




LA06 - Computer Science & Statistics

# PREDICT STUDENTS' ACADEMIC

Kelompok 6 Introduction to Data Science

# NAMA ANGGOTA

| | |
|---|---|
| Jovan Melsyah | 2802413202 |
| Melvern Gerrad | 2802417882 |
| Jonathan | 2802443621 |
| Kenneth Christian Yapri | 2802527192 |

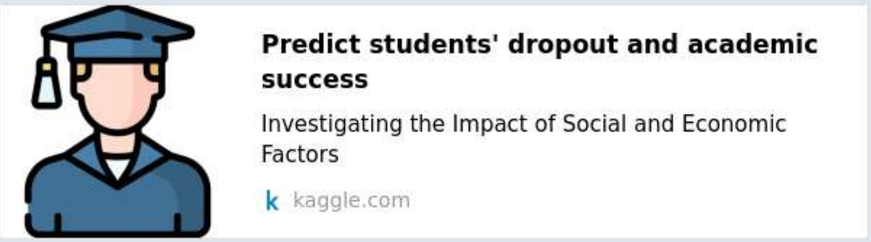
# DAFTAR ISI


# I. PENDAHULUAN

# LATAR BELAKANG

Pada penelitian yang kami lakukan, kami mendapat informasi bahwa retensi mahasiswa dalam pendidikan tinggi adalah salah satu tantangan utama yang dihadapi oleh institusi pendidikan di seluruh dunia. Mahasiswa yang tidak menyelesaikan pendidikan mereka sering kali menghadapi dampak negatif pada perkembangan karier, stabilitas finansial, dan kesejahteraan pribadi mereka. Dari sisi institusi Pendidikan menyatakan bahwa tingkat putus studi yang tinggi dapat merusak reputasi, mengurangi pendapatan, dan menimbulkan biaya tambahan untuk merekrut mahasiswa baru. Oleh karena itu, penting untuk memahami faktor-faktor yang memengaruhi retensi mahasiswa agar dapat merancang strategi yang efektif dalam mendukung keberhasilan akademik mereka. Melalui Faktor-faktor ini, kami akan memprediksi apakah mahasiswa tersebut akan Graduate, Enrolled,atau Dropout.

# PENJELASAN SINGKAT DATASET




https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention/data

## Predict students' dropout and academic success

Dataset “Predict Student’s Dropout and Academic Success” ini merupakan sebuah dataset yang berisi menenai status retensi mahasiswa beserta faktor faktor yang mempengaruhinya. Dalam Dataset ini lebih ditekankan pada faktor sosial dan ekonomi dari seorang mahasiswa tersebut.

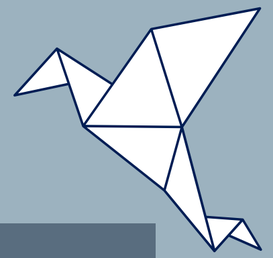
Dataset ini dibuat oleh Kaggel User dengan nickname “The Devastator” dan diupdate terkahir pada 2 tahun yang lalu, dan memiliki usability 10.0

Dataset ini terdiri dari 35 kolom dan 4424 baris. Dimana 23 kolom dari 35 kolom itu kami gunakan sebagai visualisasi, lalu melihat keterkaitannya, 12 dari 23 kolom tersebut kami gunakan untuk modeling.

# ALASAN MENGAMBIL DATASET

Kami memilih dataset ini karena data yang ada sangat relevan untuk memahami masalah mahasiswa yang drop out. Dataset ini mencakup informasi penting seperti nilai akademik, kondisi keuangan, dan latar belakang mahasiswa, Semuanya bisa membantu untuk melihat faktor-faktor yang memengaruhi mereka untuk tetap belajar atau berhenti. Dengan data ini, Kami bisa membuat sebuah prediksi kelulusan dan drop out sehingga nantinya Perguruan tinggi bisa mengambil langkah lebih cepat untuk membantu mahasiswa yang kesulitan. Dataset ini cocok untuk belajar karena lengkap dan mengandung data yang banyak.

# TUJUAN PENGGUNAAN DATASET



Dengan menggunakan Dataset ini, kami ingin menganalisa dan membuat model machine learning serta memprediksi apakah mahasiswa tersebut akan Graduate, Enrolled, atau Dropout dengan faktor faktor yang ada

## I. Descriptive & Diagnostic

- Mengidentifikasi dan mengvisualisasikan pendistribusian target (Graduate, Enrolled, Dropout)
- Mengidentifikasi dan mengvisualisasikan pendistribusian faktor yang mempengaruhi target (Graduate, Enrolled, Dropout)
- Mengvisualisasikan korelasi atau hubungan faktor dengan target

# TUJUAN PENGGUNAAN DATASET

## II. Predictive

- Membuat suatu model machine learning untuk memprediksi mahasiswa tersebut melalui dataset ini


# II. DATA PROCESSING

# PENGAMBILAN DATA

**Dari 35 Kolom yang ada pada dataset, kami mengambil sebanyak 23 kolom yang terdiri dari:**

- Marital Status
- Application Mode
- Application order
- Course
- Daytime/evening attendance
- Previous qualification
- Nacionality
- Mother's qualification
- Father's qualification
- Mother's occupation
- Father's occupation
- Displaced
- Educational special needs
- Debtor
- Tuition fees up to date
- Gender
- Scholarship holder
- Age at enrollment
- International
- Unemployment rate

# DESKRIPSI DATA

**BAGIAN 1**

- **Marital status:** Status perkawinan mahasiswa. (Kategorikal)
- **Application mode**: Metode aplikasi yang digunakan oleh mahasiswa. (Kategorikal)
- **Application order**: Urutan aplikasi yang diajukan oleh mahasiswa. (Numerikal)
- **Course**: Program studi yang diambil oleh mahasiswa. (Kategorikal)
- **Daytime/evening attendance**: Apakah mahasiswa menghadiri kelas pada siang hari atau malam hari. (Kategorikal)
- **Previous qualification**: Kualifikasi yang dimiliki mahasiswa sebelum mendaftar ke pendidikan tinggi. (Kategorikal)
- **Nationality**: Kewarganegaraan mahasiswa. (Kategorikal)
- **Mother's qualification**: Kualifikasi pendidikan ibu dari mahasiswa. (Kategorikal)
- **Father's qualification**: Kualifikasi pendidikan ayah dari mahasiswa. (Kategorikal)
- **Mother's occupation**: Pekerjaan ibu dari mahasiswa. (Kategorikal)
- **Father's occupation**: Pekerjaan ayah dari mahasiswa. (Kategorikal)
- **Displaced**: Apakah mahasiswa termasuk orang yang terlantar. (Kategorikal)

# DESKRIPSI DATA

## BAGIAN 2

- **Educational special needs:** Apakah mahasiswa memiliki kebutuhan pendidikan khusus. (Kategorikal)
- **Debtor**: Apakah mahasiswa memiliki hutang. (Kategorikal)
- **Tuition fees up to date**: Apakah pembayaran biaya kuliah mahasiswa sudah diperbarui. (Kategorikal)
- **Gender:** Jenis kelamin mahasiswa. (Kategorikal)
- **Scholarship holder:** Apakah mahasiswa menerima beasiswa. (Kategorikal)
- **Age at enrollment**: Usia mahasiswa saat pendaftaran. (Numerikal)
- **International**: Apakah mahasiswa termasuk mahasiswa internasional. (Kategorikal)
- **Unemployment Rate**: Tingkat pengangguran di wilayah asal mahasiswa pada waktu tertentu. (Numerikal)
- **Inflation Rate**: Tingkat inflasi di wilayah asal mahasiswa pada waktu tertentu. (Numerikal)
- **GDP**: Produk Domestik Bruto dari wilayah asal mahasiswa pada waktu tertentu. (Numerikal)
- **Target: Status** hasil akademik mahasiswa, yaitu apakah mahasiswa Graduate (lulus), Enrolled (masih terdaftar), atau Dropout (berhenti). (Kategorikal)

# DATA INFORMATION

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | Marital status | 4424 non-null | int64 |
| 1 | Application mode | 4424 non-null | int64 |
| 2 | Application order | 4424 non-null | int64 |
| 3 | Course | 4424 non-null | int64 |
| 4 | Daytime/evening attendance | 4424 non-null | int64 |
| 5 | Previous qualification | 4424 non-null | int64 |
| 6 | Nacionality | 4424 non-null | int64 |
| 7 | Mother's qualification | 4424 non-null | int64 |
| 8 | Father's qualification | 4424 non-null | int64 |
| 9 | Mother's occupation | 4424 non-null | int64 |
| 10 | Father's occupation | 4424 non-null | int64 |
| 11 | Displaced | 4424 non-null | int64 |
| 12 | Educational special needs | 4424 non-null | int64 |
| 13 | Debtor | 4424 non-null | int64 |
| 14 | Tuition fees up to date | 4424 non-null | int64 |
| 15 | Gender | 4424 non-null | int64 |
| 16 | Scholarship holder | 4424 non-null | int64 |
| 17 | Age at enrollment | 4424 non-null | int64 |
| 18 | International | 4424 non-null | int64 |
| 19 | Unemployment rate | 4424 non-null | float64 |
| ... | | | |
| 21 | GDP | 4424 non-null | float64 |
| 22 | Target | 4424 non-null | object |

**Dataset ini sudah bersih dan lengkap dikarenakan tidak adanya kolom yang barisnya null sehingga process data cleansing tidak diperlukan**

# DECODING DATA

## Before

| Marital status | Application mode | Application order | Course | Daytime/evening attendance | Previous qualification | Nacionality | Mother's qualification | Father's qualification | Mother's occupation | ... | Debtor | Tuition fees up to date | Gender | Scholarship holder | Age at enrollment | Internation |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Single | 8 | 5 | 2 | 1 | 1 | 1 | 13 | 10 | 6 | ... | 0 | 1 | 1 | 0 | 20 | |
| Single | 6 | 1 | 11 | 1 | 1 | 1 | 1 | 3 | 4 | ... | 0 | 0 | 1 | 0 | 19 | |
| Single | 1 | 5 | 5 | 1 | 1 | 1 | 22 | 27 | 10 | ... | 0 | 0 | 1 | 0 | 19 | |
| Single | 8 | 2 | 15 | 1 | 1 | 1 | 23 | 27 | 6 | ... | 0 | 1 | 0 | 0 | 20 | |
| Married | 12 | 1 | 3 | 0 | 1 | 1 | 22 | 28 | 10 | ... | 0 | 1 | 0 | 0 | 45 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| Single | 1 | 6 | 15 | 1 | 1 | 1 | 1 | 1 | 6 | ... | 0 | 1 | 1 | 0 | 19 | |
| Single | 1 | 2 | 15 | 1 | 1 | 19 | 1 | 1 | 10 | ... | 1 | 0 | 0 | 0 | 18 | |
| Single | 1 | 1 | 12 | 1 | 1 | 1 | 22 | 27 | 10 | ... | 0 | 1 | 0 | 1 | 30 | |
| Single | 1 | 1 | 9 | 1 | 1 | 1 | 22 | 27 | 8 | ... | 0 | 1 | 0 | 1 | 20 | |
| Single | 5 | 1 | 15 | 1 | 1 | 9 | 23 | 27 | 6 | ... | 0 | 1 | 0 | 0 | 22 | |

ows × 23 columns

**Seperti yang Terlihat data yang seharusnya categorical ditulis dalam bentuk angka sehingga harus didecode terlebih dahulu.**

**data decode diambil dari:**

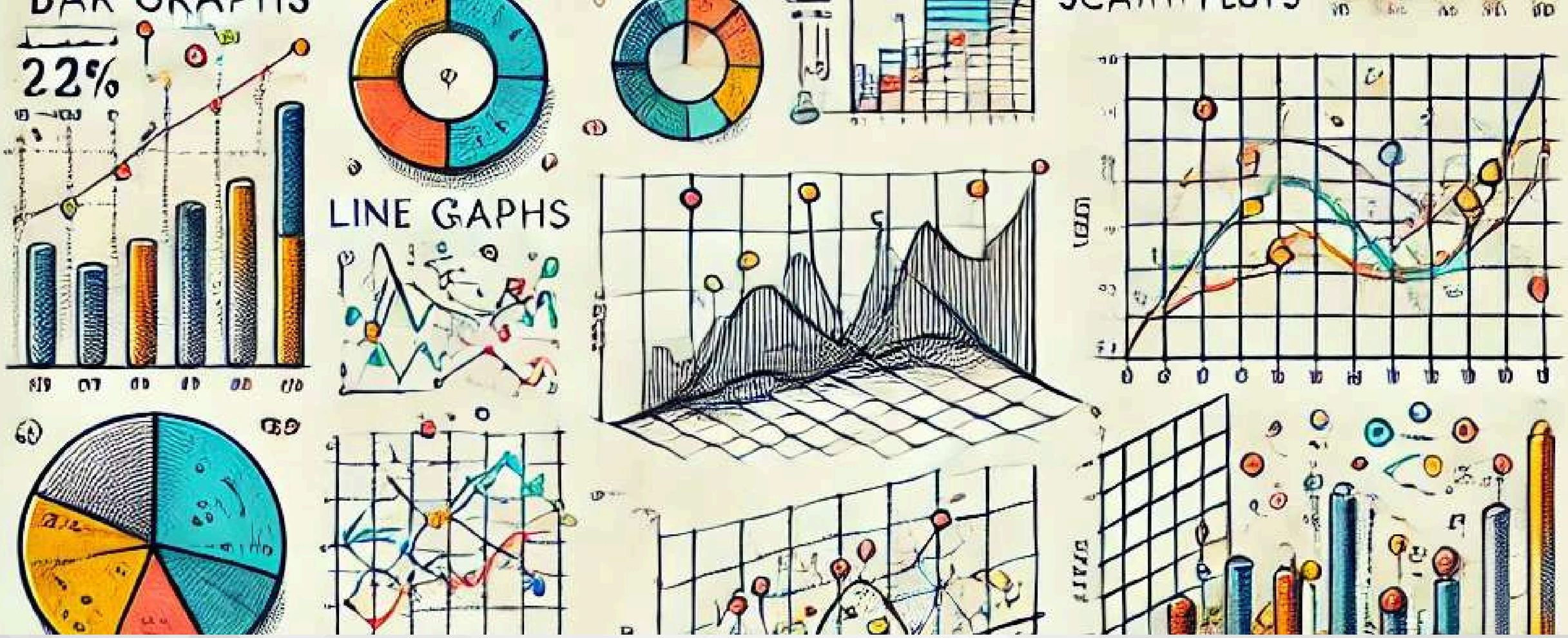
**https://github.com/carmelh/SQL_projects/blob/main/student_data_analysis/Datasets/tA10_yes_no.csv**

# DECODING DATA

## After

| Marital status | Application mode | Application order | Course | Daytime/evening attendance | Previous qualification | Nacionality | Mother's qualification | Father's qualification | Mother's occupation | ... | Debtor | Tuition fees up to date | Gender | Scholarship holder | Age at enrollment | International | Unemplo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Single | 2nd phase—general contingent | 5 | Animation and Multimedia Design | Daytime | Secondary education | Portuguese | General commerce course | Other—11th Year of Schooling | Personal Services, Security and Safety Workers... | ... | No | Yes | Male | No | 20 | No | |
| Single | International student (bachelor) | 1 | Tourism | Daytime | Secondary education | Portuguese | Secondary Education—12th Year of Schooling or ... | Higher Education—degree | Intermediate Level Technicians and Professions | ... | No | No | Male | No | 19 | No | |
| Single | 1st phase—general contingent | 5 | Communication Design | Daytime | Secondary education | Portuguese | General Course of Administration and Commerce | Basic education 1st cycle (4th/5th year) or eq... | Unskilled Workers | ... | No | No | Male | No | 19 | No | |
| Single | 2nd phase—general contingent | 2 | Journalism and Communication | Daytime | Secondary education | Portuguese | Supplementary Accounting and Administration | Basic education 1st cycle (4th/5th year) or eq... | Personal Services, Security and Safety Workers... | ... | No | Yes | Female | No | 20 | No | |
| Married | Over 23 years old | 1 | Social Service (evening attendance) | Evening | Secondary education | Portuguese | General Course of Administration and Commerce | Basic Education 2nd Cycle (6th/7th/8th Year) o... | Unskilled Workers | ... | No | Yes | Female | No | 45 | No | |

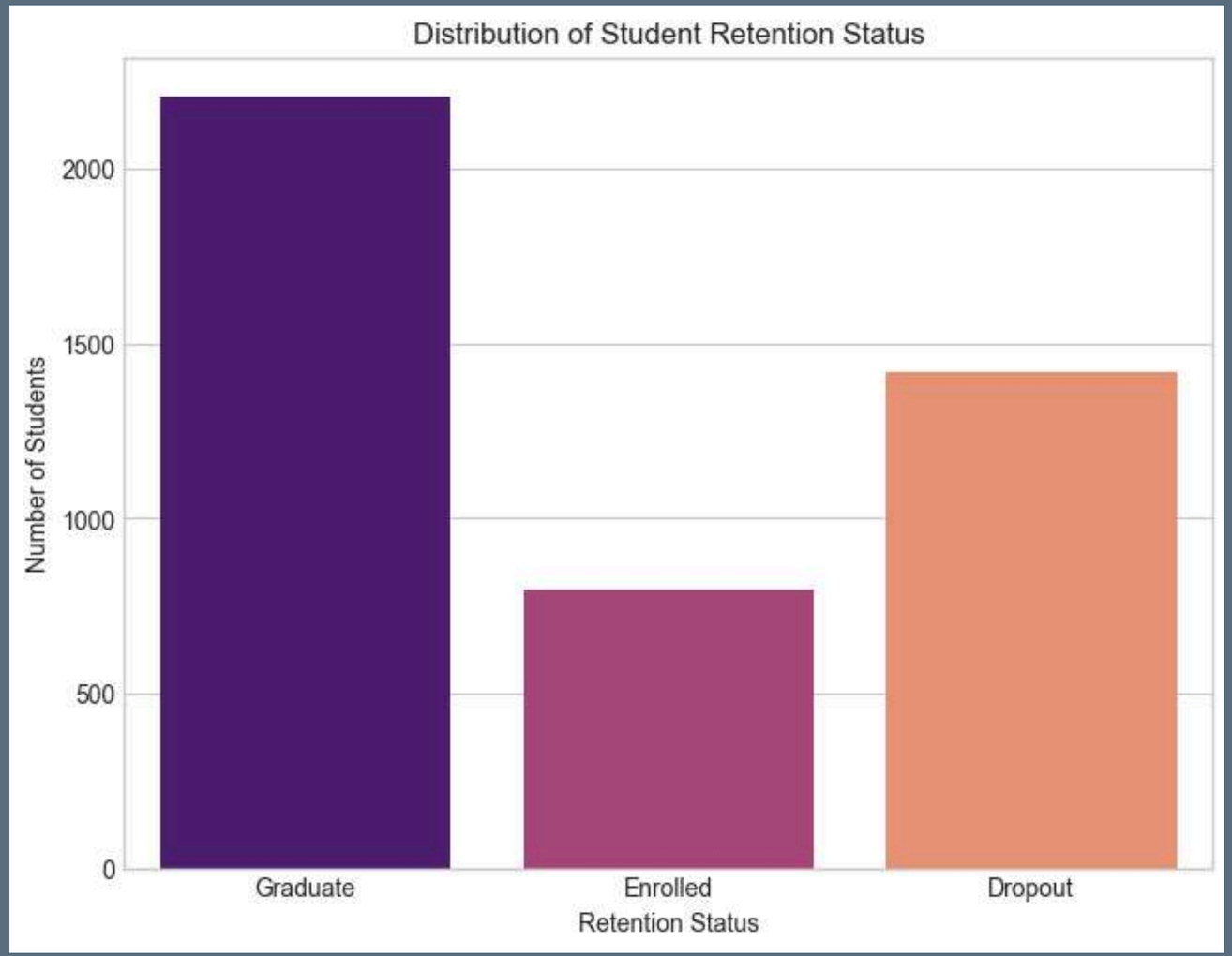
Data Setelah Didecode




# III. DATA VISUALIZATION

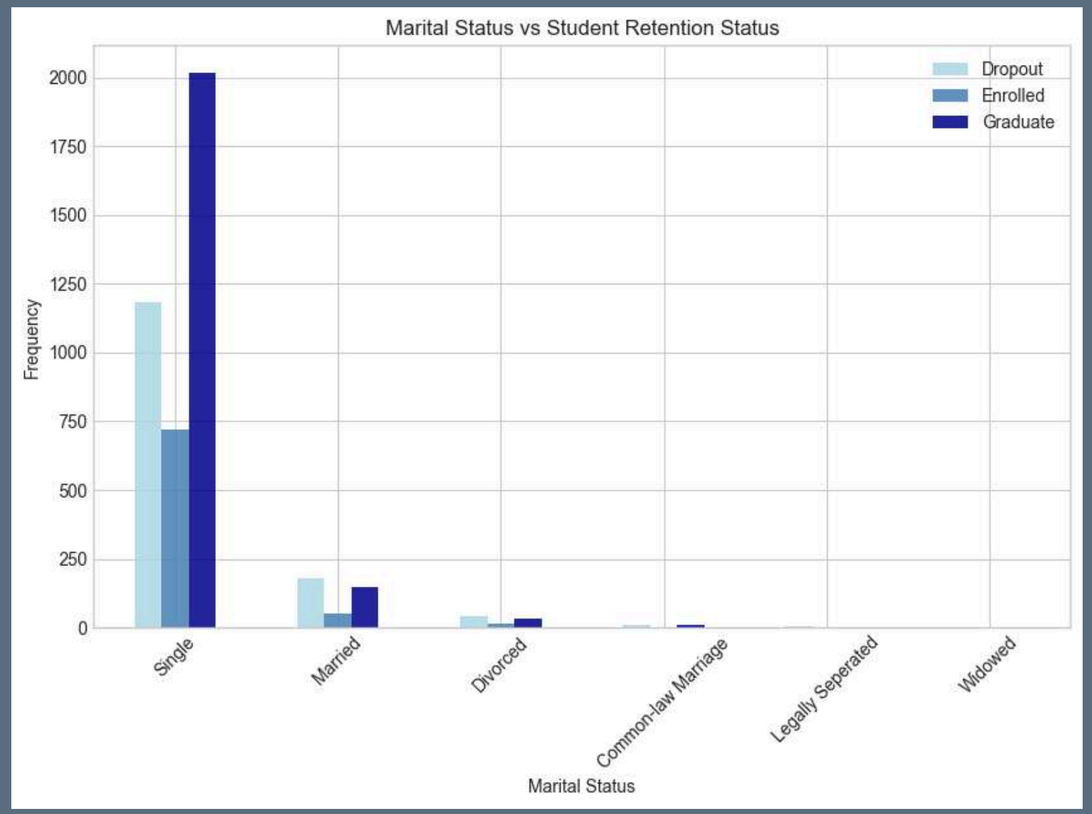
## (DESCRIPTIVE & DIAGNOSTIC)

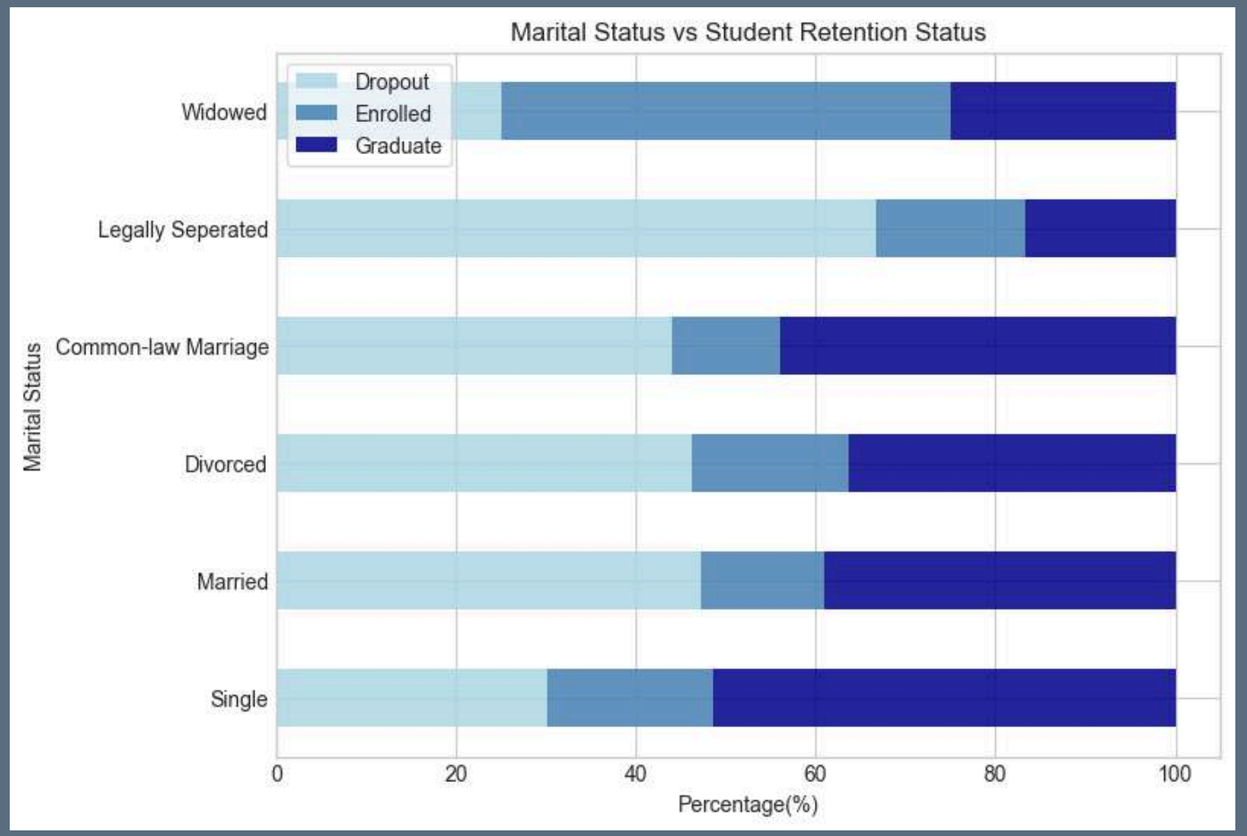
# DISTRIBUTION OF STUDENT RETENTION STATUS




Ini adalah grafik batang yang menggambarkan status retention murid-murid dengan 3 kategori: "Graduate", "Enrolled" , dan "Dropout". Kategori "Graduate" menunjukkan jumlah murid yang berhasil menyelesaikan program studi mereka dan lulus, sedangkan kategori "Enrolled" menggambarkan jumlah murid yang masih terdaftar dan aktif mengikuti program studi. Sementara itu, kategori "Dropout" menampilkan jumlah murid yang keluar dari program atau tidak melanjutkan studi mereka. Tinggi masing-masing batang pada grafik ini mencerminkan distribusi status retention murid-murid, memberikan gambaran yang jelas tentang tingkat kelulusan, keterlibatan aktif, serta angka putus sekolah dalam institusi pendidikan tersebut. Grafik ini bisa dipakai untuk untuk mengevaluasi efektivitas program pendidikan dan membantu dalam perencanaan kebijakan terkait peningkatan retention dan pengurangan angka putus sekolah.

# MARITAL STATUS VS STUDENT RETENTION STATUS

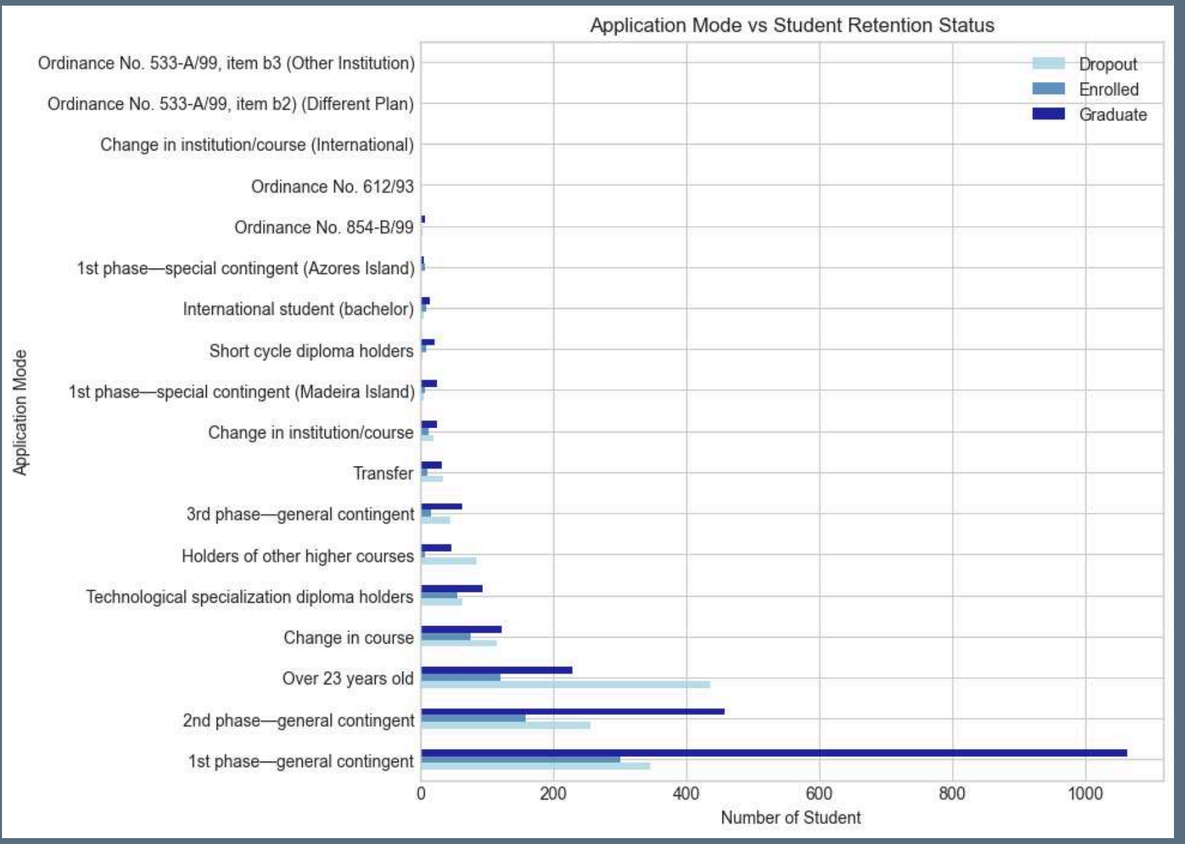


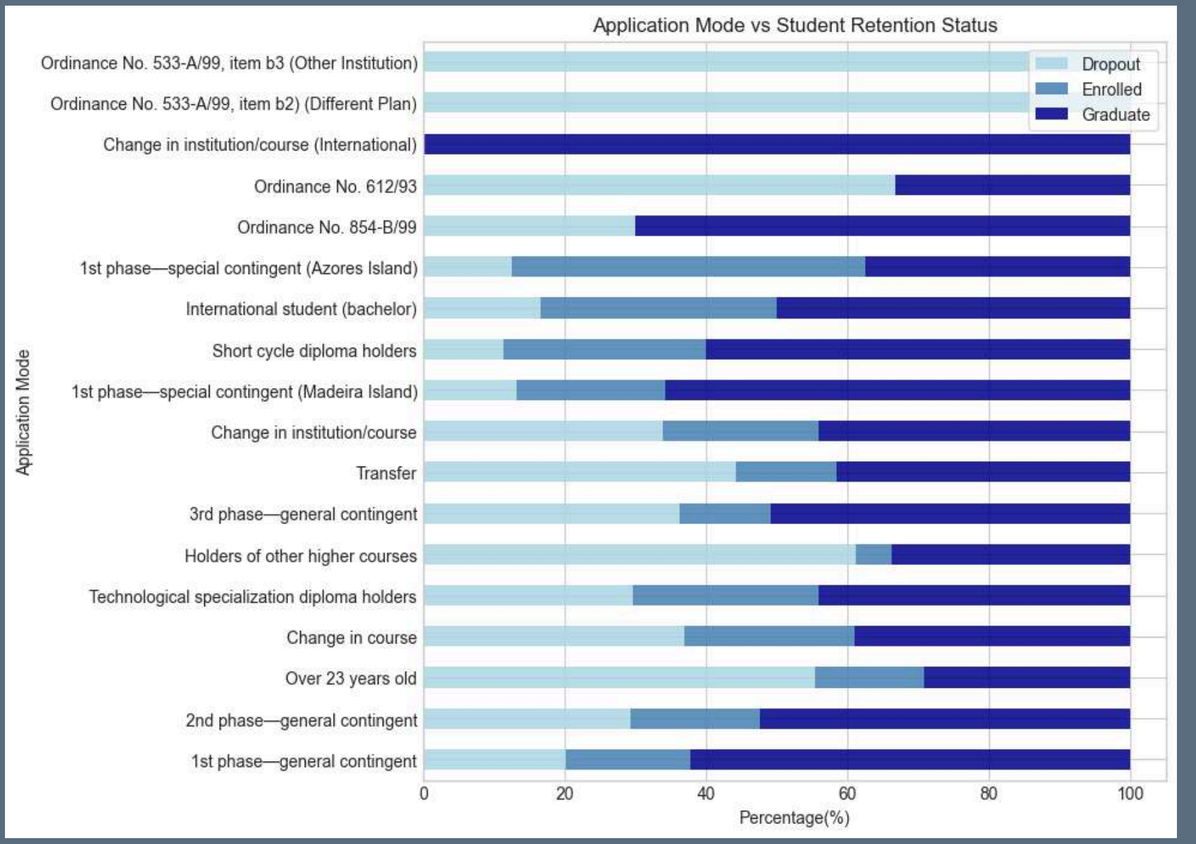




Grafik ini menunjukkan frequency marital status murid dan juga hubunganya dengan student retention status. Bisa dilihat dari grafik ini bahwa lebih banyak murid yang masih single, ini mungkin bisa karena murid-murid yang masih single bisa focus dengan mendapatkan edukasi

Grafik ini menggambarkan hubungan antara status pernikahan murid dan status retensi murid. Dari grafik ini, terlihat bahwa sebagian besar murid yang berstatus single berhasil lulus, sementara murid yang berstatus pisah rumah cenderung lebih banyak yang drop out. Di sisi lain, murid yang berstatus janda lebih banyak yang tetap terdaftar atau enrolled dalam program studi mereka.

# APPLICATION MODE VS STUDENT RETENTION STATUS

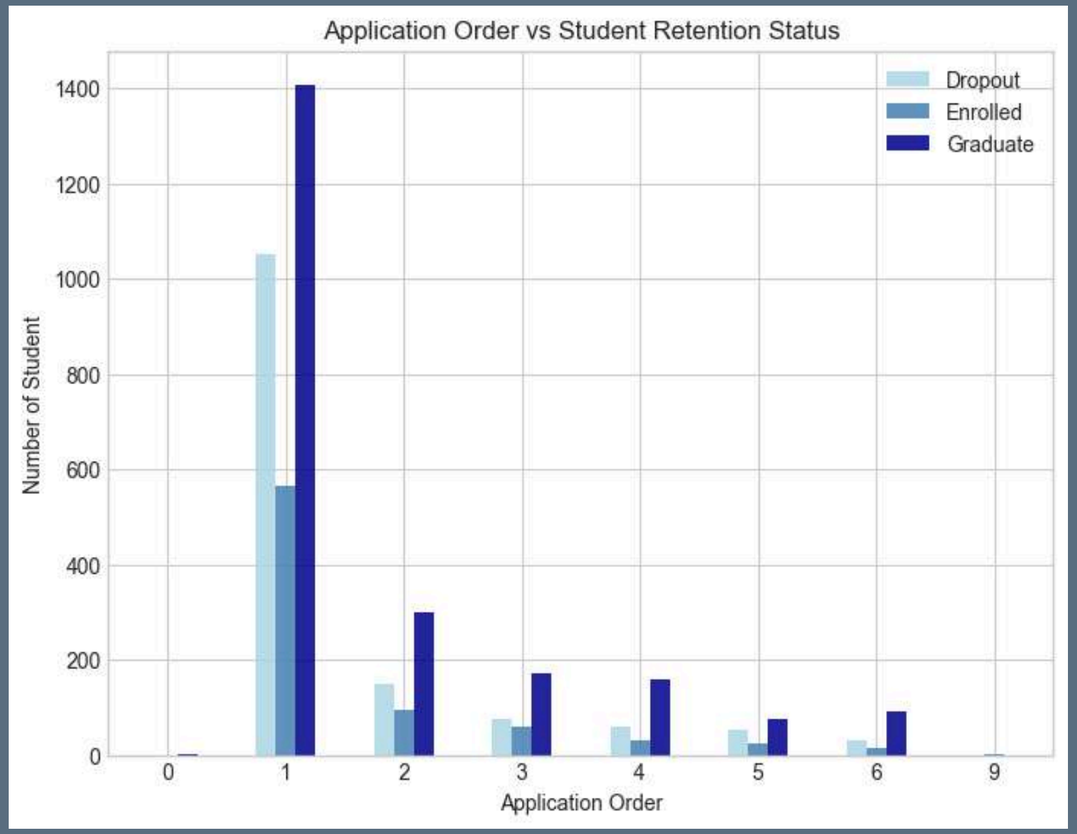


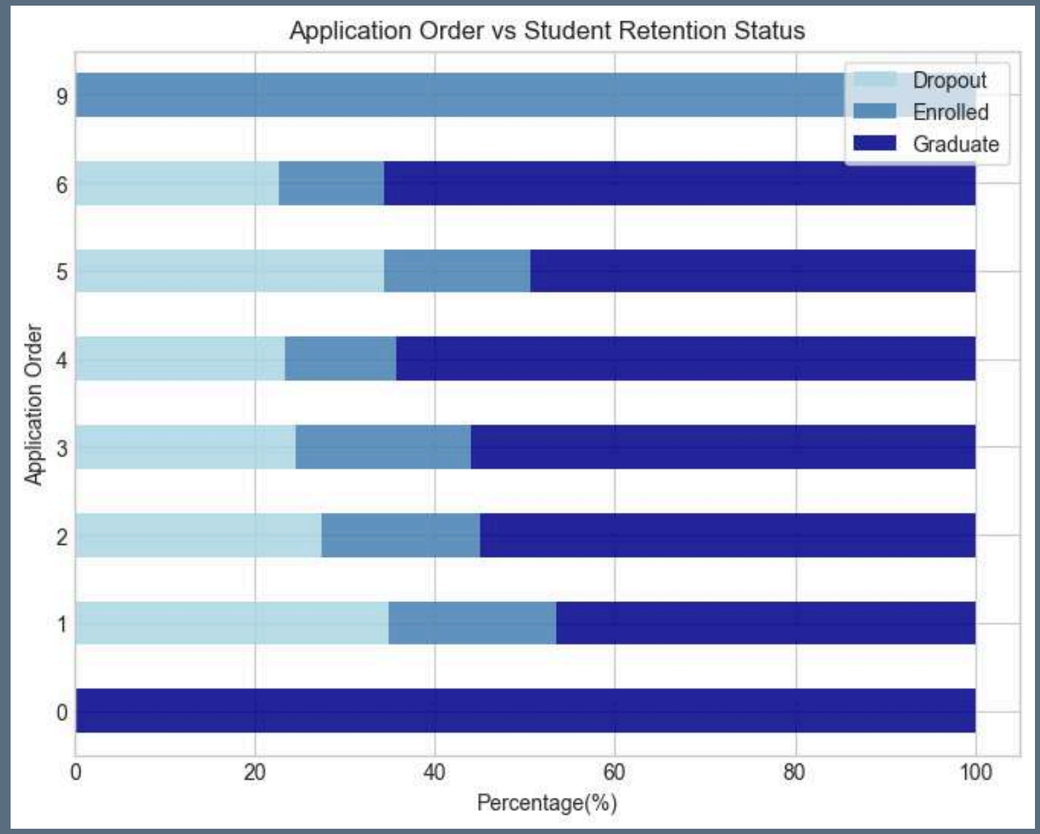




Grafik ini menggambarkan frekuensi mode aplikasi dan hubungannya dengan status retensi murid. Dari grafik ini, dapat dilihat bahwa sebagian besar murid memilih mode aplikasi 1st Phase - General Contingent. Grafik ini juga memperlihatkan bagaimana pemilihan mode aplikasi ini berhubungan dengan status retensi murid.

Grafik ini menggambarkan hubungan antara mode aplikasi dan status retensi murid. Dari grafik ini, terlihat bahwa semua murid yang memilih mode aplikasi " (Other Institution)" dan "(Different Plan)" mengalami dropout, sementara semua murid yang memilih mode aplikasi "International" berhasil lulus. Namun, penting untuk dicatat bahwa hasil ini dipengaruhi oleh ukuran sampel yang terlalu kecil, sehingga kesimpulan yang diambil harus berhati-hati. Dari ini bisa dilihat tidak ada correlasi antara application mode dan student retention status.

# APPLICATION ORDER VS STUDENT RETENTION STATUS



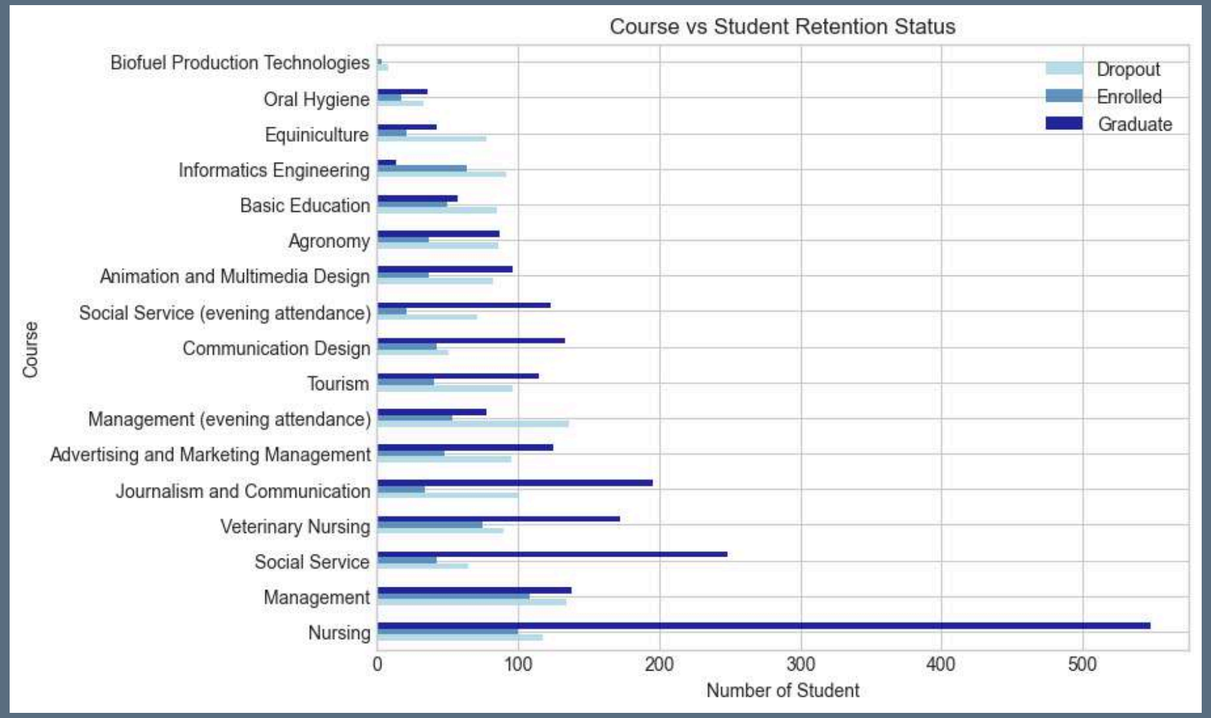



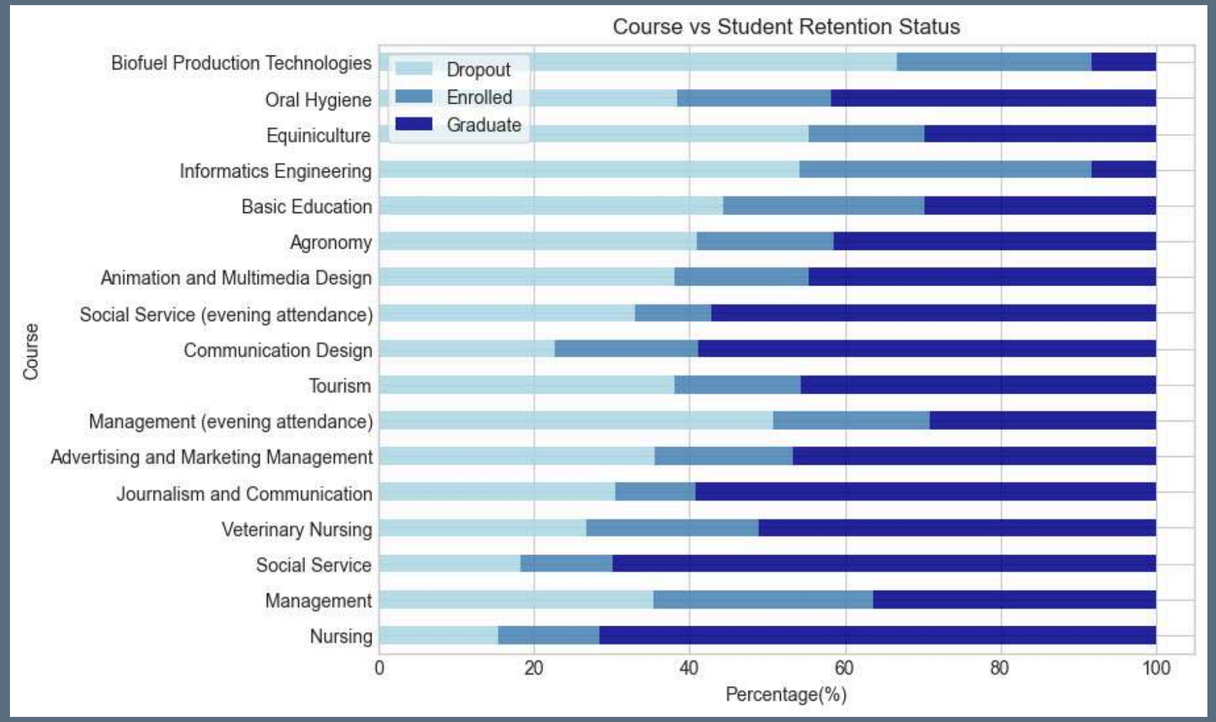

Grafik ini menunjukkan frequency application order dan juga hubunganya dengan student retention status. Bisa dilihat dari grafik ini bahwa lebih banyak murid yang mengambil application order yang pertama.

Grafik ini menunjukkan hubungan application order dengan student retention status. Bisa dilihat dari grafik ini bahwa semua murid yang mengambil application order yang ke-9 enrolled dan juga murid yang mengambil application order yang ke-0 graduate. Namun, penting untuk dicatat bahwa hasil ini dipengaruhi oleh ukuran sampel yang terlalu kecil, sehingga kesimpulan yang diambil harus berhati-hati.Dari ini bisa dilihat tidak ada correlasi antara application mode dan student retention status.

# COURSE VS STUDENT RETENTION STATUS

Dari kedua grafik di bawah ini, dapat disimpulkan bahwa program studi Nursing (Keperawatan) adalah yang paling banyak dipilih oleh murid, serta memiliki tingkat kelulusan yang tertinggi. Hal ini menunjukkan bahwa murid yang memilih program studi ini cenderung memiliki tingkat retensi yang baik. Di sisi lain, program studi Biofuel Production Technologies menunjukkan persentase dropout yang paling tinggi, mengindikasikan tantangan tertentu yang dihadapi oleh murid dalam mempertahankan studi mereka di bidang ini. Sementara itu, program studi Informatics Engineering memiliki persentase retensi yang cukup signifikan, meskipun tidak setinggi Nursing.

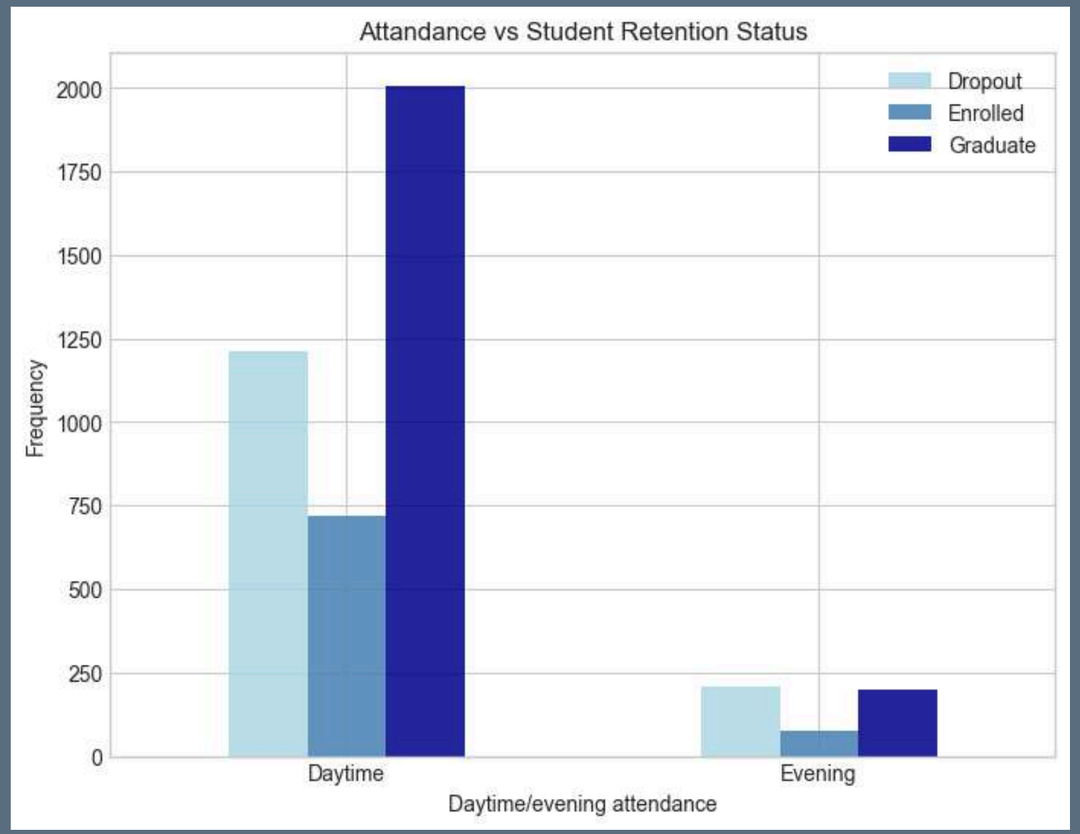


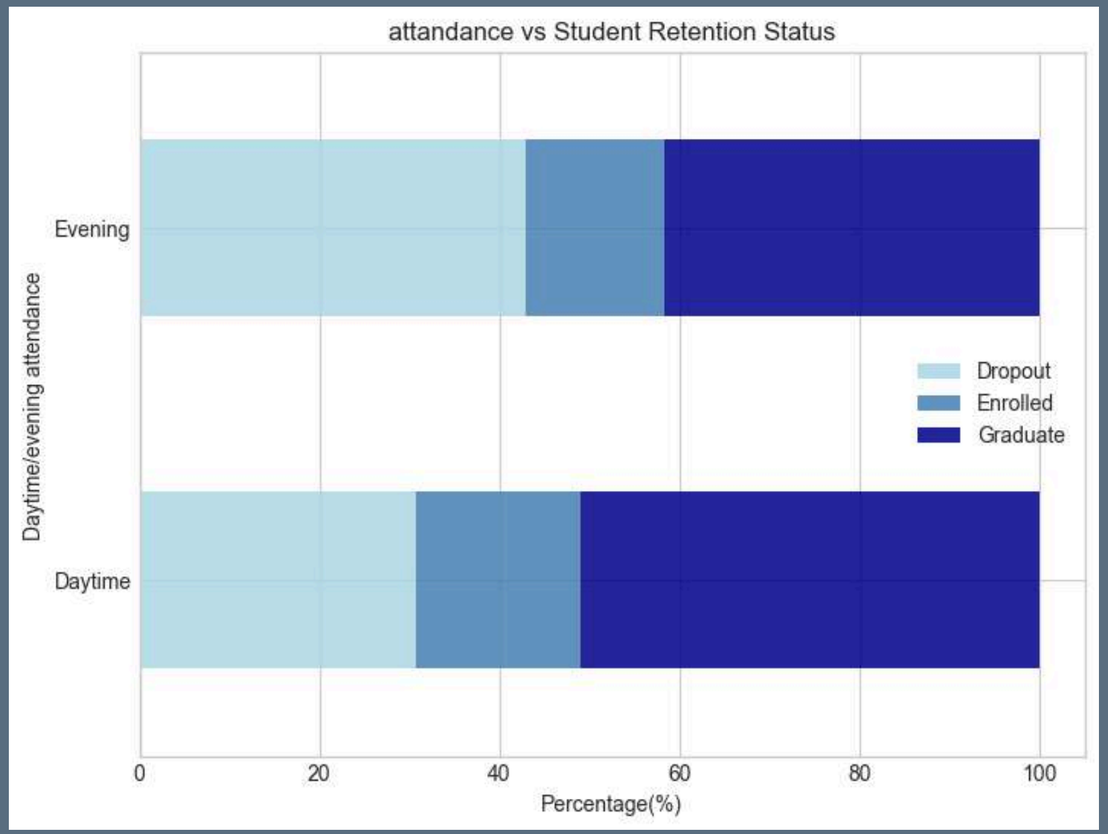



# ATTANDANCE VS STUDENT RETENTION STATUS

Grafik ini menunjukkan frequency attendace dan juga hubunganya dengan student retention status. Bisa dilihat dari grafik ini bahwa semua murid mengambil attendance di daytime sedangkan Hanya sedikit yang mengambil di evening.

Grafik ini menunjukkan hubungan attendance dengan student retention status. Bisa dilihat dari grafik ini bahwa murid yang mengambil attendance di evening lebih banyak dropout daripada murid yang mengambil attendance di daytime dimana persentase murid disana lebih banyak yang lulus.

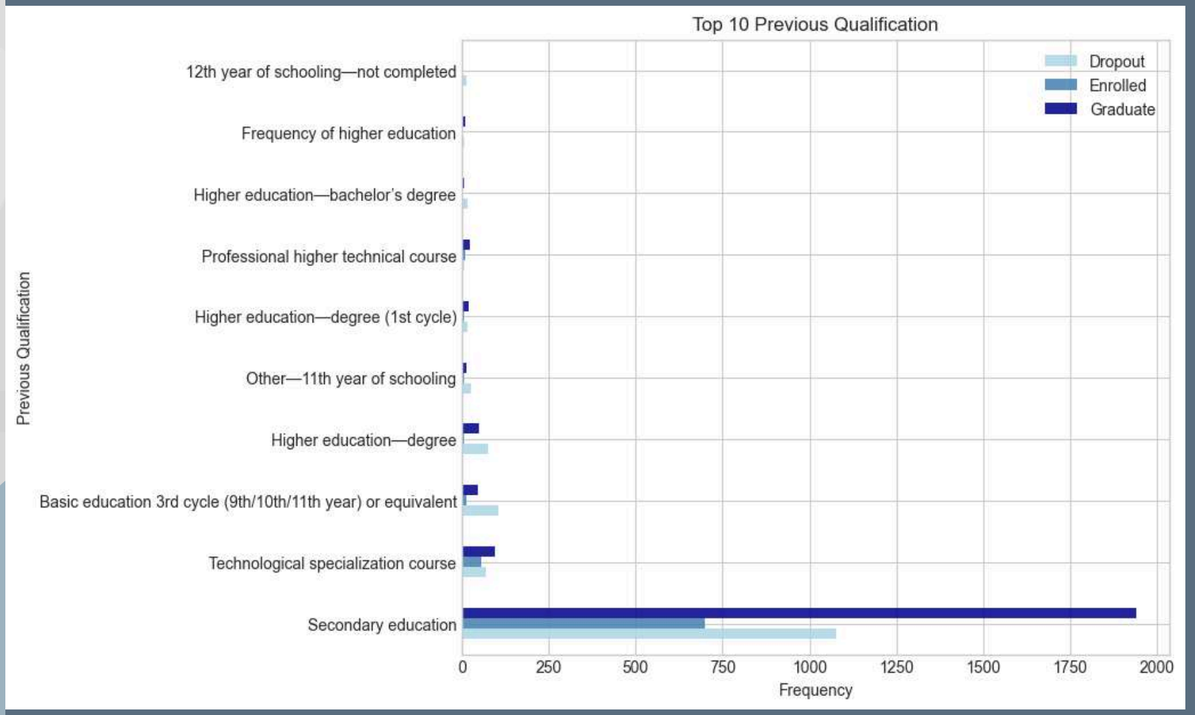


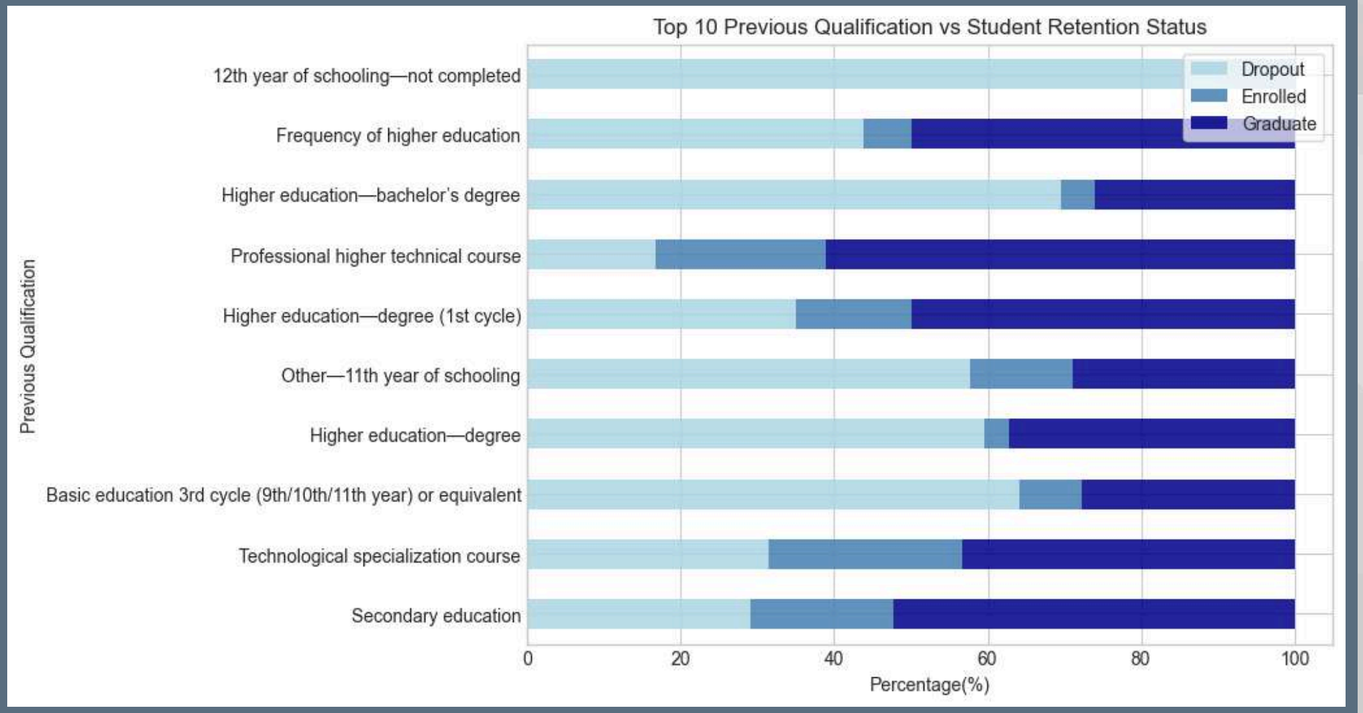



# PREVIOUS QUALIFICATION VS STUDENT RETENTION STATUS

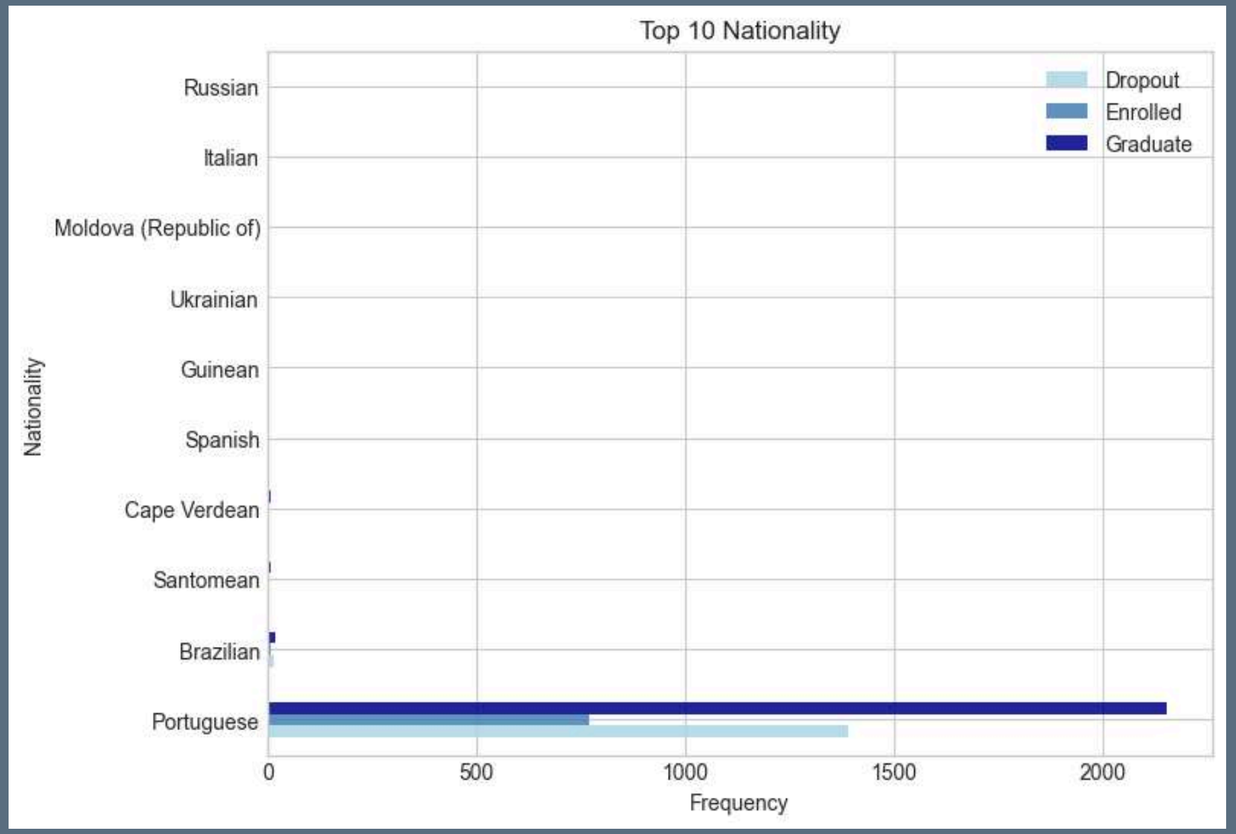
Grafik ini menunjukkan frequency of previous qualification dan juga hubunganya dengan student retention status. Bisa dilihat dari grafik inj bahwa mahasiswa yang memiliki seconday education yang paling banyak dan juga memiliki banyak jumlah kelulusan dibandingkan dengan qualifications yang lain.

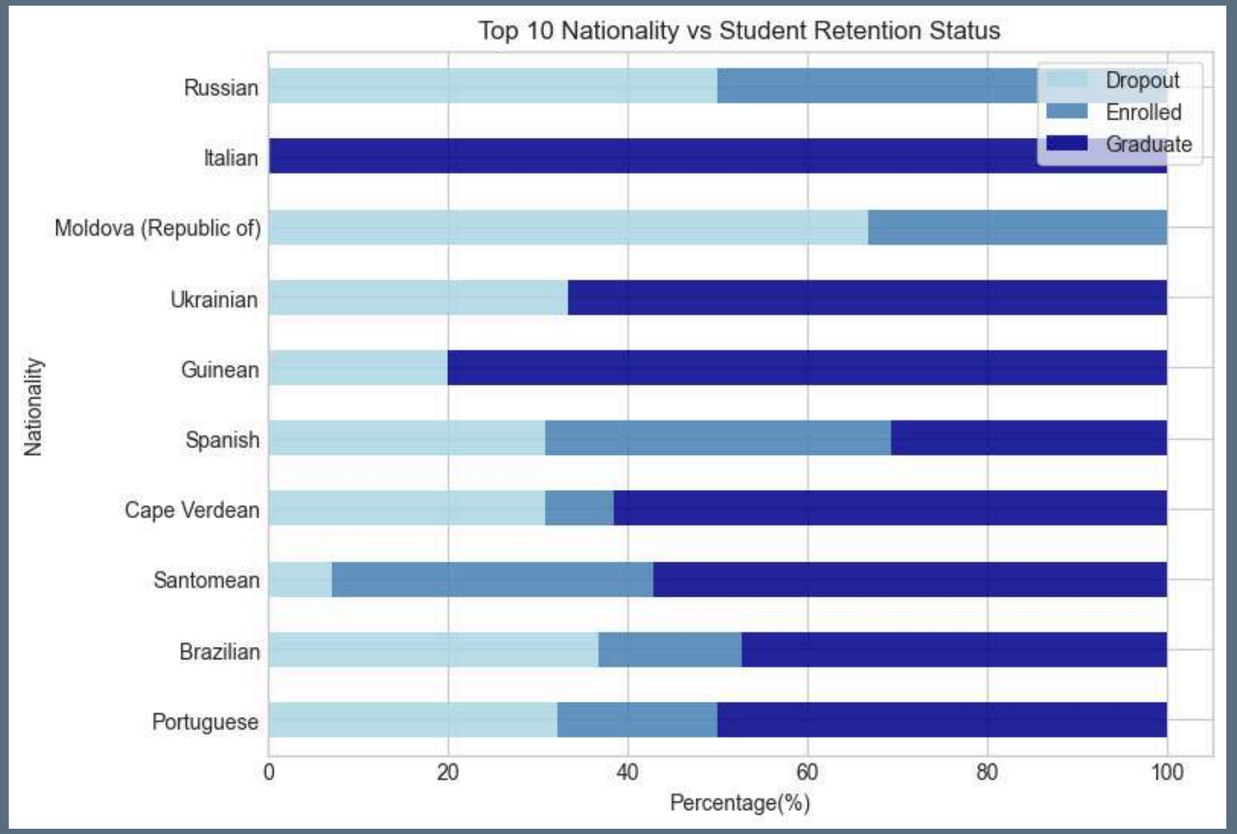
Grafik ini menunjukkan hubungan frequency previous qualification dengan student retention status. Bisa dilihat dari grafik ini bahwa murid yang tidak menyelesaikan kelas 12 dropout sedangkan murid yang memiliki professional higher technical course dan secondary education memiliki persentase kelulusam yang besar.



# NATIONALITY VS STUDENT RETENTION STATUS

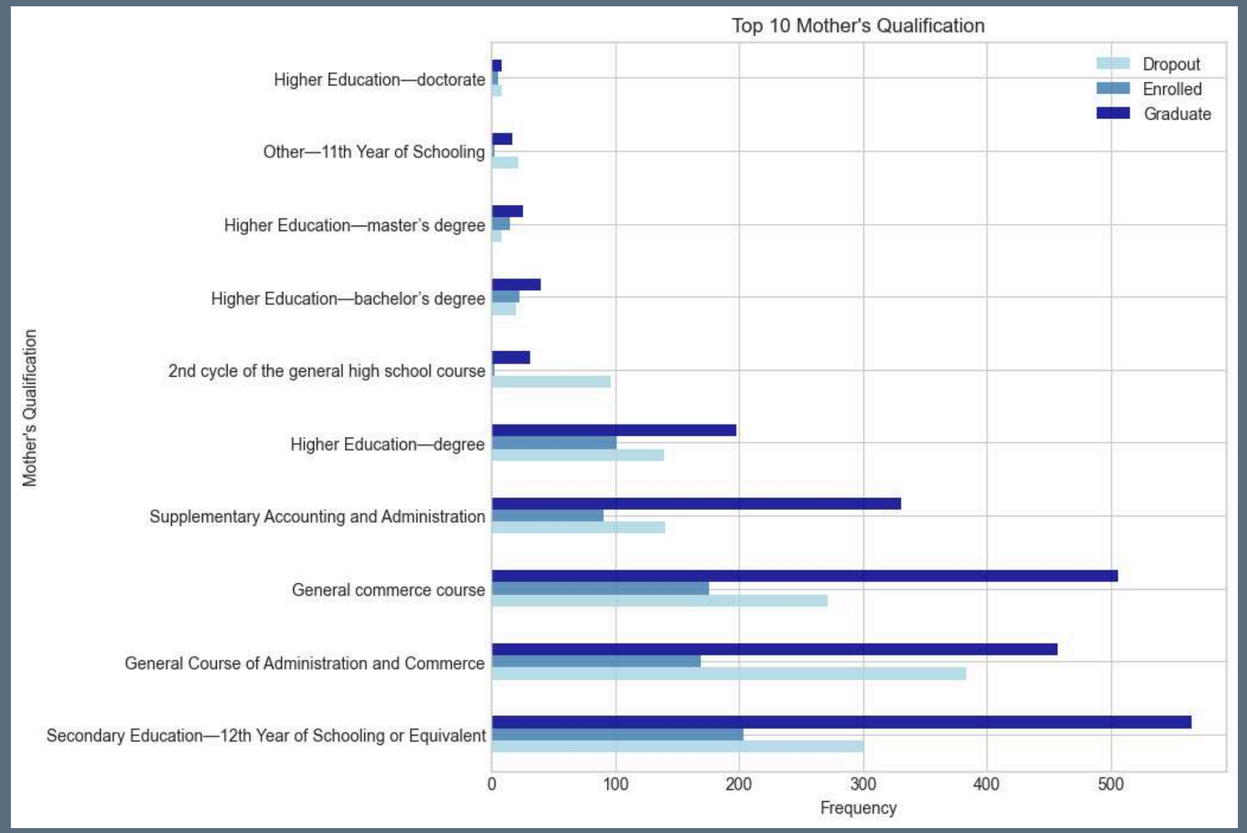


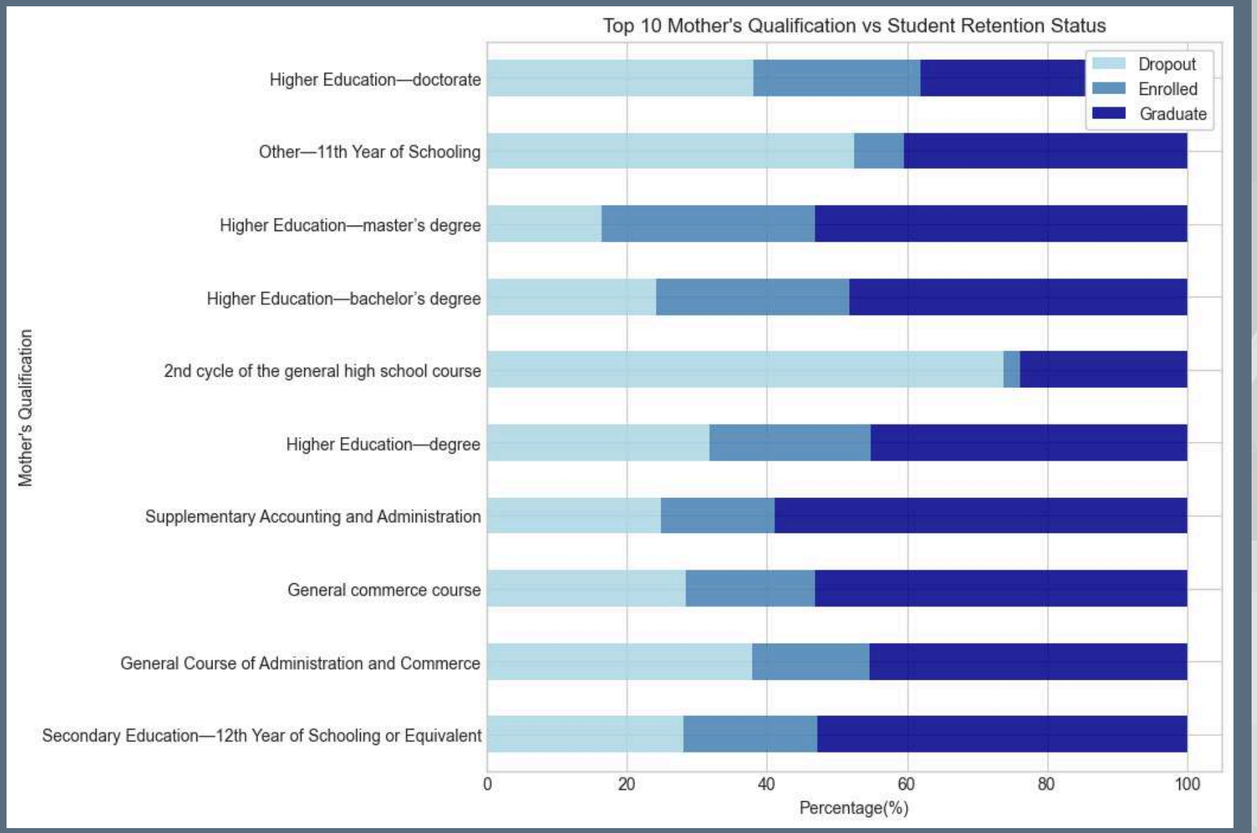




Grafik ini menunjukkan hubungan nationality dengan student retention status. Bisa dilihat dari grafik ini bahwa murid yang memiliki nationality portugese memiliki angka graduate yang paling tinggi .Namun, penting untuk dicatat bahwa hasil ini dipengaruhi oleh ukuran sampel yang terlalu kecil dan berspesifik di satu negara  sehingga kesimpulan yang diambil harus berhati-hati.

dari grafik diatas kita bisa melihat bahwa murid yang berasal dari Italian memiliki 100% graduate dibandingkan dari murid murid yang lain sedangkan murid yang berasal dari republic of Moldova tidak ada yang graduate.

# MOTHER'S QUALIFICATION VS STUDENT RETENTION STATUS

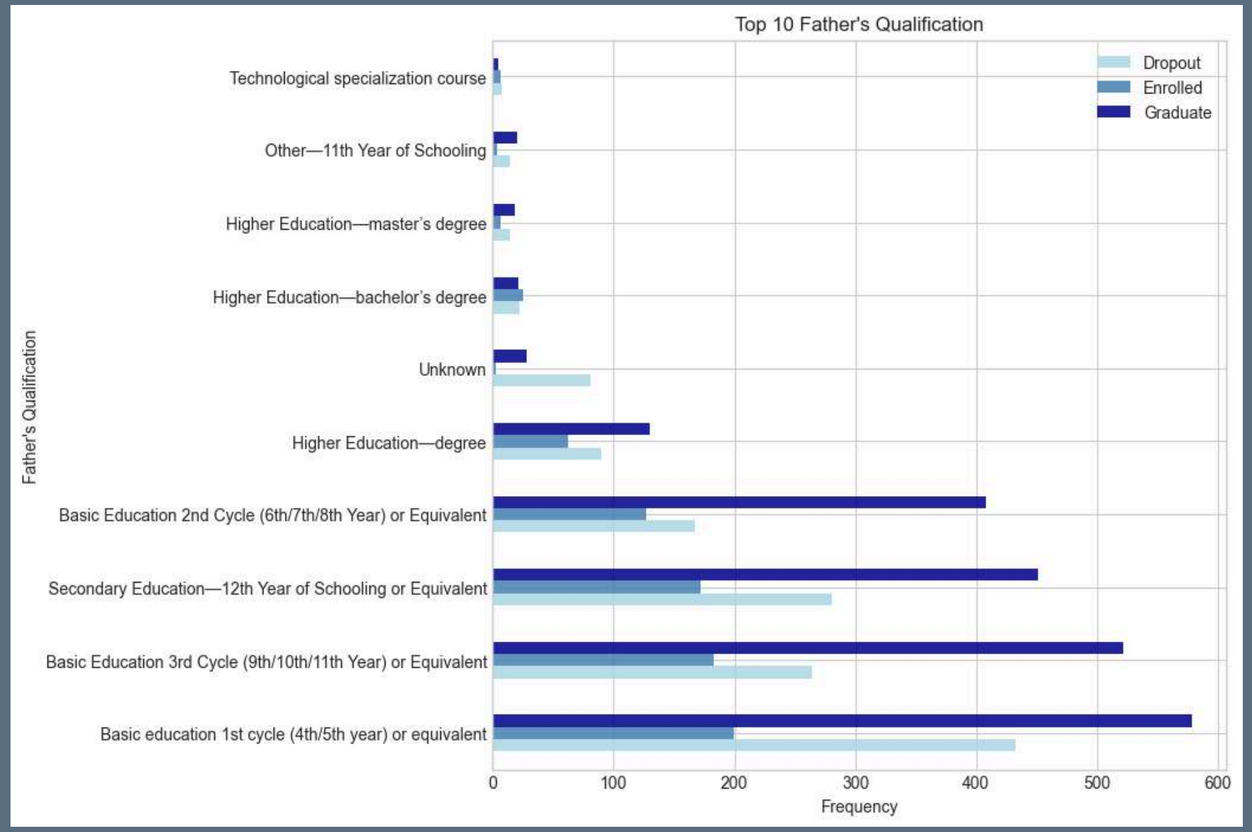


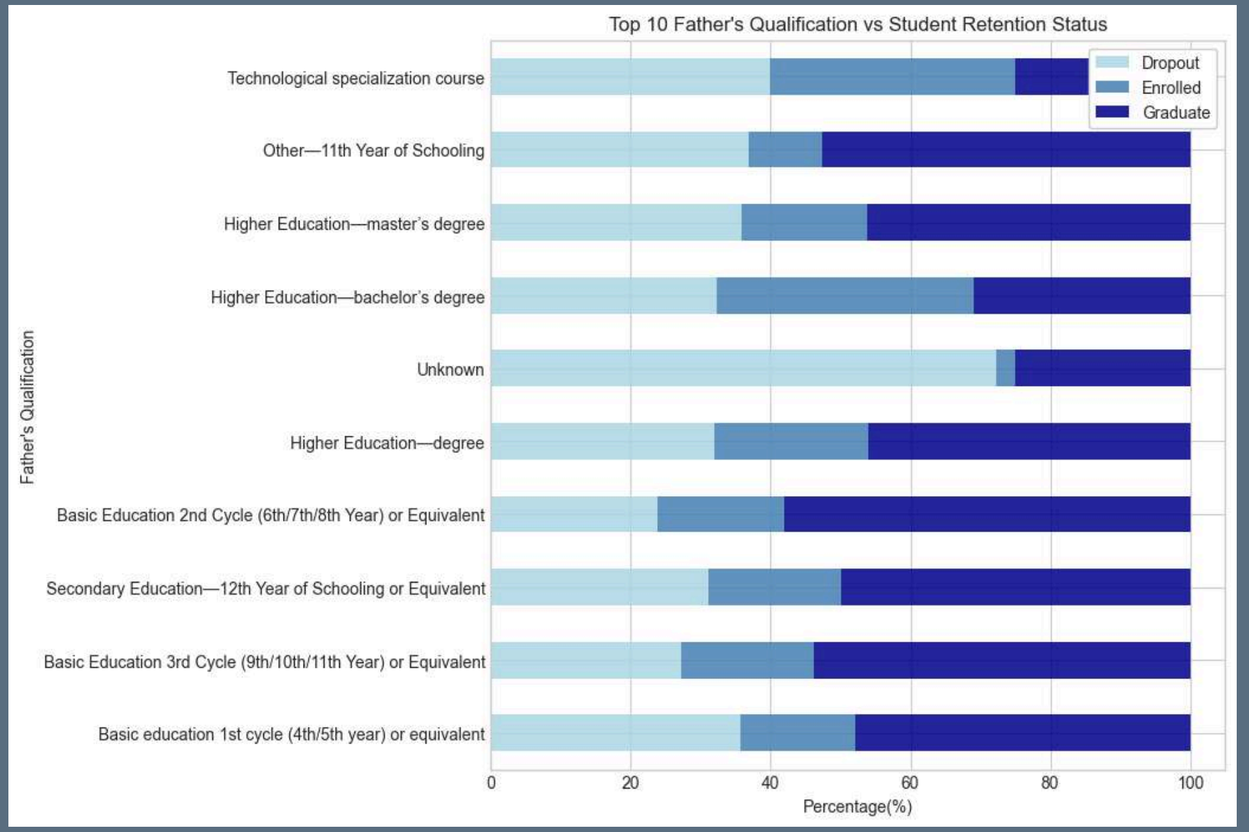




Grafik ini menunjukkan hubungan antara mother qualifications dengan student retention status, kita bisa melihat bahwa pendidikan ibu memiliki pengaruh terhadap jumlah siswa. Ibu dengan pendidikan tinggi lebih banyak memiliki anak yang bersekolah.

Grafik ini menunjukkan frequency of mother qualifications dan juga hubunganya dengan student retention status. Bisa Grafik ini menunjukkan bahwa pendidikan ibu memiliki pengaruh terhadap status retensi siswa. Ibu dengan pendidikan tinggi lebih banyak memiliki anak yang lulus sekolah.

# FATHER'S QUALIFICATION VS STUDENT RETENTION STATUS

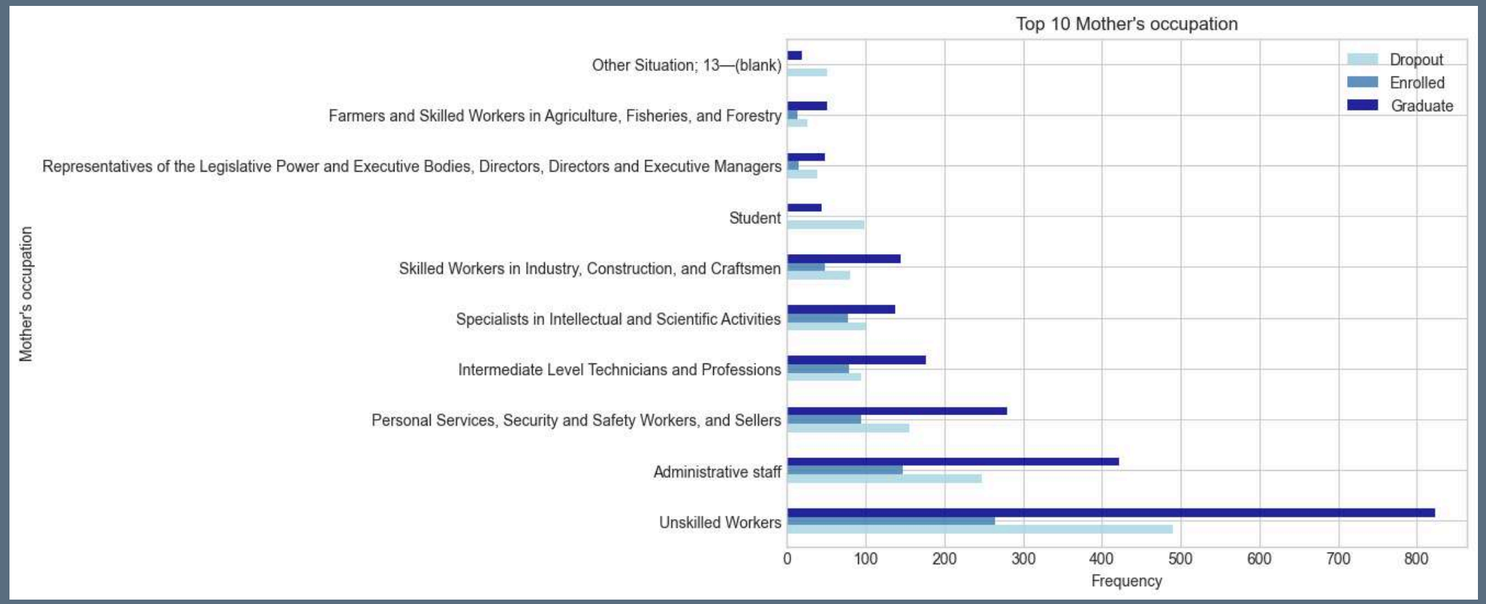


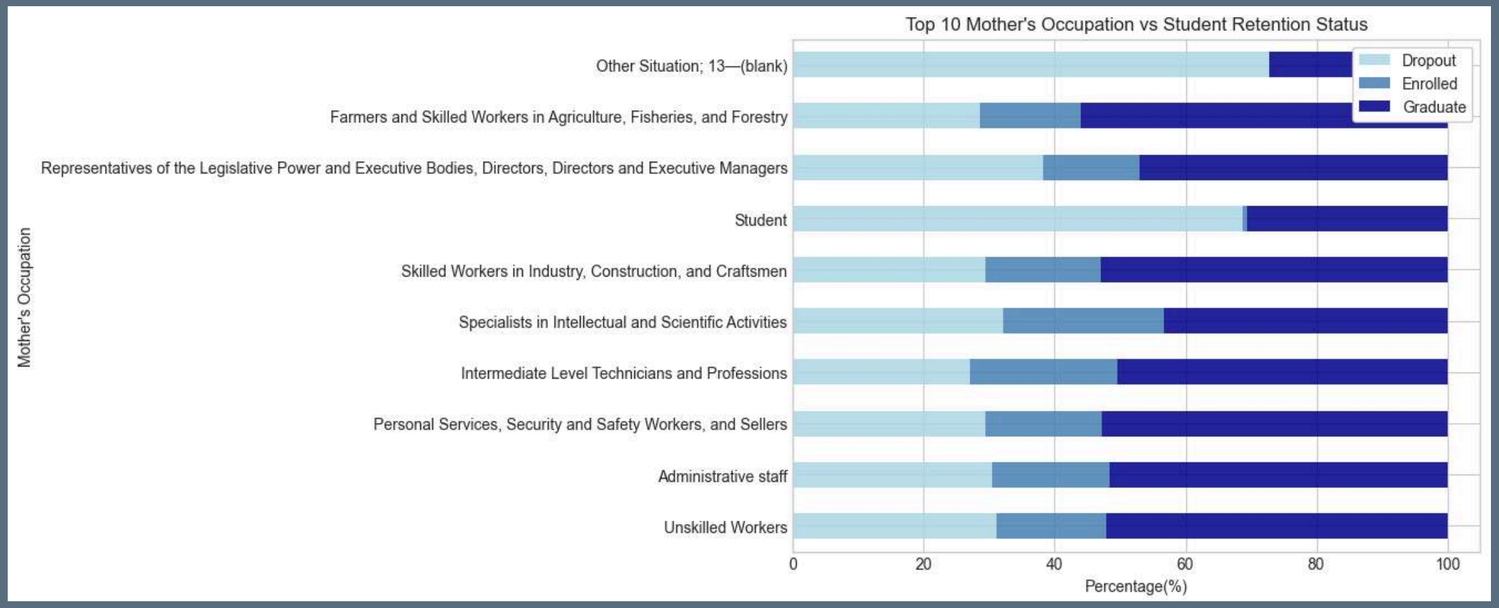

Graf ini menunjukan hubungan father qualificationsdan bisa dilihat bahwa sebagian besar ayah memiliki tingkat pendidikan yang cukup tinggi, dengan puncaknya di pendidikan menengah pertama. Meskipun frekuensi pendidikan tinggi lebih rendah, tetap terdapat proporsi yang signifikan dari ayah dengan pendidikan tersebut.

Graf ini menunjukkan hubungan yang kuat antara kualifikasi ayah dan status retensi siswa. Siswa yang memiliki ayah dengan tingkat pendidikan lebih tinggi cenderung tetap bersekolah dan mencapai tingkat kelulusan yang lebih tinggi. Namun, perlu diingat bahwa terdapat pengecualian pada tingkatan pendidikan dasar 1st Cycle yang perlu diperhatikan lebih lanjut

# MOTHER OCCUPATION VS STUDENT RETENTION STATUS

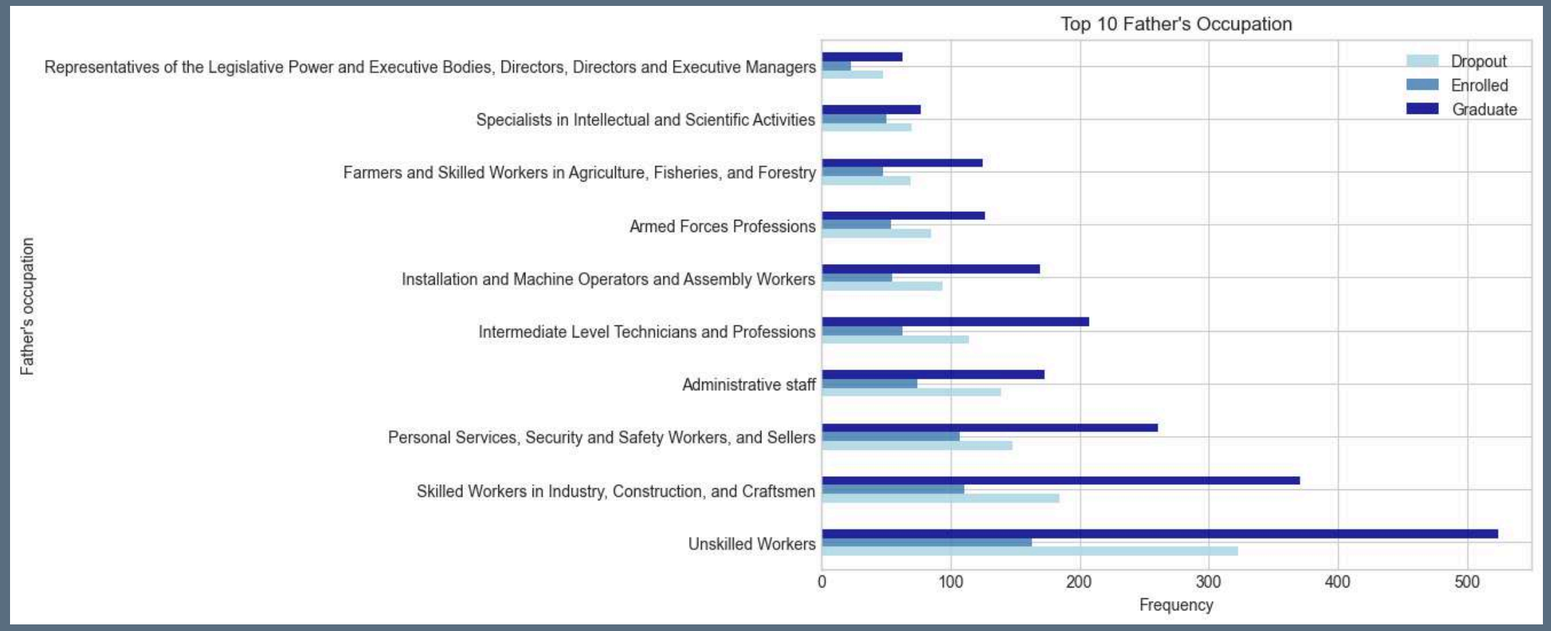



Grafik ini menunjukkan frequency mother qualifications dan juga hubunganya dengan student retention status. Bisa dilihat dari grafik ini bahwa lebih banyak murid yang ibunya adlaah unskilled worker graduate dibandingkan dengan yang lain .

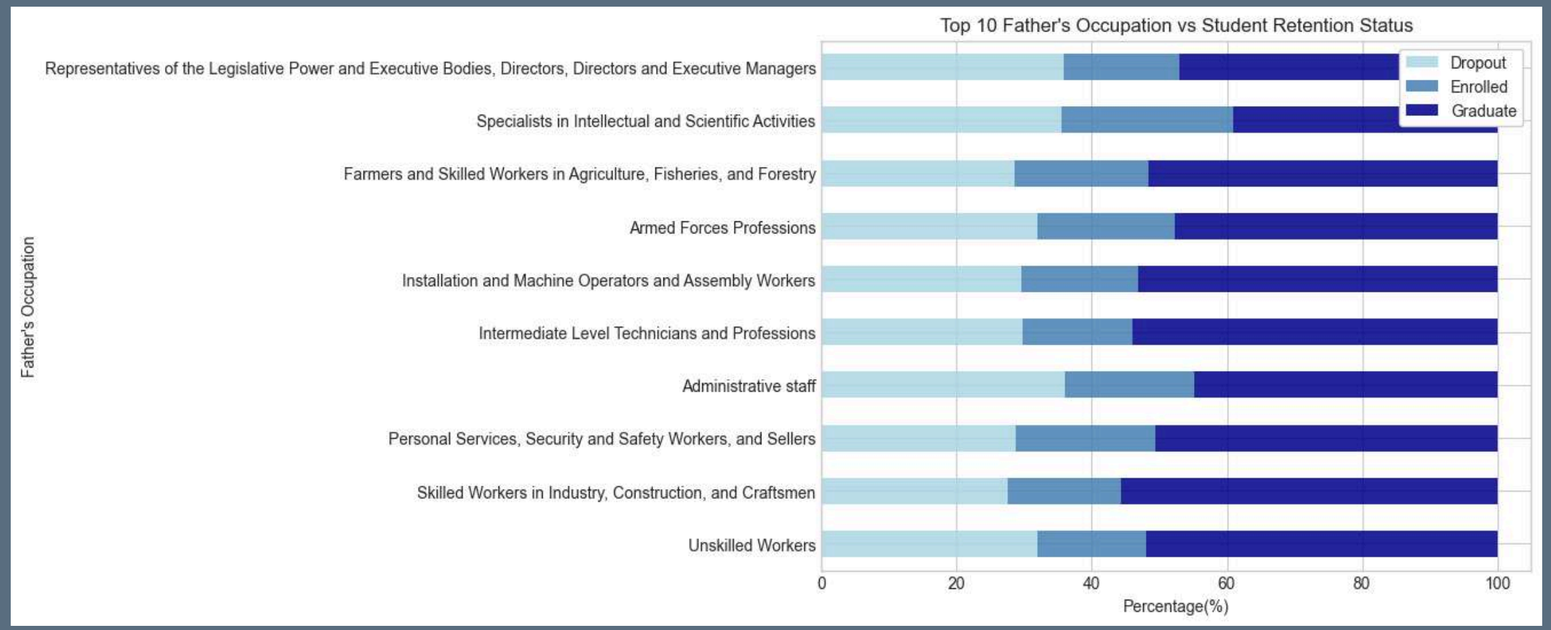



Grafik ini menunjukkan hubungan antara pekerjaan ibu dan status studi siswa. Dari grafik ini, terlihat bahwa persentase dropout lebih tinggi pada kelompok ibu yang bekerja sebagai Unskilled Worker,Administrative Staff", dan Personal Service Workers and Safety Workers. Sebaliknya, persentase siswa yang enrolled dan graduate lebih tinggi pada kelompok ibu yang bekerja di bidang Specialists in Intellectual and Scientific Activities, Intermediate Level Technicians and Professionals, dan Skilled Workers in Industry, Construction, and Craftmen.

# FATHER OCCUPATION VS STUDENT RETENTION STATUS

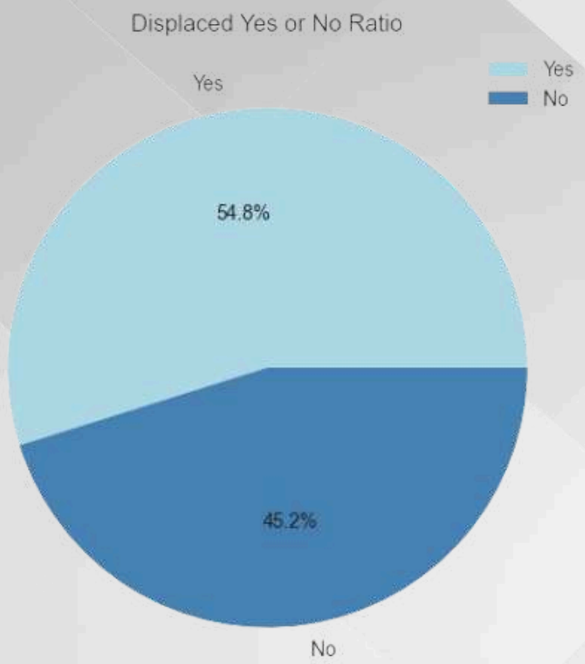


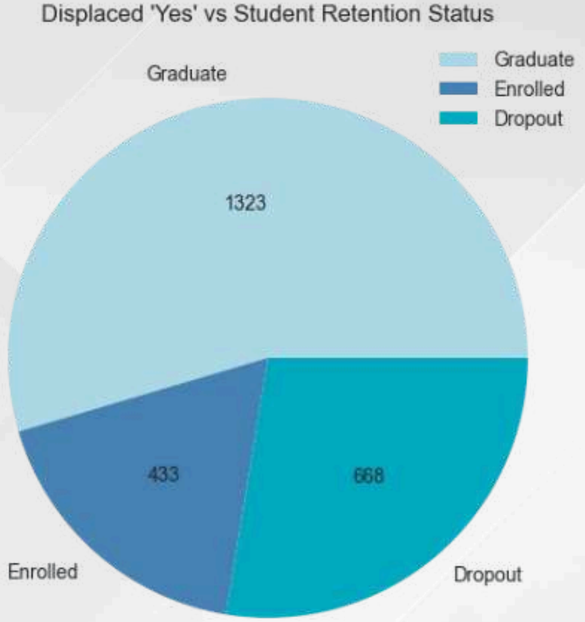

Grafik ini menunjukkan frequency father qualifications dan juga hubunganya dengan student retention status. Bisa dilihat dari grafik ini bahwa lebih banyak murid yang ayahnya adlaah unskilled worker graduate dibandingkan dengan yang lain .

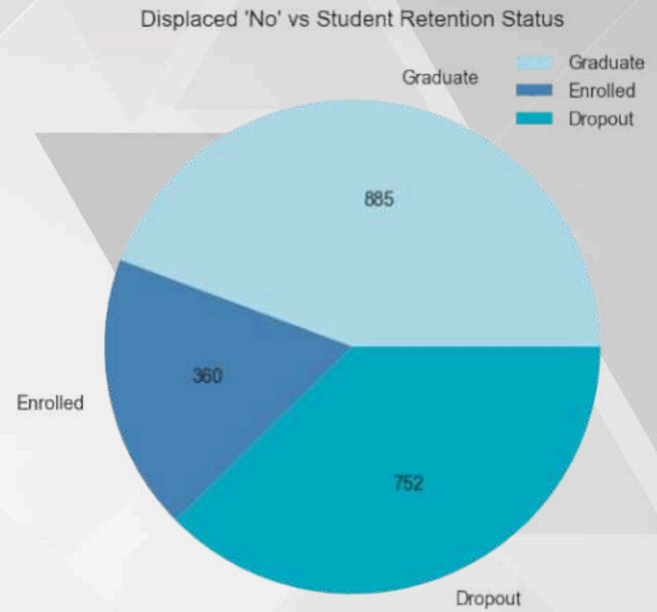



Grafik ini menunjukkan hubungan father qualifications dengan student retention status. Bisa dilihat dari grafik ini bahwa semua data tersebut memiliki angka yang sangat rata di dalam semua tipe qualifications.dengan jumlah drop out di kisaran 25% sedangkan enrolled di kisaran 10 samai 15% dan jumlah yang paling banyak adalah graduate demgan kisaran 50%.

# DISPLACED VS STUDENT RETENTION STATUS




Grafik ini adalah diagram lingkaran yang menunjukkan persentase siswa yang "displaced" (ya atau tidak).

Hasilnya:

54.8% dari siswa  termasuk dalam kategori "displaced" (yes).
45.2% dari siswa termasuk dalam kategori "displaced" (no).

Diagram ini menunjukkan distribusi status retensi siswa yang "displaced" (Yes).

Hasilnya:

Sebanyak 1323 siswa lulus (graduate).
Sebanyak 433 siswa masih terdaftar (enrolled).
Sebanyak 668 siswa putus sekolah (dropout).

Diagram ini menunjukkan distribusi status retensi siswa yang tidak "displaced" (No).

Hasilnya:

Sebanyak 885 siswa lulus (graduate).
Sebanyak 360 siswa masih terdaftar (enrolled).
Sebanyak  752 siswa putus sekolah (dropout).

# EDUCATION SPECIAL NEEDS VS STUDENT RETENTION STATUS

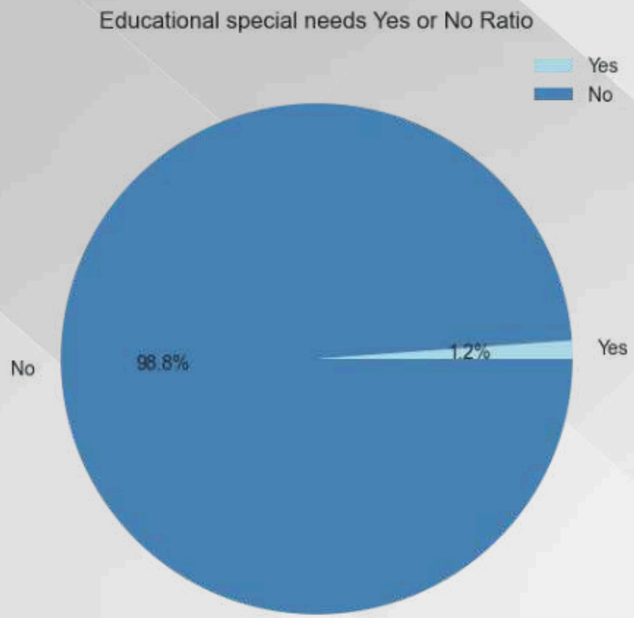



Diagram lingkaran ini menunjukkan distribusi siswa berdasarkan apakah mereka memiliki kebutuhan pendidikan khusus atau tidak.

Hasilnya:

Sebagian besar siswa (98.8%) tidak memiliki kebutuhan pendidikan khusus (No).

Hanya 1.2% siswa yang memiliki kebutuhan pendidikan khusus (Yes).

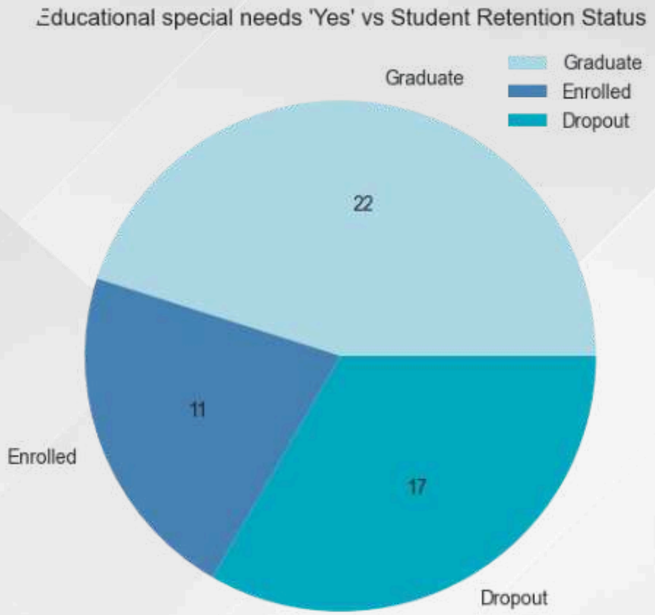



Diagram ini menunjukkan distribusi status retensi siswa yang memiliki kebutuhan pendidikan khusus (Yes).

Hasilnya:
Sebanyak 22 siswa lulus (graduate).
Sebanyak 11 siswa masih terdaftar (enrolled).
Sebanyak 17 siswa putus sekolah (dropout).

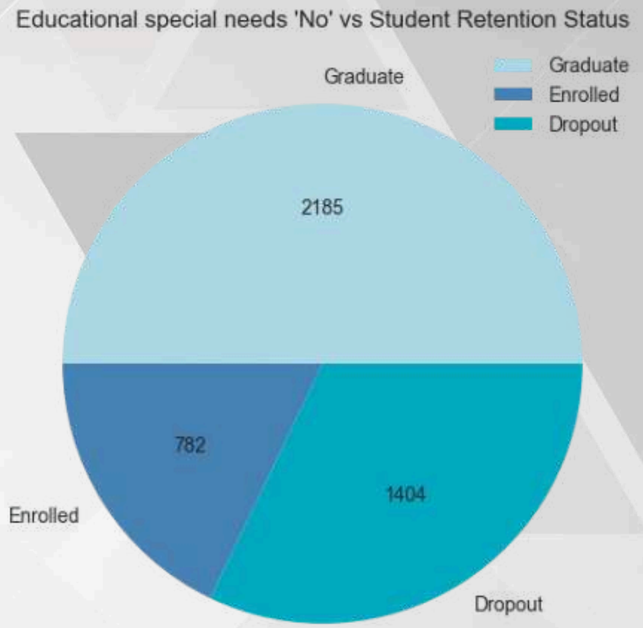
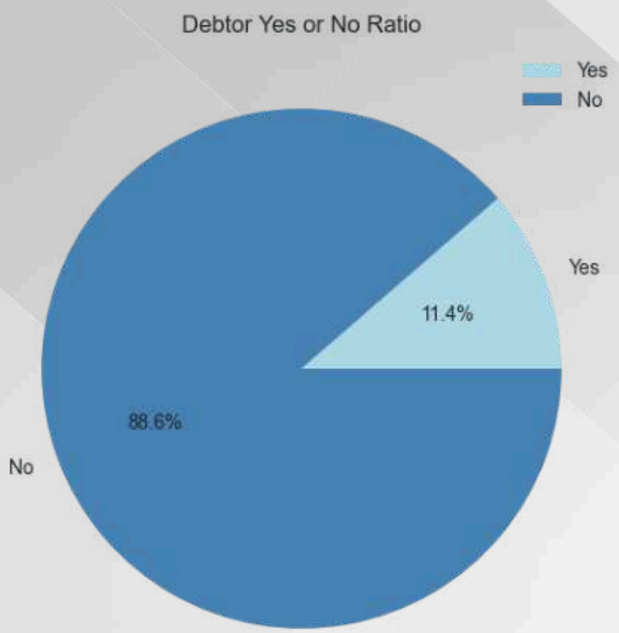
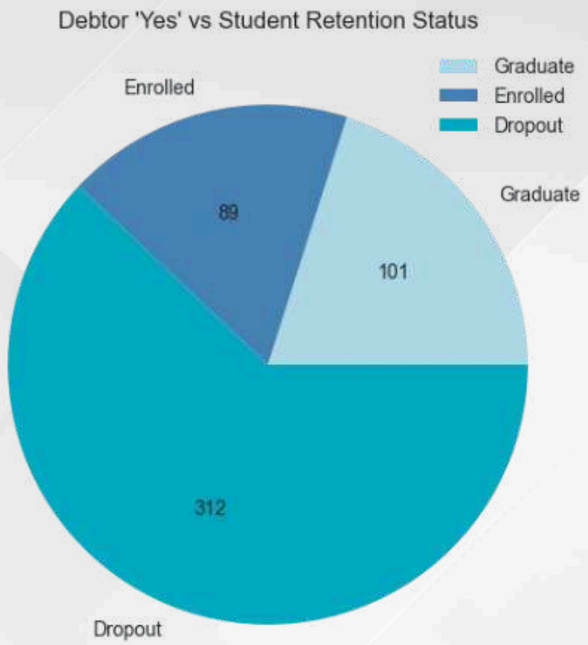



Diagram ini menunjukkan distribusi status retensi siswa yang tidak memiliki kebutuhan pendidikan khusus (No).

Hasilnya:
Sebanyak 2185 siswa lulus (graduate).
Sebanyak 782 siswa masih terdaftar (enrolled).
Sebanyak 1404 siswa putus sekolah (dropout).

# DEBTOR VS STUDENT RETENTION STATUS



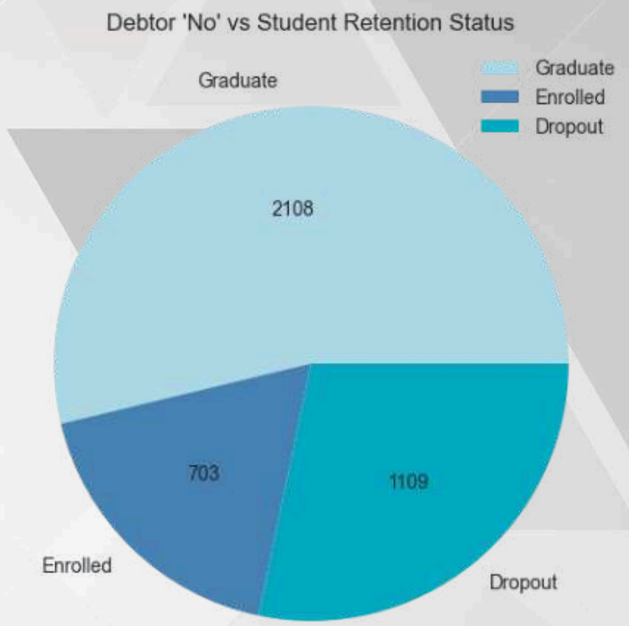




Diagram lingkaran ini menunjukkan distribusi siswa berdasarkan status debitur mereka (Yes/No).

Hasilnya:

88.6% siswa tidak memiliki status debitur (No).

11.4% siswa memiliki status debitur (Yes).

Diagram ini menunjukkan distribusi status retensi siswa yang memiliki status debitur (Yes).

Hasilnya:

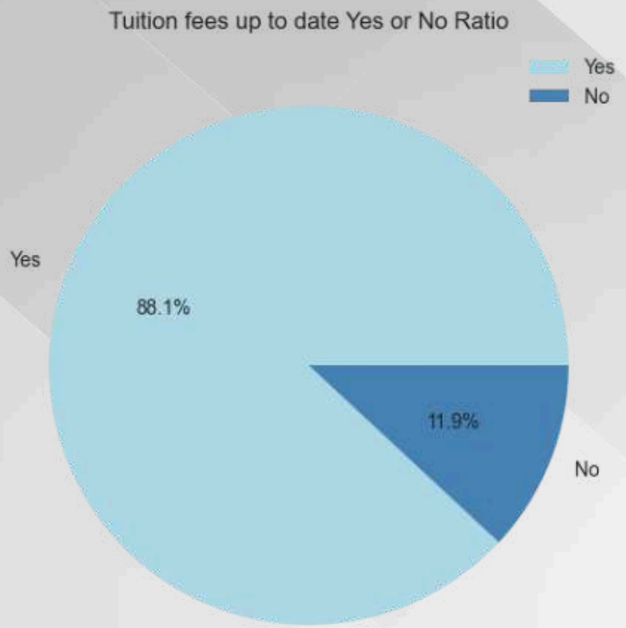
101 siswa lulus (graduate).
89 siswa masih terdaftar (enrolled).
312 siswa putus sekolah (dropout).

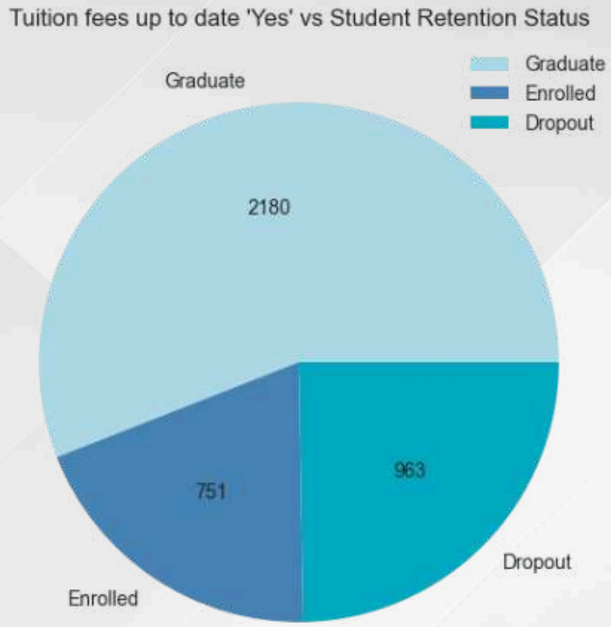
Diagram ini menunjukkan distribusi status retensi siswa yang tidak memiliki status debitur (No).

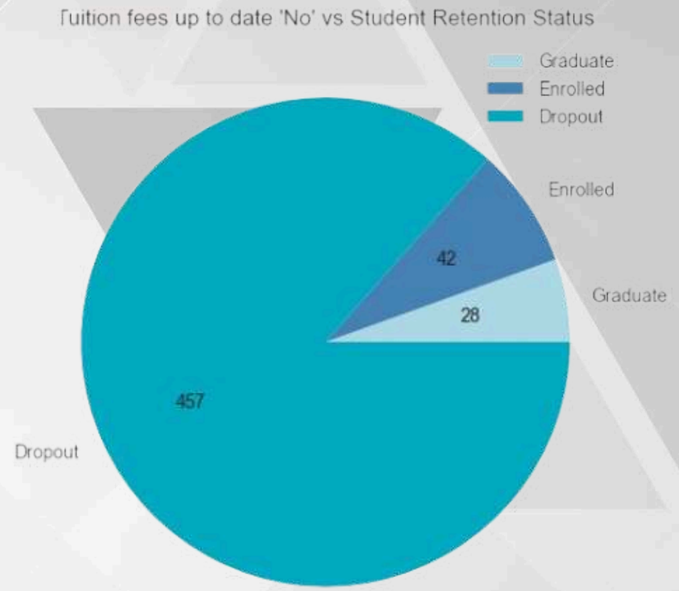
Hasilnya:

2108 siswa lulus (graduate).
703 siswa masih terdaftar (enrolled).
1109 siswa putus sekolah (dropout).

# TUITION FEES UP TO DATE VS STUDENT RETENTION STATUS




Menunjukkan proporsi siswa yang telah membayar uang kuliah tepat waktu (Yes) dan yang belum (No).

Interpretasi:

Yes (88.1%): Sebagian besar siswa telah membayar uang kuliah tepat waktu.

No (11.9%): Sebagian kecil siswa belum membayar uang kuliah tepat waktu.

Mengelompokkan siswa yang membayar uang kuliah tepat waktu (Yes)

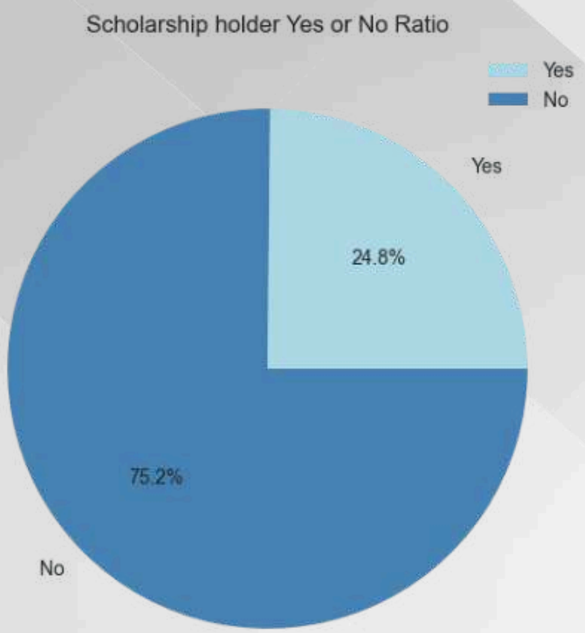
Interpretasi:

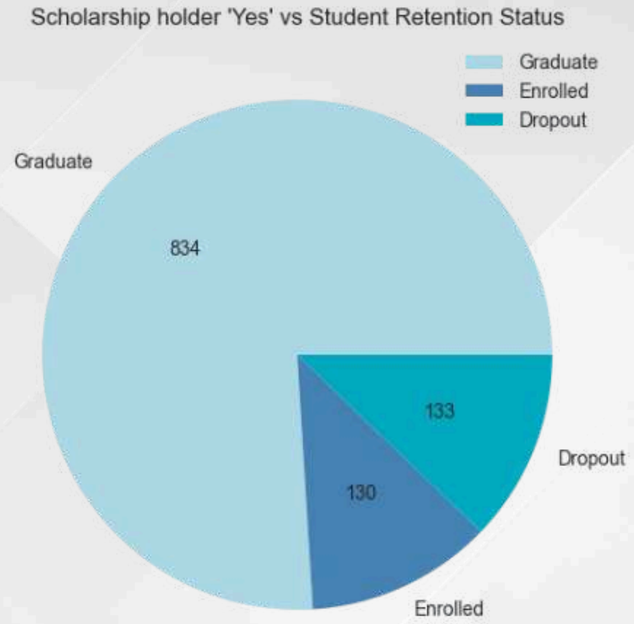
Graduate (2180): Mayoritas siswa yang membayar tepat waktu berhasil lulus.
Enrolled (751): Sebagian siswa masih terdaftar.
Dropout (963): Sebagian kecil siswa berhenti meskipun telah membayar tepat waktu.

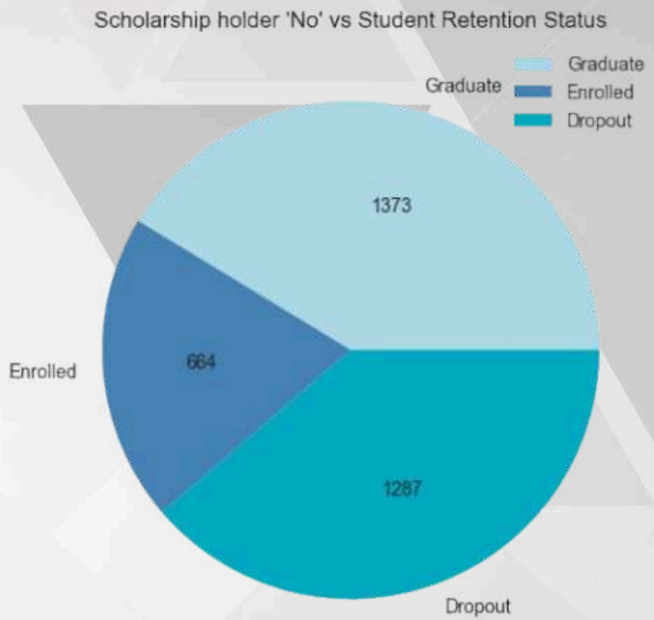
Mengelompokkan siswa yang belum membayar uang kuliah tepat waktu (No) berdasarkan status mereka:

Interpretasi:
Graduate (28): Sangat sedikit siswa yang lulus meskipun belum membayar tepat waktu.
Enrolled (42): Hanya sedikit siswa yang masih terdaftar tanpa pembayaran tepat waktu.
Dropout (457): Sebagian besar siswa berhenti jika tidak membayar uang kuliah tepat waktu.

# SCHOLARSHIP HOLDER VS STUDENT RETENTION STATUS




Menunjukkan proporsi siswa yang menerima beasiswa (Yes) dan yang tidak menerima beasiswa (No).

Interpretasi:

Yes (24.8%): Hanya sebagian kecil siswa yang menerima beasiswa.
No (75.2%): Sebagian besar siswa tidak menerima beasiswa.

Mengelompokkan siswa penerima beasiswa (Yes) berdasarkan status mereka

Interpretasi:

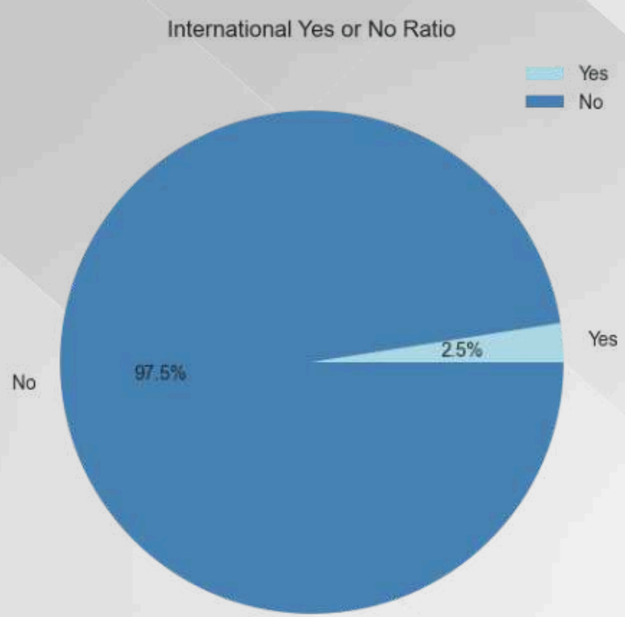
Graduate (834): Sebagian besar siswa penerima beasiswa berhasil lulus.
Enrolled (130): Sebagian kecil siswa penerima beasiswa masih terdaftar.
Dropout (133): Sebagian kecil siswa penerima beasiswa berhenti dari studi.

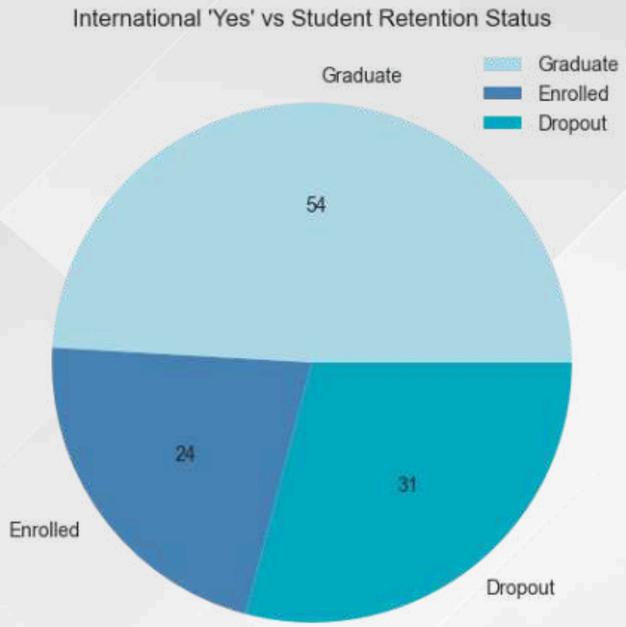
Mengelompokkan siswa yang tidak menerima beasiswa (No) berdasarkan status mereka

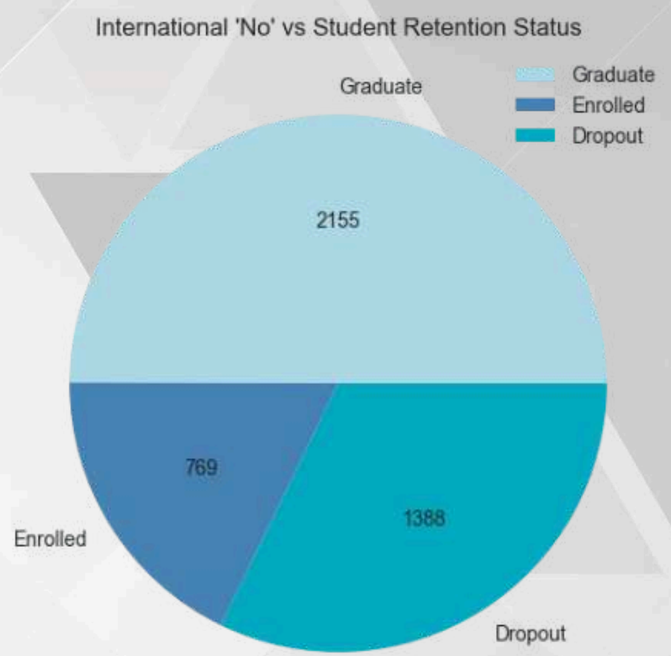
Interpretasi:

Graduate (1373): Sebagian besar siswa tanpa beasiswa tetap dapat lulus.
Enrolled (664): Sebagian siswa tanpa beasiswa masih terdaftar.
Dropout (1287): Banyak siswa tanpa beasiswa yang berhenti dari studi.

# INTERNATIONAL VS STUDENT RETENTION STATUS




Menunjukkan proporsi siswa internasional (Yes) dan non-internasional (No).

Interpretasi:

Yes (2.5%): Hanya sebagian kecil siswa yang merupakan mahasiswa internasional.
No (97.5%): Mayoritas siswa adalah non-internasional.

Mengelompokkan mahasiswa internasional (Yes) berdasarkan status mereka:

Interpretasi:

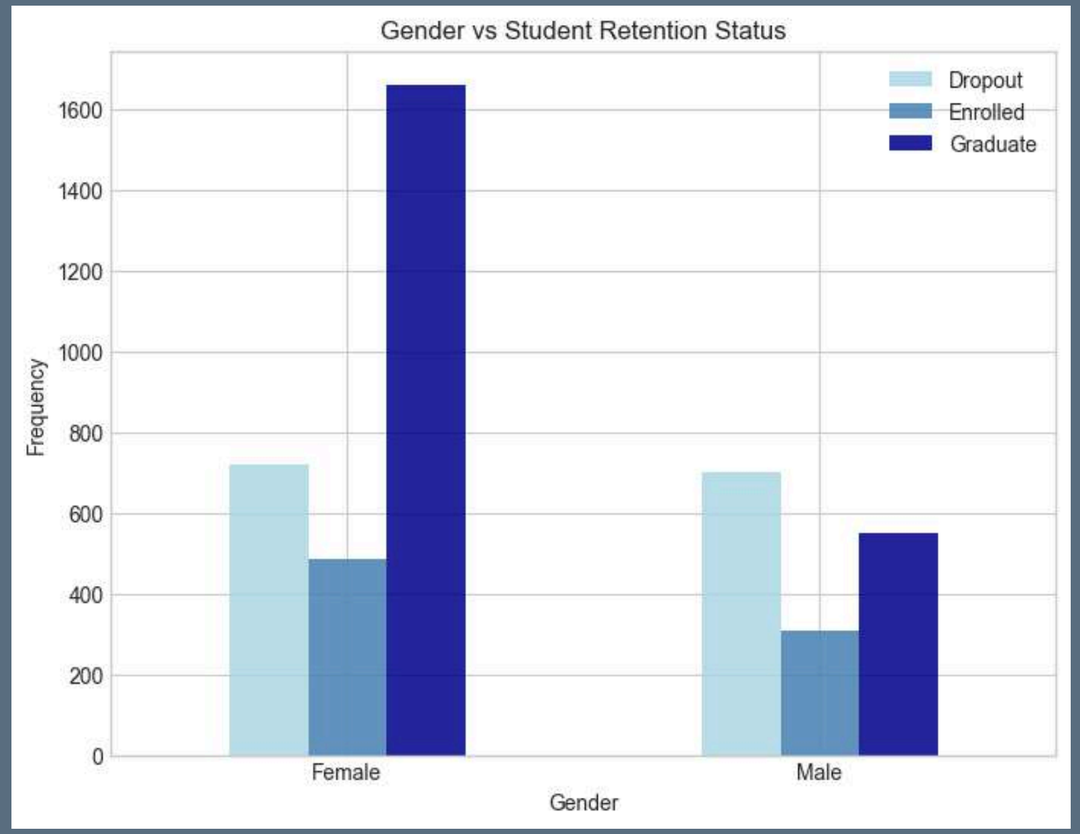
Graduate (54): Sebagian besar mahasiswa internasional berhasil lulus.
Enrolled (24): Sebagian kecil mahasiswa internasional masih terdaftar.
Dropout (31): Ada beberapa mahasiswa internasional yang berhenti dari studi mereka.

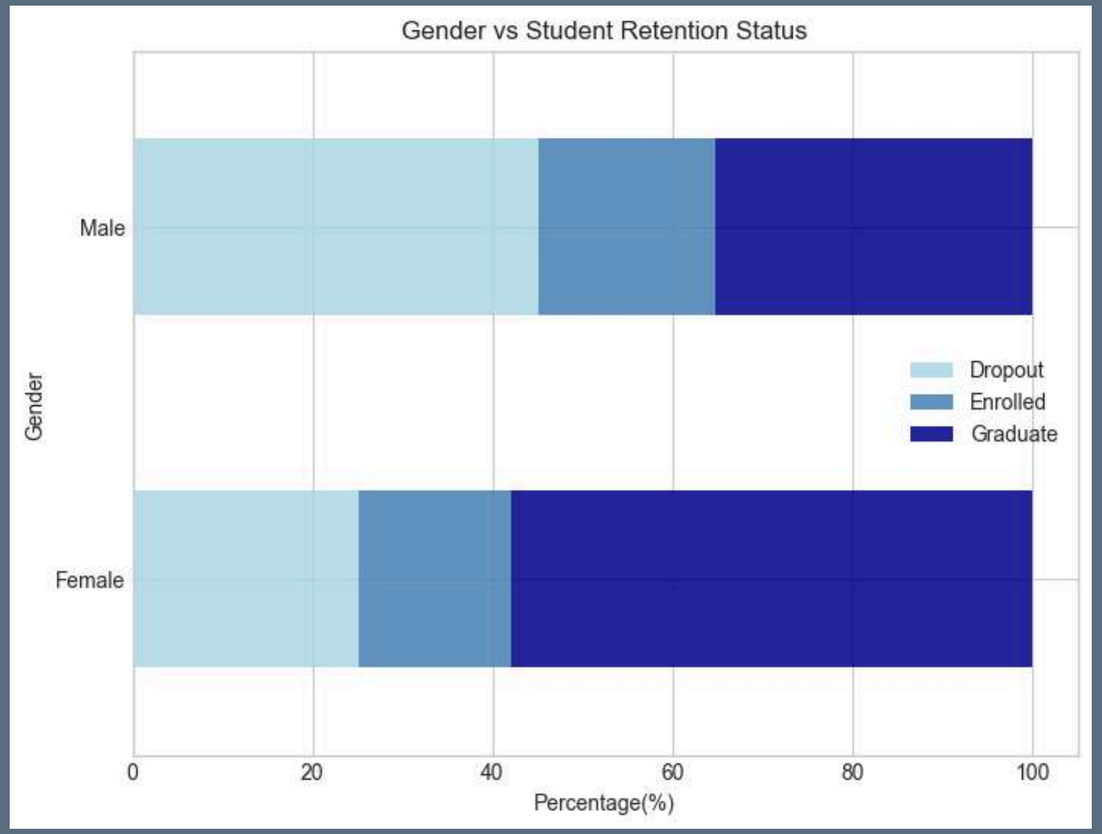
Mengelompokkan siswa non-internasional (No) berdasarkan status mereka

Interpretasi:

Graduate (2155): Sebagian besar siswa non-internasional berhasil lulus.
Enrolled (769): Sebagian siswa non-internasional masih terdaftar.
Dropout (1388): Terdapat jumlah signifikan siswa non-internasional yang berhenti

# GENDER VS STUDENT RETENTION STATUS

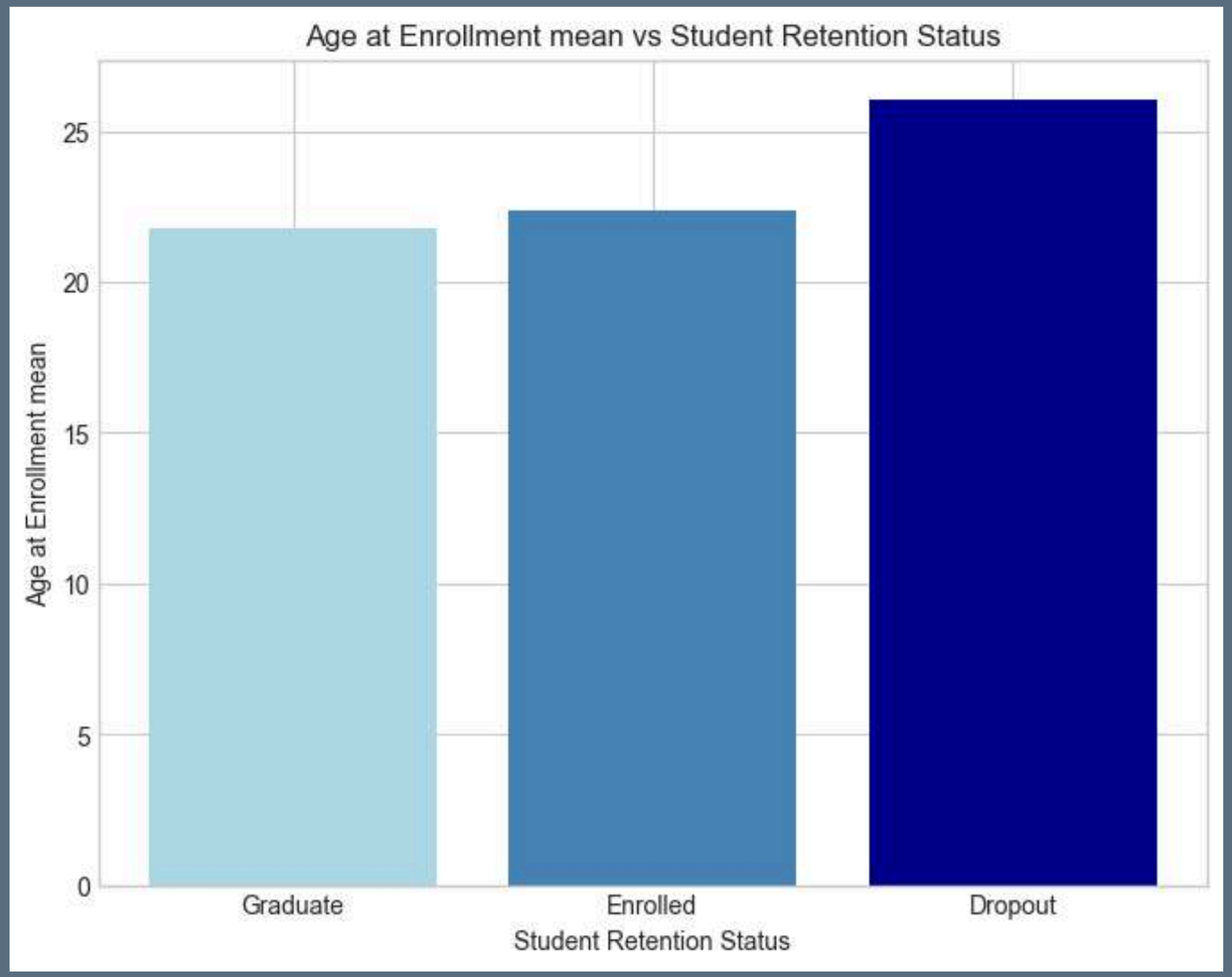



Dari grafik ini, dapat dilihat frekuensi gender murid serta hubungannya dengan status retensi murid. Grafik ini menunjukkan bahwa jumlah murid perempuan lebih banyak dibandingkan dengan murid pria. Selain itu, grafik ini juga memberikan gambaran mengenai bagaimana perbedaan gender mempengaruhi status retensi murid, seperti kelulusan, keterdaftaran, atau dropout.

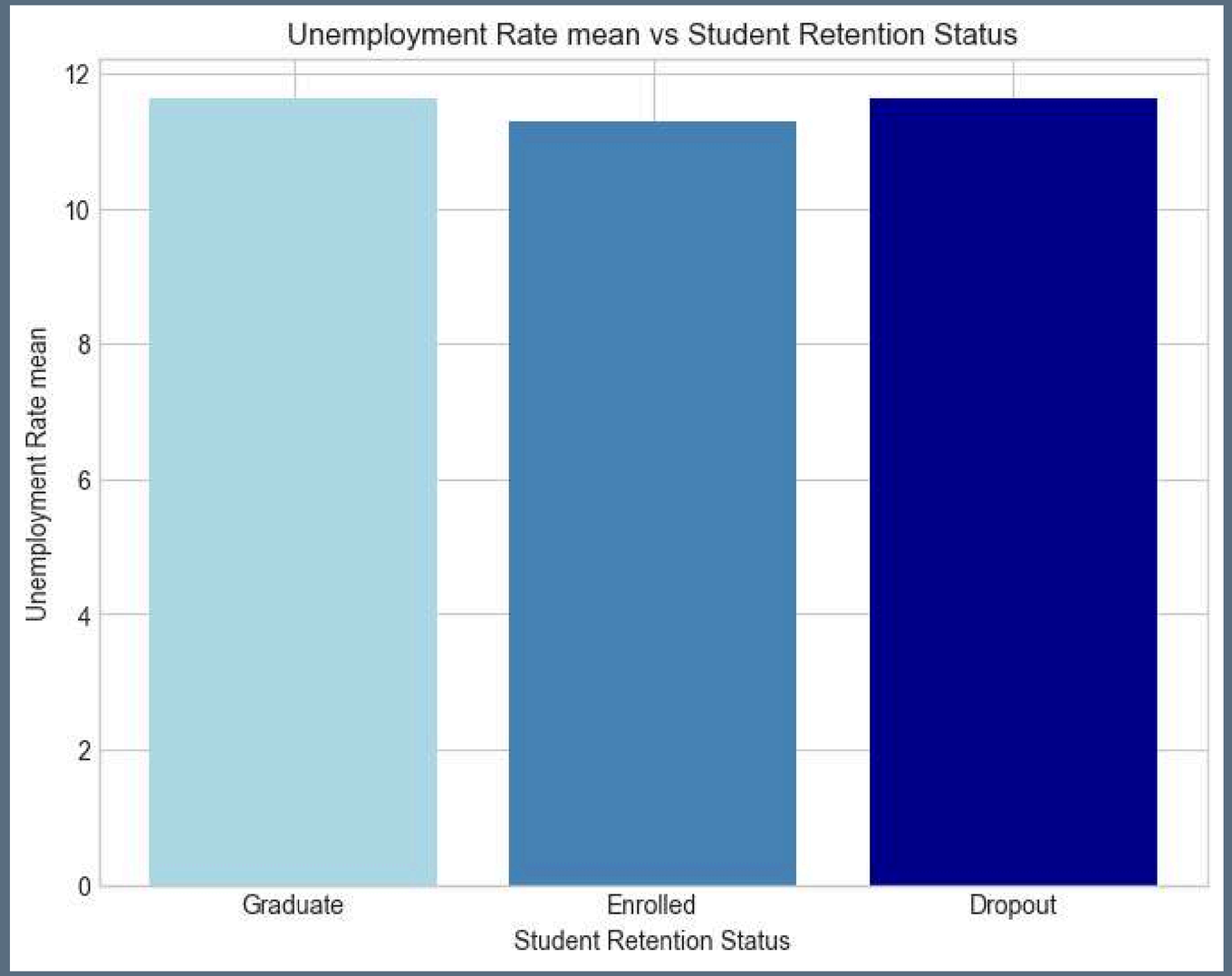
Grafik ini menggambarkan hubungan antara gender dan status retensi murid. Dari grafik tersebut, dapat dilihat bahwa murid perempuan memiliki persentase kelulusan yang lebih tinggi dibandingkan dengan murid pria. Sebaliknya, murid pria memiliki persentase dropout yang lebih besar dibandingkan dengan murid perempuan.

# AGE AT ENROLLMENT MEAN VS STUDENT RETENTION STATUS



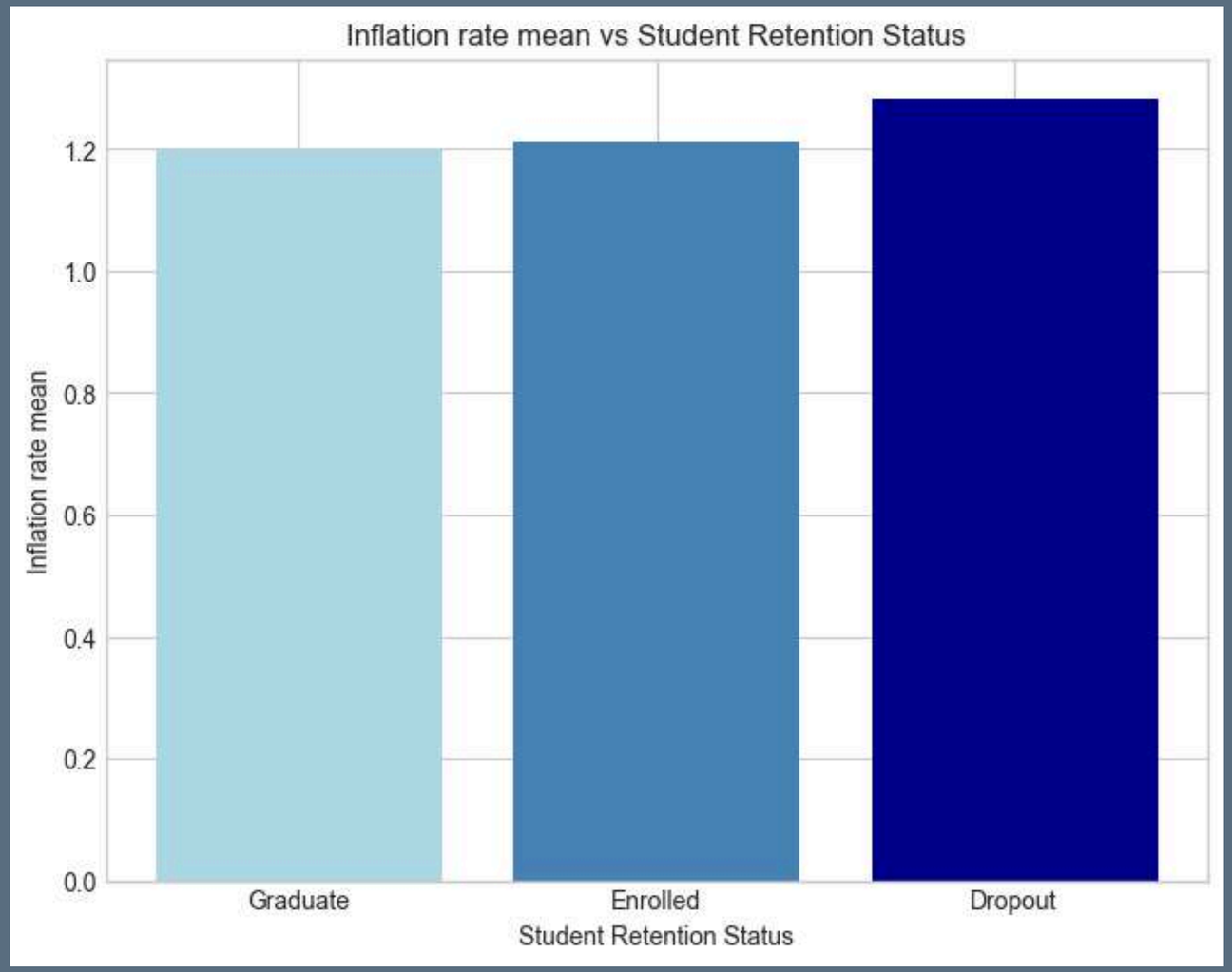

Grafik ini menggambarkan hubungan antara umur dan status retensi murid. Dari grafik tersebut, dapat dilihat bahwa rata-rata umur (mean) untuk murid yang lulus (graduate) adalah sekitar 22 tahun, sementara murid yang masih terdaftar (enrolled) memiliki rata-rata umur sekitar 23 tahun. Di sisi lain, murid yang mengalami dropout memiliki rata-rata umur yang lebih tinggi, yaitu sekitar 27 tahun. Hal ini menunjukkan adanya perbedaan usia yang signifikan antara murid yang berhasil menyelesaikan studi, yang masih aktif terdaftar, dan yang memutuskan untuk keluar dari program.

# UNEMPLOYMENT RATE MEAN VS STUDENT RETENTION STATUS



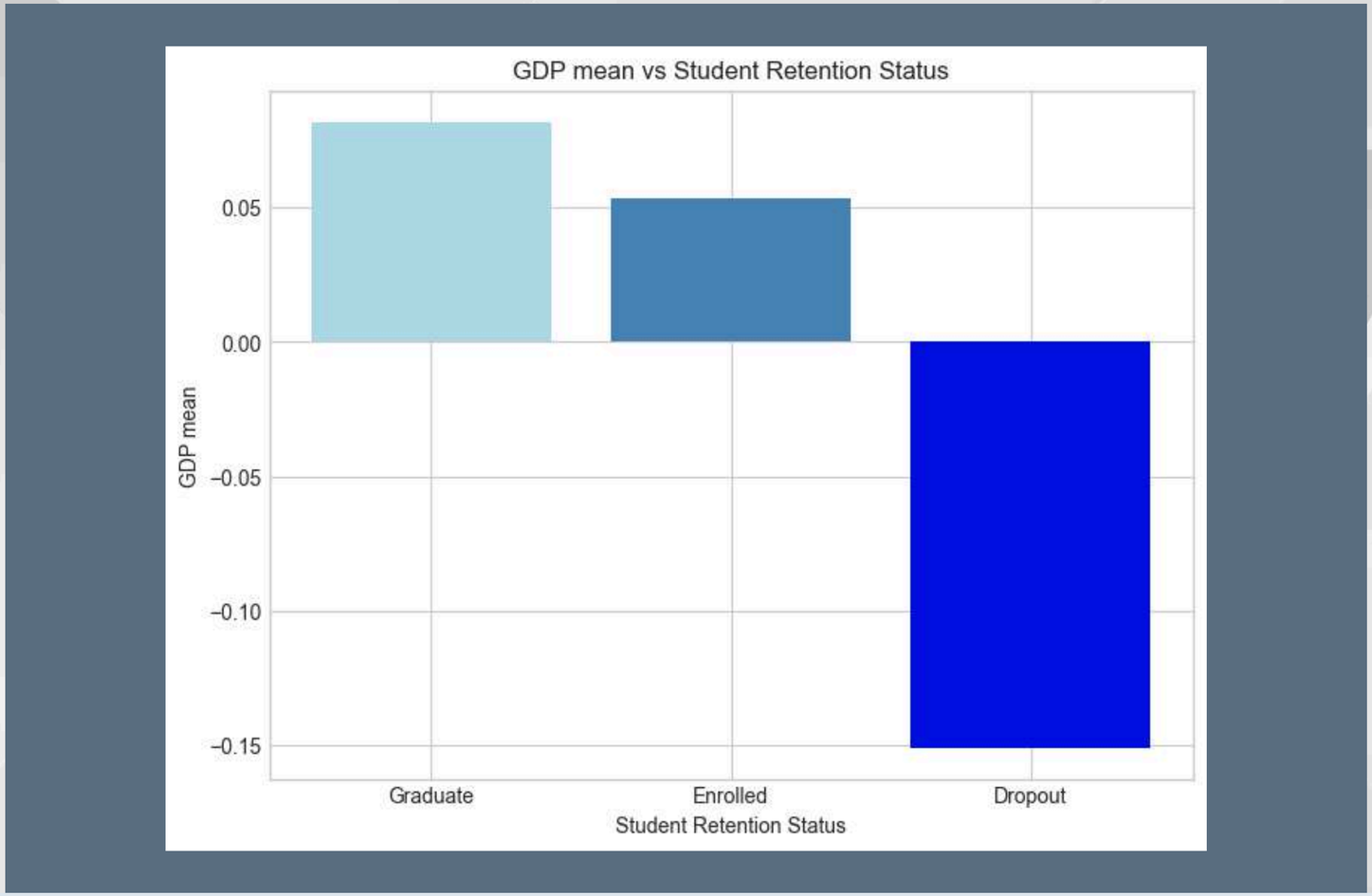

Dari grafik ini memberi gambaran untuk hubungan antara student retention status dan unemployemnt rate, terlihat bahwa mean unemployment rate untuk murid yang graduate dan yang dropout hampir sama, sementara tingkat unemployment rate untuk yang masih enrolled sedikit lebih rendah. Hal ini menunjukkan bahwa tidak ada hubungan yang signifikan antara student retention status dan tingkat unemployment rate siswa.

# INFLATION RATE MEAN VS STUDENT RETENTION STATUS

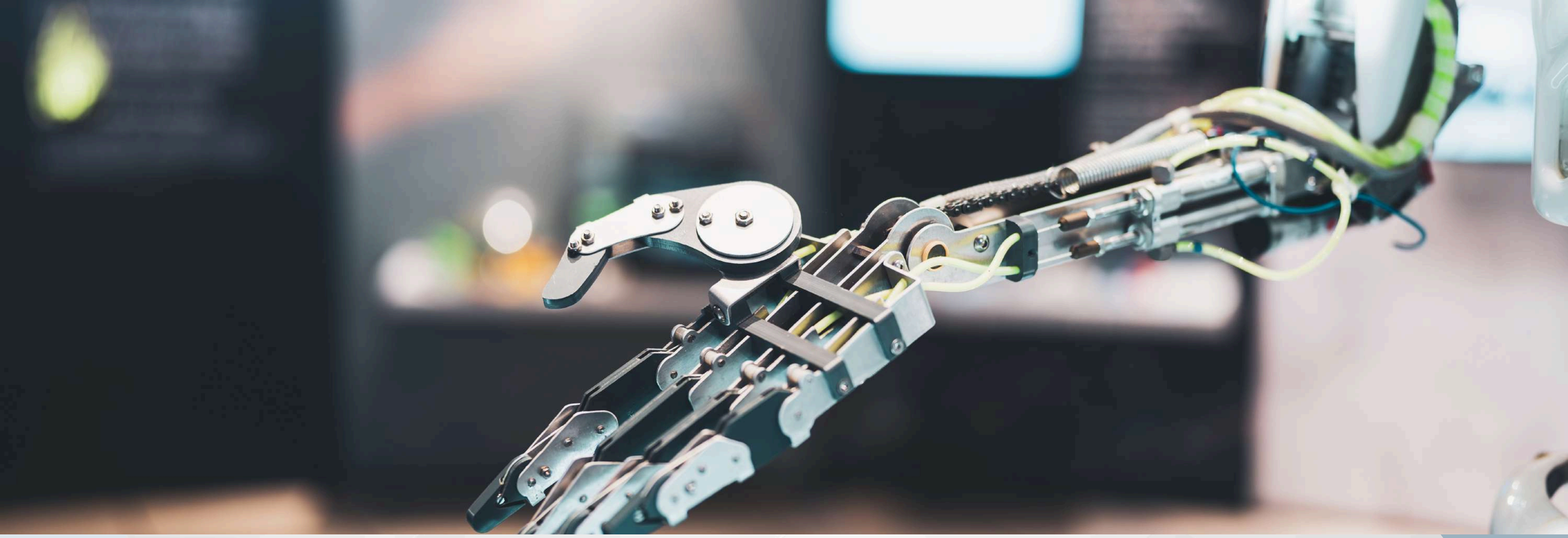



Dari grafik ini memberi gambaran untuk hubungan antara student retention status dan inflation rate, terlihat bahwa mean inflation rate untuk murid yang lulus dan yang enroleed hampir sama dengan yang enrolled lebih tinggi sedikit, sementara tingkat inflation rate untuk yang dropout sedikit lebih tinggi. Grafik ini menunjukkan bahwa tidak ada hubungan yang signifikan antara student retention status kelulusan dan tingkat inflation rate siswa.

# GPD MEAN VS STUDENT RETENTION STATUS




Dari grafik ini memberi gambaran untuk hubungan antara student retention status dan GDP, terlihat bahwa mean GDP untuk murid yang graduate paling tinggi, sementara tingkat GDP untuk yang dropout paling rendah. Ini bisa berarti bahwa murid yang graduate memiliki standar hidup yang lebih tinggi disbanding yang dropout atau enrolled. Grafik ini menunjukkan bahwa ada hubungan yang signifikan antara student retention status kelulusan dan tingkat GDP siswa.



# IV. MODELING
## (PREDICTIVE)

# DATA YANG DIAMBIL UNTUK MODELING

Dari 23 kolom yang diambil untuk Visualisasi, kami kemudian mengambil 12 kolom dari 23 kolom tersebut yang memiliki korelasi terhadap target. 12 Kolomnya yaitu

1. GDP
2. Inflation rate
3. Unemploymen rate
4. Age
5. Gender
6. Scholarship holder
7. Tuition Fees Up to date
8. Debtor
9. attendance
10. course
11. Marital Status

```
#Encode all the object type column

encoder = OneHotEncoder()
df_encode = encoder.fit_transform(df[['Marital status', 'Dayt
df_encode = pd.DataFrame(df_encode.toarray())

df_encode
```

✓ 0.0s

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ... | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | ... | 0.0 | 0.0 | 0.0 |
| 1 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | ... | 0.0 | 0.0 | 0.0 |
| 2 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | ... | 0.0 | 0.0 | 0.0 |
| 3 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | ... | 0.0 | 1.0 | 0.0 |
| 4 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 4419 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | ... | 0.0 | 1.0 | 0.0 |
| 4420 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | ... | 0.0 | 1.0 | 0.0 |
| 4421 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 |
| 4422 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | ... | 0.0 | 0.0 | 1.0 |
| 4423 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | ... | 0.0 | 1.0 | 0.0 |

4424 rows × 27 columns

# ENCODING

Setelah menentukan data mana yang menjadi bahan untuk model machine learning kita.  Selanjutnya, kita encoding data data object kita menjadi angka menggunakan one Hot Encoder agar dapat kebaca oleh model yang akan kita buat nantinya.

# SPLITING DATA

Setelah Encoding dilakukan selanjutnya kami melakukan spliting data menjadi data untuk training dan data untuk validation. kami membaginya sebesar 80% untuk training dan 20% untuk validation

```
#spliting the data

df_train, df_val = train_test_split(df_merge, test_size= 0.2, random_state=99)

x_train = df_train.drop(['Target'], axis = 1).copy()
y_train = df_train['Target'].copy()
x_val = df_val.drop(['Target'], axis = 1).copy()
y_val = df_val['Target'].copy()
```

0.0s

# MODELING

```
#modeling (default)

randomForest = RandomForestClassifier()
randomForest.fit(x_train, y_train)
print(randomForest.score(x_train,y_trai
print(randomForest.score(x_val, y_val))
```

0.4s

```
#modeling with K fold cross validation
x = df_merge.drop(['Target'], axis = 1).copy()
y = df_merge['Target'].copy()

best_model = grid_search.best_estimator_
score = cross_val_score(best_model, x, y, cv=10, scoring= 'accu
y_pred = cross_val_predict(best_model, x, y, cv=10)
print(score)
print(score.mean())
print(f1_score(y, y_pred, average='weighted'))
```

0.0s

**Default**

**K Fold Cross Validation**

Setelah Spliting Data, kami melakukan modeling,  kami membuat 2 model yang satu menggunakan metode default dengan spliting data biasa dan satunya lagi menggunakan metode k-Fold Cross Validation dimana data yang menjadi Training dan validation bersifat dinamis atau berubah ubah

# MODEL (DEFAULT)

**Menggunakan Model metode Default diperoleh:**

Training Accuracy : 88.78%

Validation Accuracy : 62.03%

F1 Score : 60.70%

# MODEL (K FOLD CROSS VALIDATION)

**Menggunakan Model metode Cross Validation Score diperoleh:**

Overall Accuracy (cross-validation score) : 64.13%

Overall F1 Score (cross-validation score) : 59.64%

**Menggunakan Model metode One-vs-All metrics diperoleh:**

Class-Specific Accuracy (One-vs-All metrics): 76.14%

Class-Specific Accuracy (One-vs-All metrics) : 49.93%

**Dengan Confusion Table Sebagai Berikut:**

| | Class | TP | FP | TN | FN | Accuracy | F1 Score |
|---|---|---|---|---|---|---|---|
| 0 | Dropout | 870 | 502 | 2501 | 551 | 0.761980 | 0.622986 |
| 1 | Enrolled | 60 | 92 | 3538 | 734 | 0.813291 | 0.126850 |
| 2 | Graduate | 1911 | 989 | 1226 | 298 | 0.709087 | 0.748092 |

# TERIMA KASIH

Kelompok 6 ~~~~~~