machine LEARNING

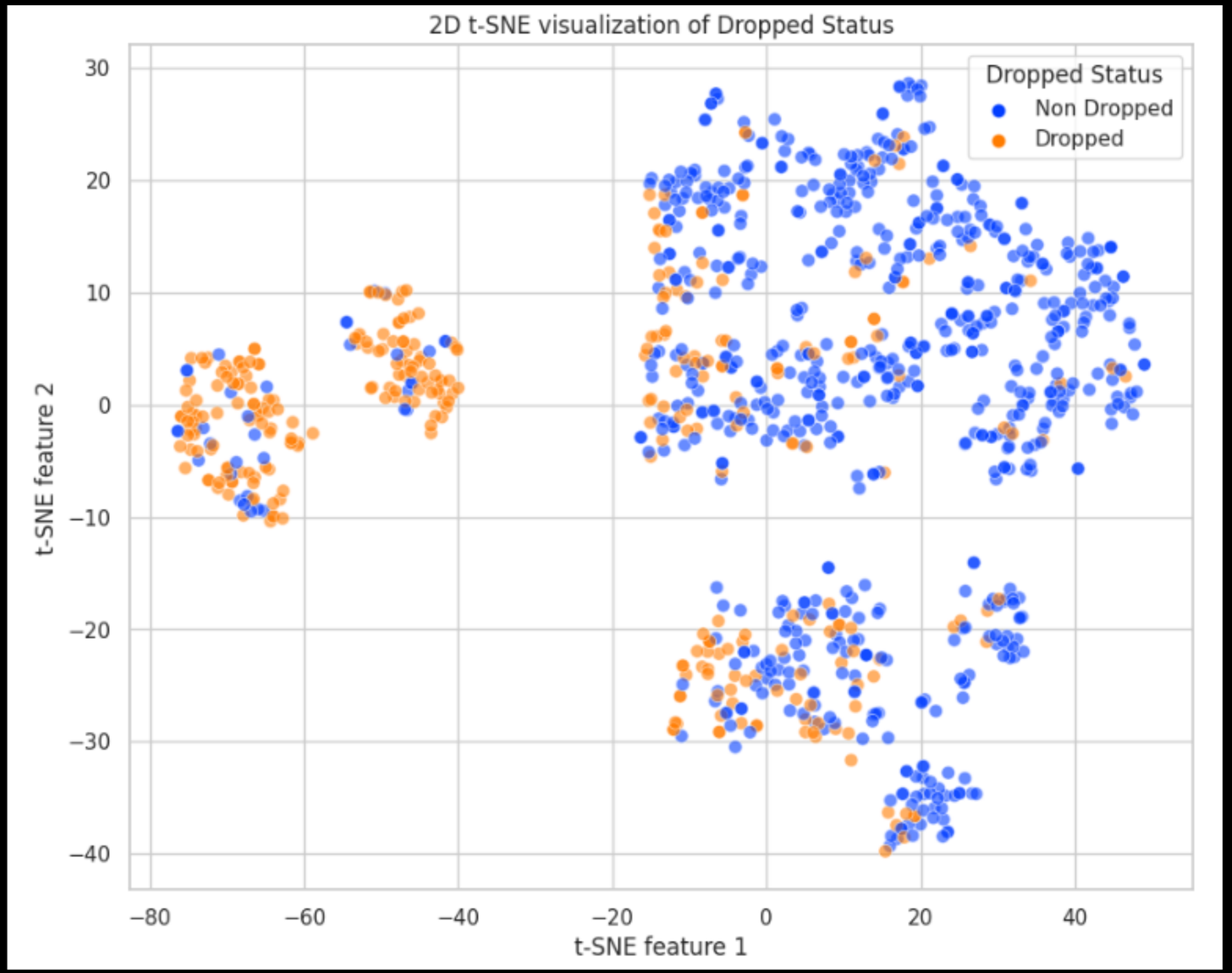
ROC



Algorithmic OPTIMIZATION

Machine LEARNING

T-SNE



Algorithmic OPTIMIZATION

Machine LEARNING

Conclusion

Impactful Factors:

Academic Performance

1. **Interaction_CU_1st_2nd_Grade (0.57)**
2. **Total_CU_Approved (0.55)**
3. **Total_CU_Grade (0.54)**
4. **Interaction_CU_1st_2nd_Approved(0.35)**
5. **Application Mode (0.2)**

Ethical CONSIDERATION

Intelligent DECISION

Conclusion

Impactful Factors:

Economic Factors

1. **Tuition fees up to date (0.43)**
2. **Scholarship holder (0.25)**
3. **Debtor (0.23)**

Ethical CONSIDERATION

Intelligent DECISION

Conclusion

Impactful Factors:

Demographic Analysis

1. **Age at Enrollment (0.25)**
2. **Gender (0.2)**

Ethical CONSIDERATION

Intelligent DECISION

Conclusion

Machine Learning

Best one:

- **Logistic Regression**
- **Ensemble Model**

Consideration:

- **SVM**
- **Random Forest**

Ethical CONSIDERATION

Intelligent DECISION

Conclusion

Further Action

1. Academic Support Programs
2. Financial Aid and Scholarship
3. Mental health and Wellnes Services

Ethical CONSIDERATION

Intelligent DECISION

References

https://educationdata.org/college-dropout-rates

https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success



Thankyou :)