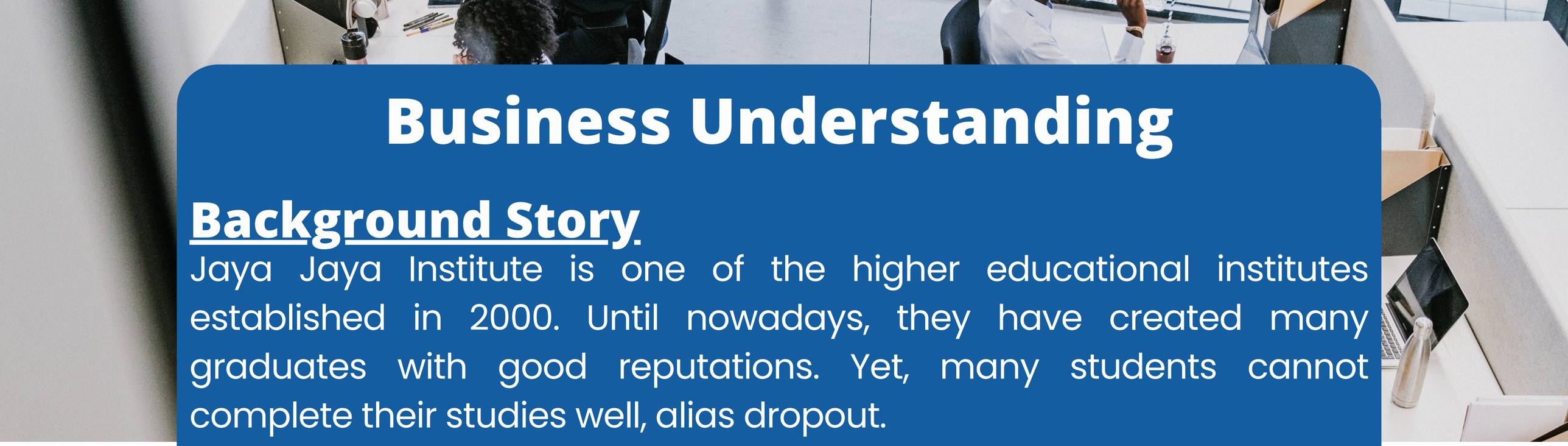
Educational Institution Problem

By: Norman Dwi Febrio



Business Understanding

Background Story

Jaya Jaya Institute is one of the higher educational institutes established in 2000. Until nowadays, they have created many graduates with good reputations. Yet, many students cannot complete their studies well, alias dropout.

This high dropout number indeed become a big problem for the institute. Therefore, Jaya Jaya Institute wants to detect students who are probably will dropout ASAP so that the institute can give special guidance.

Business Understanding

Business Problems

Problem 1 : High Dropoutted Students Number

The number of dropout student needs to be reduced to keep the institute's reputation and attraction. The higher dropout number reflects the dissatisfaction of students regarding education quality or other supports. This problem could cause a reduction in new registrants, a decrement of prospective students and their parents, and also make the institute reputation’s bad in a public point of view and other stakeholders.

Problem 2 : No Device nor System Which Could Monitor The Student’s Condition

Without an effective monitoring system, institutions find it difficult to identify potential dropout students in a timely manner and provide necessary interventions. This makes dropout prevention efforts less efficient and undirected. without a monitoring system, institutions cannot proactively address the problems of students experiencing academic or personal difficulties that could lead to dropouts

Data Understanding

A dataset created from a higher education institution (acquired from several disjoint databases) related to students enrolled in different undergraduate degrees, such as agronomy, design, education, nursing, journalism, management, social service, and technologies. The dataset includes information known at the time of student enrollment (academic path, demographics, and social-economic factors) and the students' academic performance at the end of the first and second semesters. The data is used to build classification models to predict students' dropout and academic sucess.

Feature Name	Description
Marital_status	**Marital status of Student**
Application_mode	**Kind of appliaction used by student to apply on the institute**
Application_order	**Application order (0 - 9)**
Course	**Kind of course choose by student**

Data Understanding

Feature Name	Description
Daytime_evening_attendance	Time attendance
Previous_qualification	Previous qualification of student
Previous_qualification_grade	Grade of the previous qualification
Nacionality	Nationality of student
Mothers_qualification	Student’s mother qualification
Fathers_qualification	Student’s father qualification
Mothers_occupation	Student’s mother occupation

Data Understanding

Feature Name	Description
Fathers_occupation	Student’s father occupation
Admission_grade	Admission grade (0 - 200)
Displaced	Is the student a displaced person?
Educational_special_ needs	Does the student with educational special need?
Debtor	Is the student a debtor?
Tuition_fees_up_to_date	Is the student’s tuition fees up to date?
Gender	Gender of the student

Data Understanding

Feature Name	Description
Scholarship_holder	Is the student a scholarship holder?
Age at enrollment	Previous qualification of student
Previous_qualification_grade	Grade of the previous qualification
Nacionality	Nationality of student
Mothers_qualification	Student’s mother qualification
Fathers_qualification	Student’s father qualification
Mothers_occupation	Student’s mother occupation

Data Understanding

Feature Name	Description
International	Is the student an international student?
Curricular_units_1st_sem_credited	Number of curricular units credited in the 1st semester
Curricular_units_1st_sem_enrolled	Number of curricular units enrolled in the 1st semester
Curricular_units_1st_sem_evaluations	Number of evaluations to curricular units in the 1st semester
Curricular_units_1st_sem_approved	Number of curricular units approved in the 1st semester
Curricular_units_1st_sem_grade	Grade average in the 1st semester (between 0 and 20)

Data Understanding

Feature Name	Description
Curricular_units_1st_sem_without_evaluations	Number of curricular units without evalutions in the 1st semester
Curricular_units_2nd_sem_credited	Number of curricular units credited in the 2nd semester
Curricular_units_2nd_sem_enrolled	Number of curricular units enrolled in the 2nd semester
Curricular_units_2nd_sem_evaluations	Number of evaluations to curricular units in the 2nd semester
Curricular_units_2nd_sem_approved	Number of curricular units approved in the 2nd semester
Curricular_units_2nd_sem_grade	Grade average in the 2nd semester (between 0 and 20)

Data Understanding

Feature Name	Description
Curricular_units_2nd_sem_without_evaluations	Number of curricular units without evalutions in the 2nd semester
Unemployment_rate	Unemployment rate (%)
Inflation_rate	Inflation rate (%)
GDP	GDP
Status	Graduate, Dropout, or Enrolled

Project Scope

Exploratory Data Analysis (EDA)

Develop A Predictive Model

Develop Prescriptive Dashboards

INSIGHTS

01
4424
Total Students

02
54.8%
Displaced Students

03
24.8%
Scholarship Holders

04
11.4%
Debtor Students

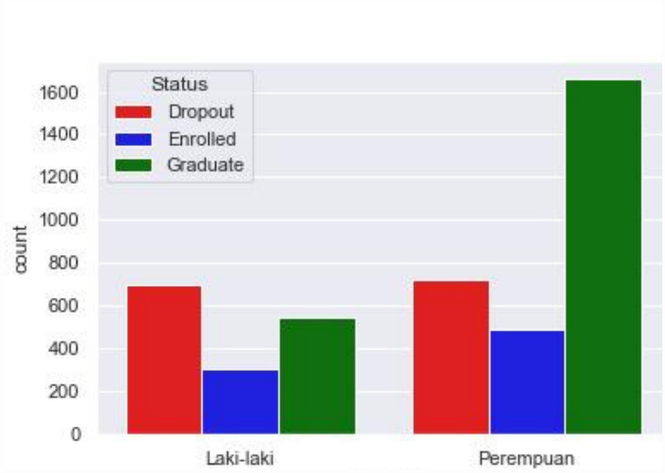
05
2.5%
International Students

06
1.2%
Students with Education Special Needs

Interesting Patterns

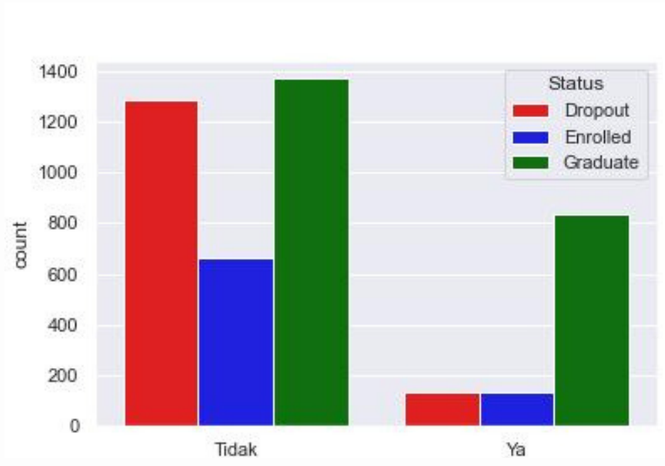
Student’s Status Distribution by

Gender



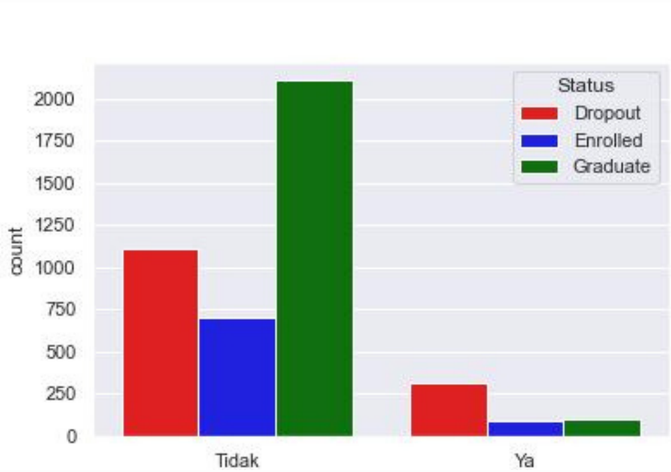
Men got more dropout student rather than women got

Scholarship Holder



Scholarship-holder students are more likely to graduate

Debtor

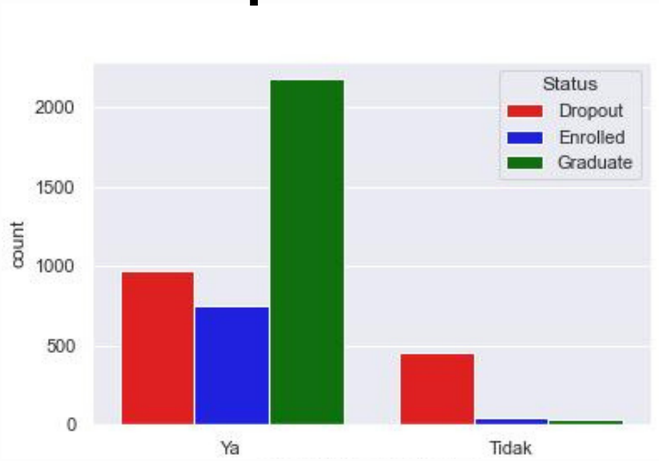


Debtor students tend to be dropped out students

Interesting Patterns

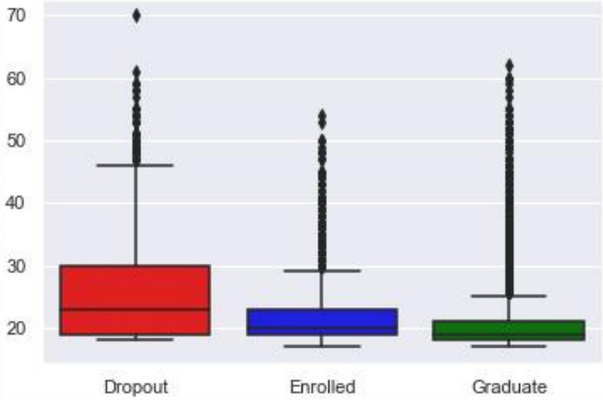
Student's Status Distribution by

Tuition Fees Up To Date



Students whose tuition fees are not up to date are more likely to dropout

Age at Enrollment



The distribution of dropped out students have wider range

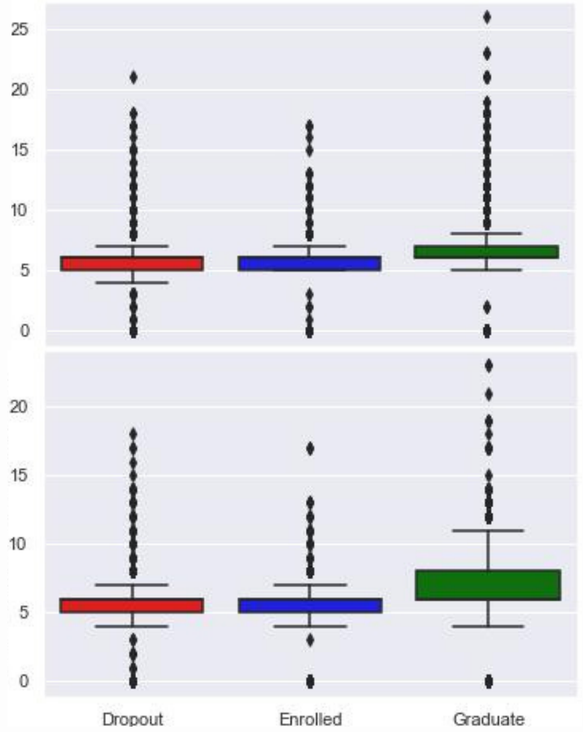
Interesting Patterns

Curricular units

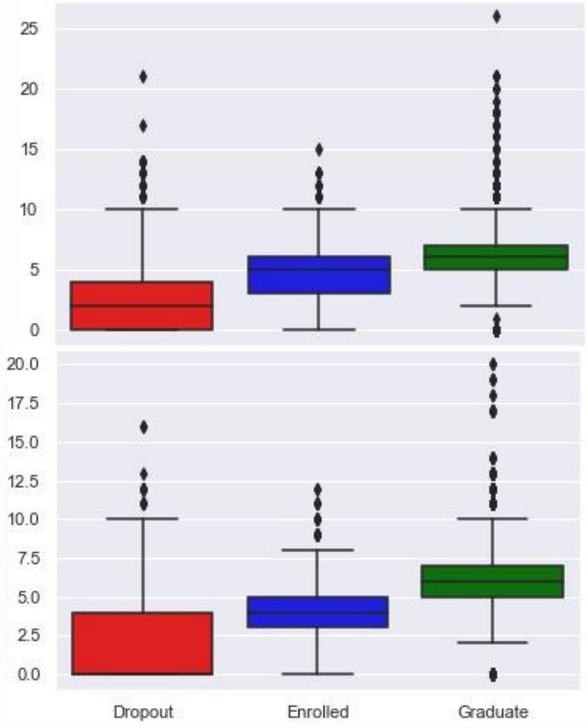
1st Sem

2nd Sem

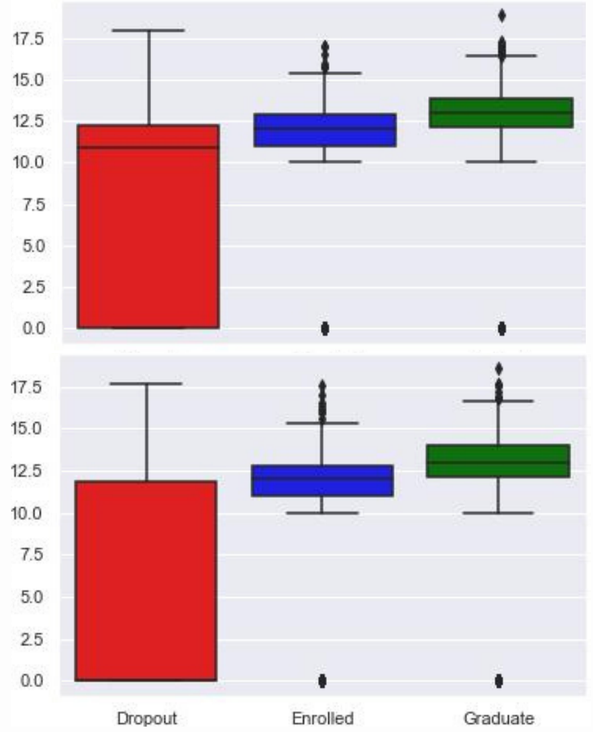
Enrolled



Approved



Grade

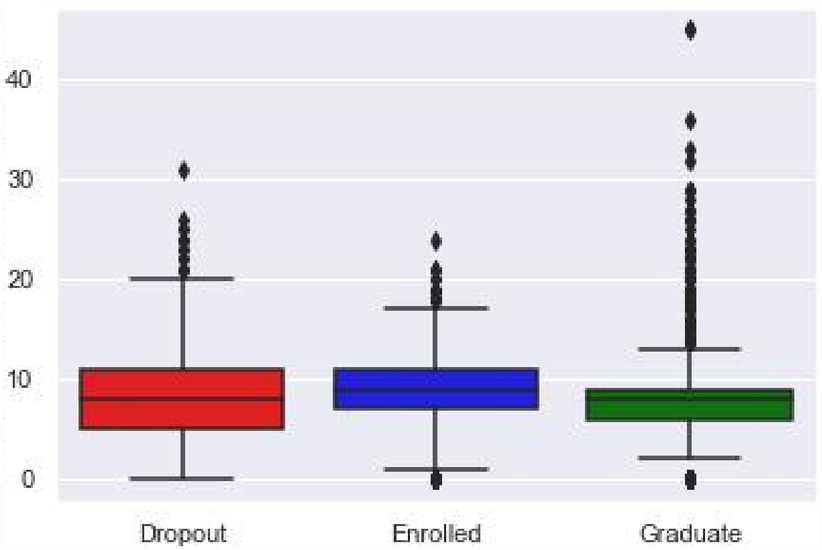


Similar patterns? Graduate students always have a higher median, either on enrolled, approved, or grade

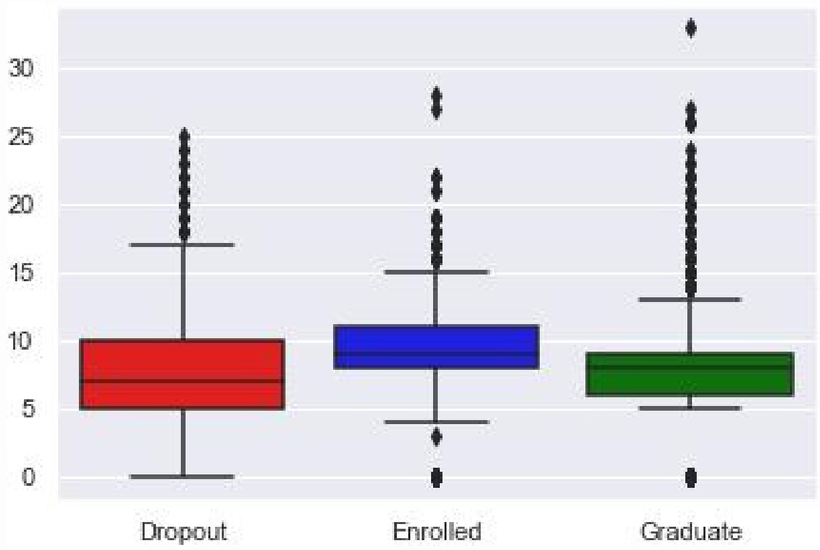
Interesting Patterns

Curricular units

1st Semester Evaluations

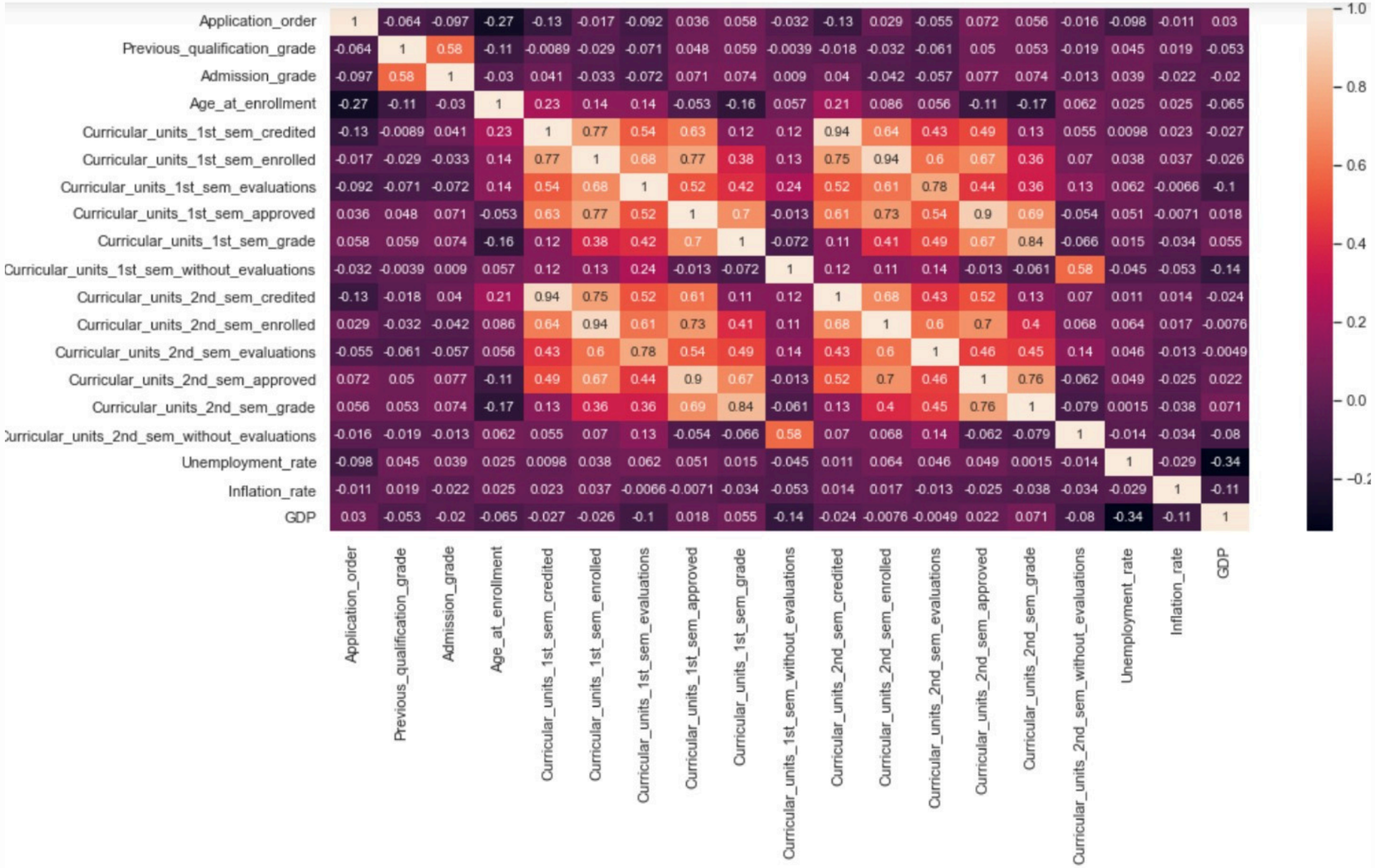


2nd Semester Evaluations



Dropout students always have wider spread and graduate students always have narrower spread.
Means that graduate students have more consistent in their evaluations rather than dropout students

Numeric Feature's Correlations



Some features are highly correlated. Those are features in the 1st semester correlated with its 2nd semester. This problem could cause multicollinearity and reduce the model's accuracy. Therefore, the Principle Component Analysis (PCA) technique will be used to preprocess these features.

Model Development

Data Preprocessing

Feature Selection:
All features in The Interesting Patterns

Numeric Features:
- PCA
- PowerTransformer

Categorical Features:
- One-hot Encoding

Data Splitting

X_train
X_test
y_train
y_test

Model Training

Use Randomized SearchCV to get the best params and train the model with this best params
- Decision-trees Classifier
- Random Forest Classifier
- Gradient Boosting

Evaluations

Choose a model with the highest Recall score on the “Dropout“ status

Evaluation Metrics

Precision

Precision is the ratio of correctly predicted positive observations to the total predicted positives. High precision indicates a low false positive rate. It answers the question, "Of all the instances predicted as positive, how many were actually positive?"

Recall

Recall is the ratio of correctly predicted positive observations to all the observations in the actual class. High recall indicates a low false negative rate. It answers the question, "Of all the instances that were actually positive, how many were correctly predicted as positive?"

F1-Score

The F1-score is the harmonic mean of precision and recall. The F1-score is a balance between precision and recall. It is useful when the class distribution is imbalanced. It answers the question, "How well does the model balance precision and recall?"

Accuracy

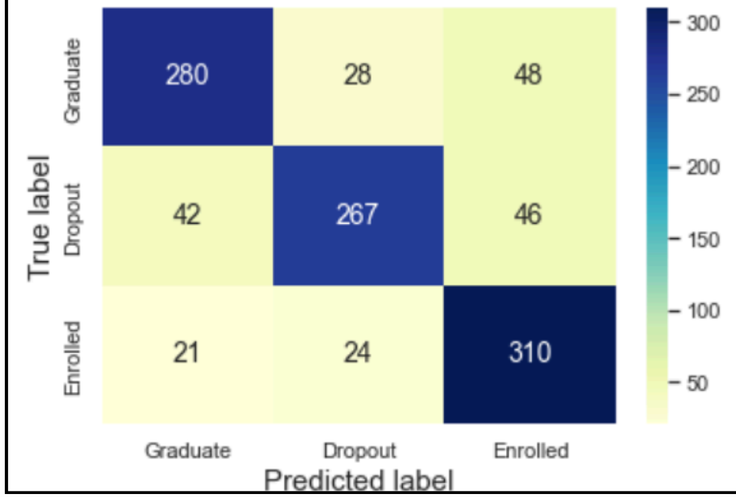
Accuracy is the ratio of correctly predicted observations to the total observations. It is the most intuitive performance measure and represents the overall effectiveness of the model. It answers the question, "How often is the model correct?"

Among these evaluation metrics, recall will be used to evaluate the model's performance. Why recall? We want a model that could accurately predict the "Dropout" status so that the institute could give guidance and direction for students classified as dropout

Evaluations

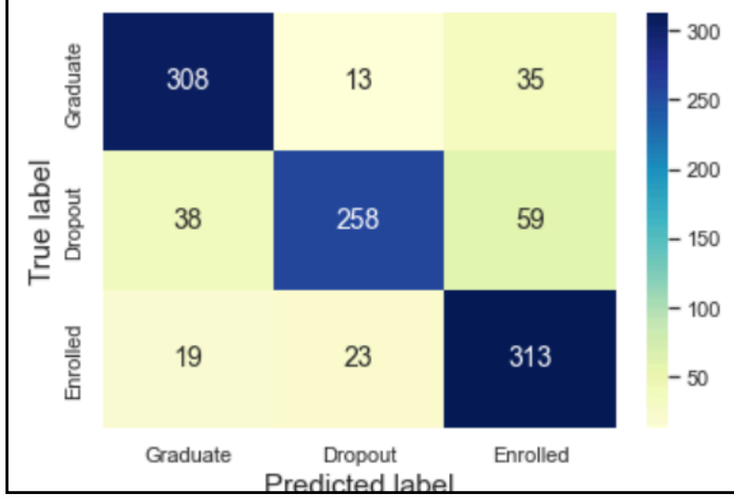
Decision-trees Classifier

	precision	recall	f1-score	support
Dropout	0.84	0.75	0.79	355
Enrolled	0.77	0.87	0.82	355
Graduate	0.82	0.79	0.80	356
accuracy			0.80	1066
macro avg	0.81	0.80	0.80	1066
weighted avg	0.81	0.80	0.80	1066



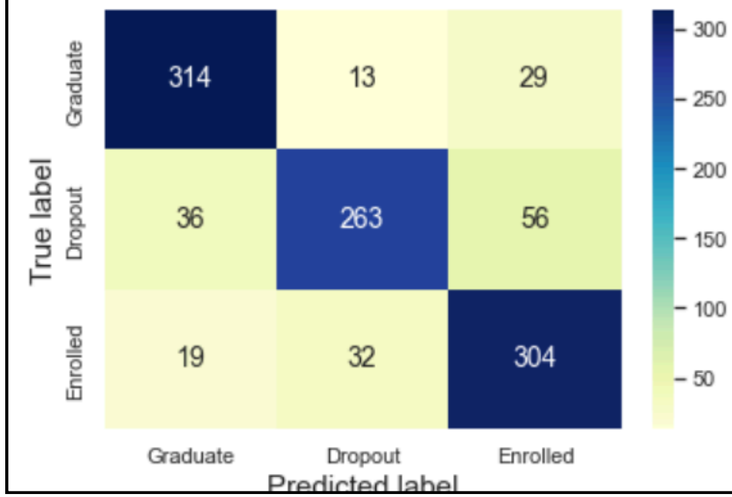
Random Forest Classifier

	precision	recall	f1-score	support
Dropout	0.88	0.73	0.80	355
Enrolled	0.77	0.88	0.82	355
Graduate	0.84	0.87	0.85	356
accuracy			0.82	1066
macro avg	0.83	0.82	0.82	1066
weighted avg	0.83	0.82	0.82	1066



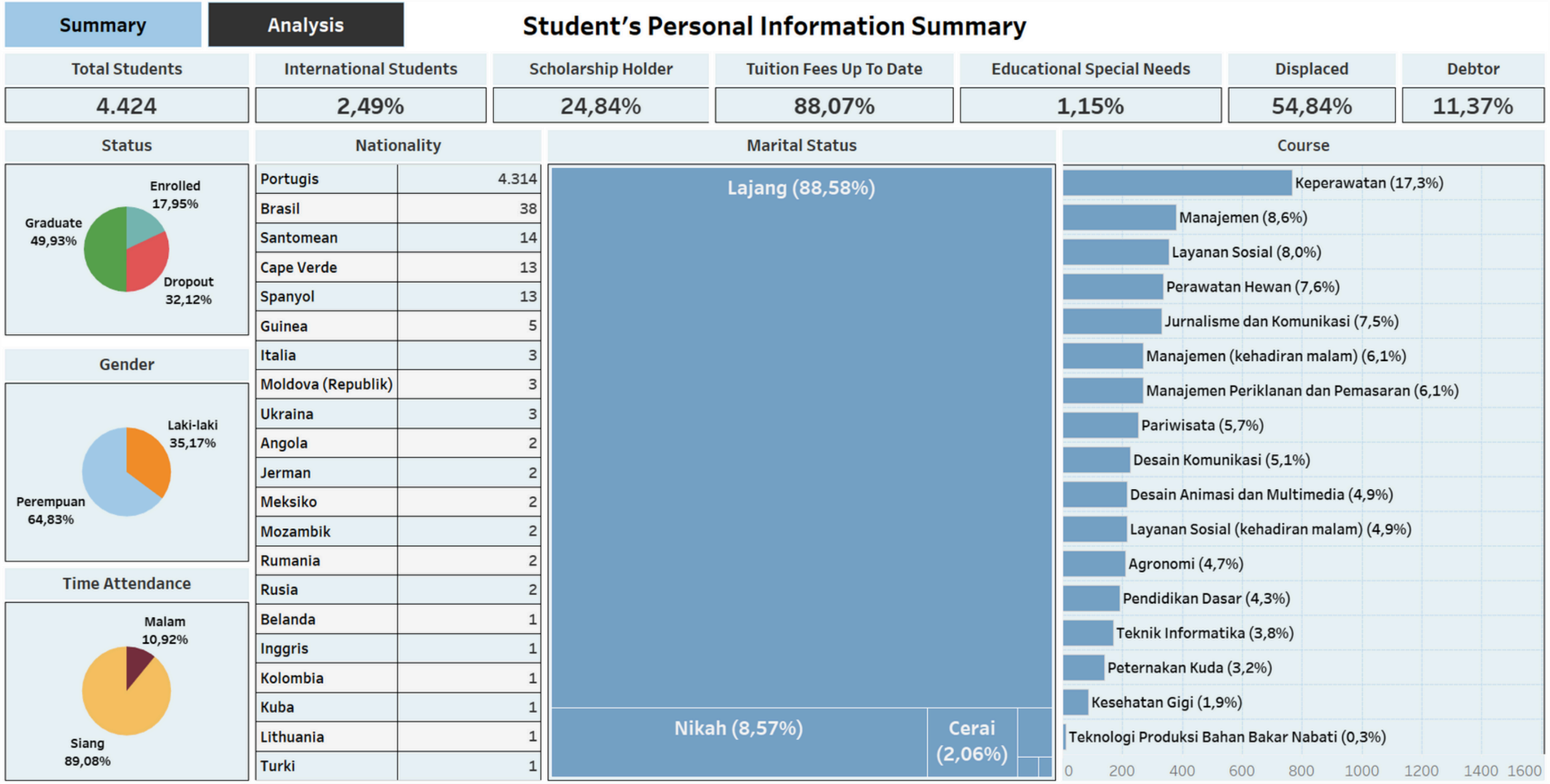
Gradient Boosting

	precision	recall	f1-score	support
Dropout	0.85	0.74	0.79	355
Enrolled	0.78	0.86	0.82	355
Graduate	0.85	0.88	0.87	356
accuracy			0.83	1066
macro avg	0.83	0.83	0.83	1066
weighted avg	0.83	0.83	0.83	1066



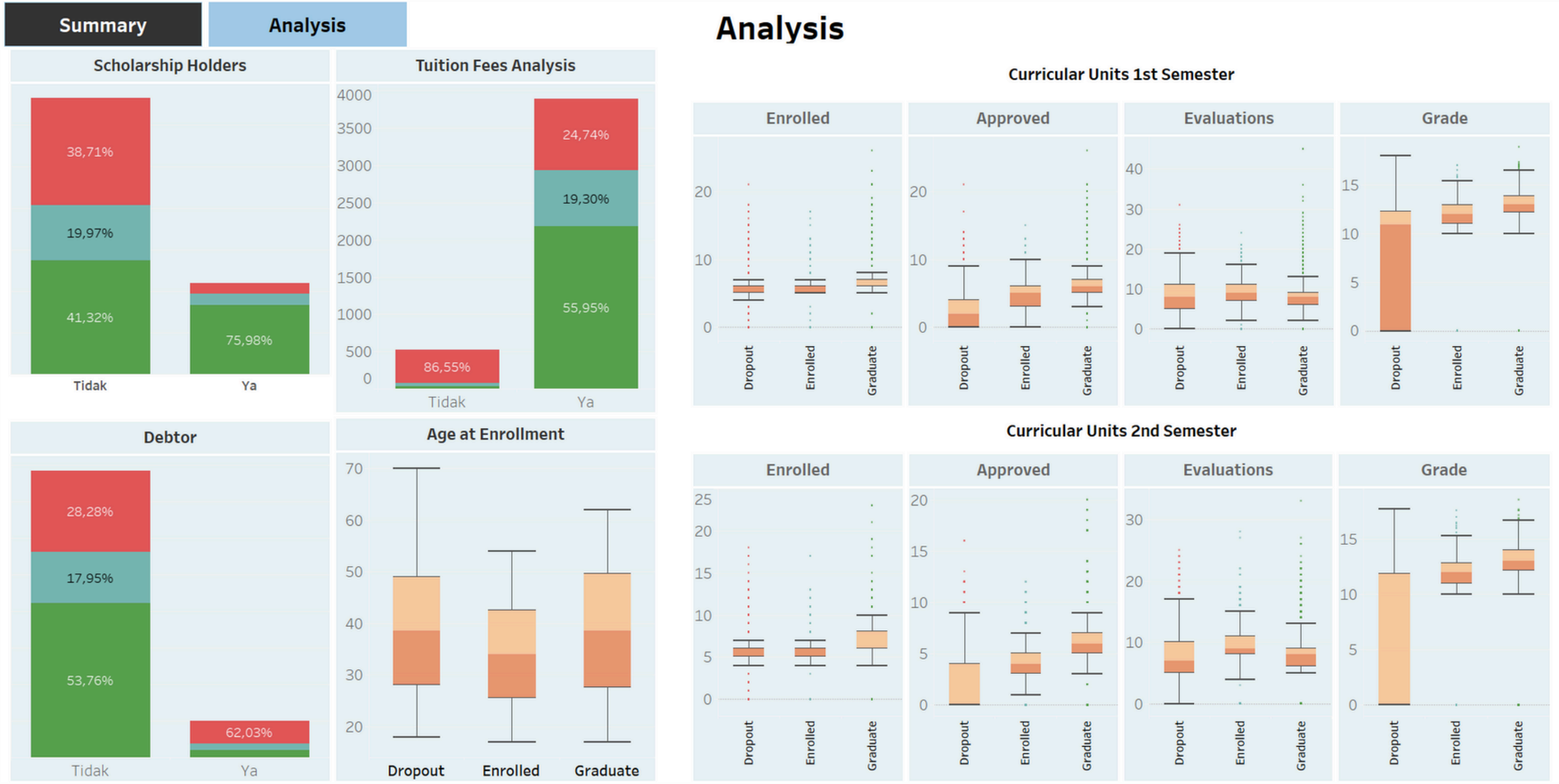
Based on the previous explanation, we need a model with the highest recall score on the “Dropout” status. From the pictures above, a model with the highest recall score on the “Dropout” status is **Decision-trees Classifier** model. This model will be deployed on Streamlit to predict the student’s status

Dashboard

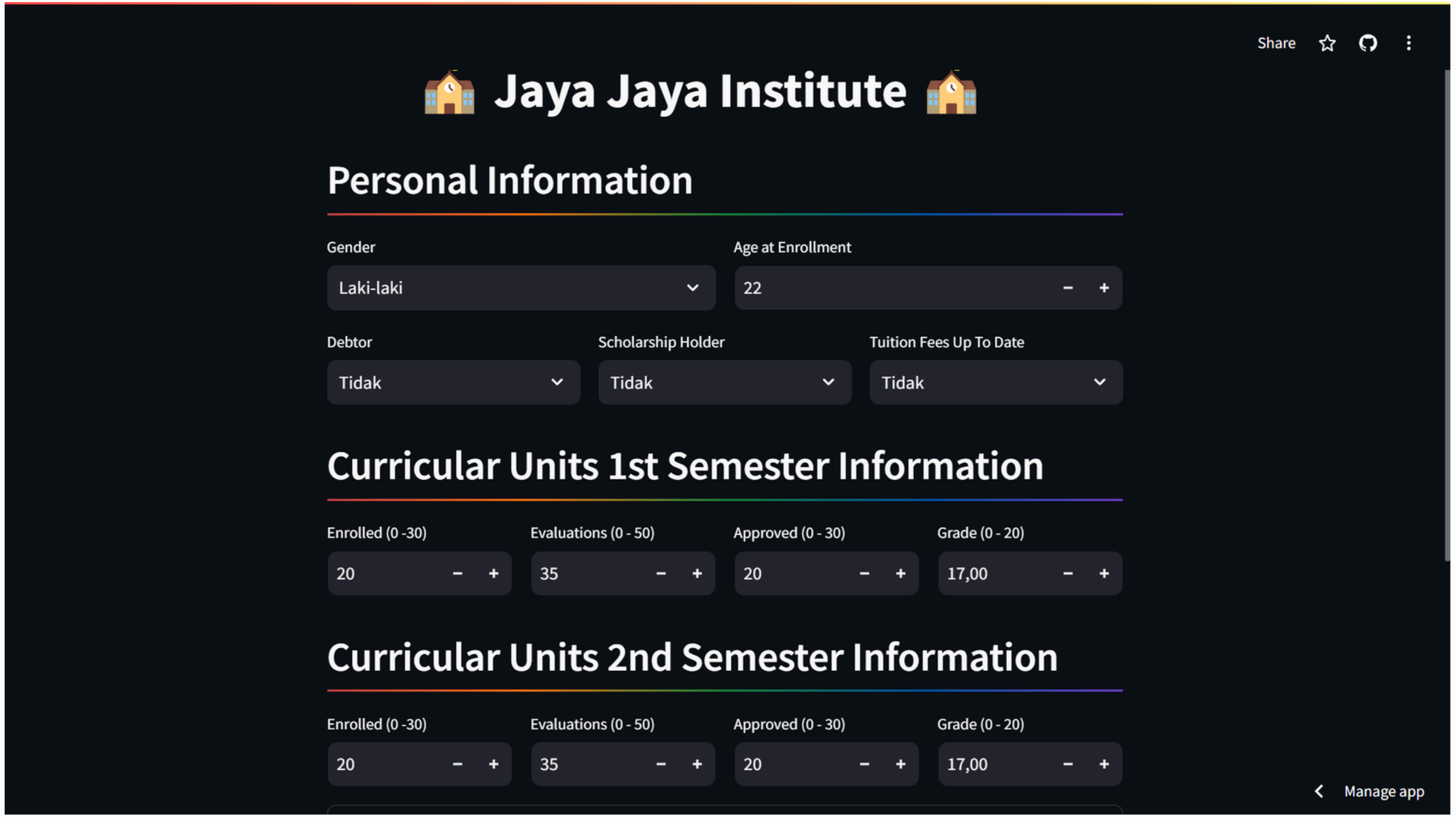


*This is an interactive dashboard. The information will automatically filtered if you clicked one or more of the charts (pie charts, table, treemaps, bar chart).

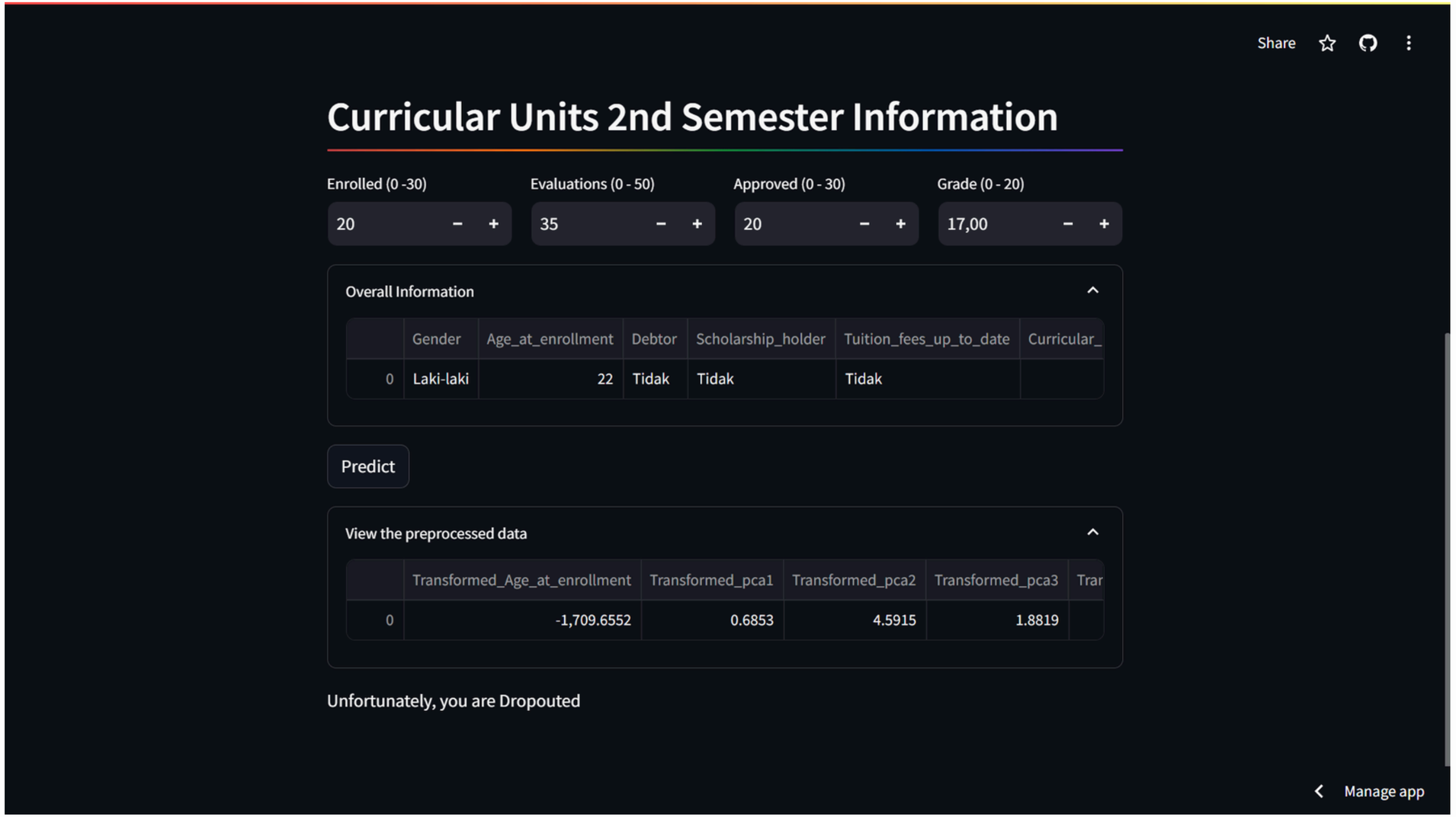
Dashboard

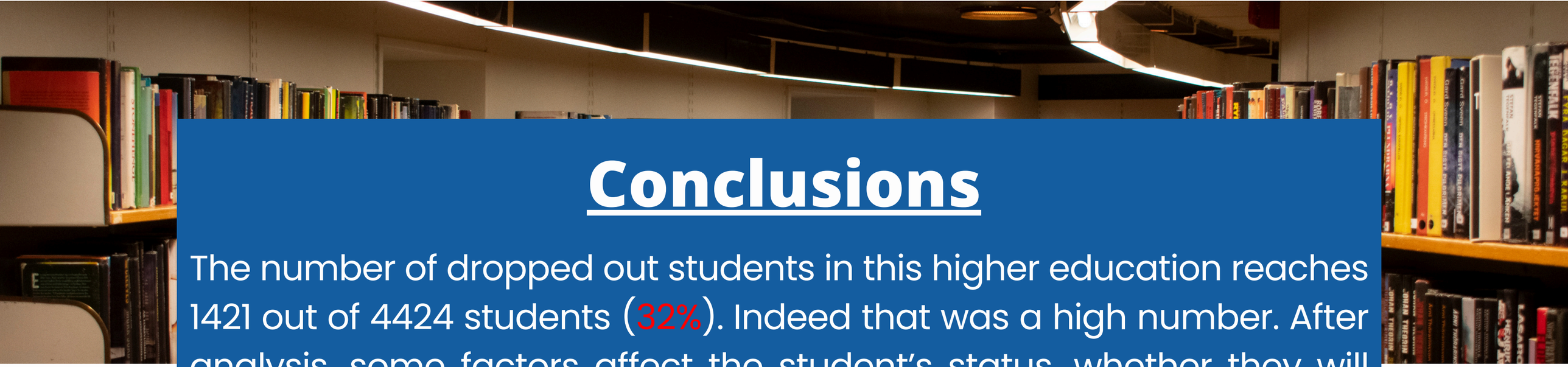

Summary
Analysis
Scholarship Holders
38,71%
19,97%
41,32%
75,98%
Tidak
Ya
Tuition Fees Analysis
4000
3500
3000
2500
2000
1500
1000
500
0
24,74%
19,30%
55,95%
86,55%
Tidak
Ya
Debtor
28,28%
17,95%
53,76%
62,03%
Tidak
Ya
Age at Enrollment
70
60
50
40
30
20
Dropout
Enrolled
Graduate
Analysis
Curricular Units 1st Semester
Enrolled
Approved
Evaluations
Grade
Dropout
Enrolled
Graduate
Curricular Units 2nd Semester
Enrolled
Approved
Evaluations
Grade
Dropout
Enrolled
Graduate

Model Prototype



Model Prototype





Conclusions

The number of dropped out students in this higher education reaches 1421 out of 4424 students (32%). Indeed that was a high number. After analysis, some factors affect the student's status, whether they will dropout or not. The strongest factor that affects the student will dropout is tuition fees up to date feature. Students who dropped out because the tuition fees un-updated reaches 86%. Next, there are some interesting patterns obtained by students in the 1st and 2nd semester. Thee graduated students tend to have higher grade rather than enrolled and dropped out students.

Recommendation Action Items

Keep tuition fees updated to the recent info

An up-to-date tuition fee will ensure that students do not experience unexpected financial problems, which may cause them to drop out. A possible action is to conduct regular audits every semester to ensure that all tuition fees data has been updated. The finance team can send automated reminders to students and parents about tuition fees updates.

Pay more attention to students who get low grades (below 13)

Actions that can be given to students with these criteria are such as providing academic guidance, encouraging students to consult with supervisors regarding problems or obstacles faced by students when studying, and providing remedial classes for students with low scores.

THANK YOU!

Norman Dwi Febrio

Prospective Data Scientist

linkediin.com/in/norman-dwi-febrio

github.com/NormanFebrio