Analisis Prediksi Keberhasilan Akademik Mahasiswa Menggunakan Pendekatan Klasifikasi Random Forest

Vanny Khairunnisaa

2206051506

Dosen Pengampu

Sarini Abdullah, S.Si, M.Stats., Ph.D.

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS INDONESIA

2024

Ringkasan

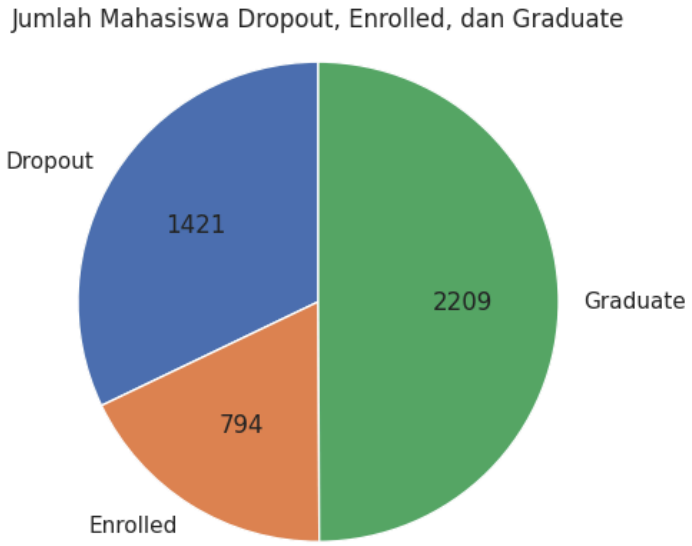
Keberhasilan akademik mahasiswa merupakan salah satu tujuan utama institusi pendidikan tinggi yang mencerminkan efektivitas proses pembelajaran dan kualitas pendukung akademik. Risiko *drop out* merupakan tantangan yang memengaruhi keberhasilan akademik. Dalam penelitian ini, algoritma *Random Forest* digunakan untuk memprediksi keberhasilan akademik mahasiswa dengan mengidentifikasi faktor-faktor yang berkontribusi terhadap risiko *drop out*. Dataset yang digunakan mencakup informasi demografis, sosial-ekonomi, serta jalur akademik mahasiswa yang berasal dari berbagai program studi.

Hasil penelitian menunjukkan bahwa klasifikasi menggunakan algoritma *Random Forest* dapat memformulasikan prediksi dalam tiga kategori keberhasilan akademik mahasiswa, yaitu *drop out*, *enrolled*, dan *graduated*. Dengan tingkat akurasi mencapai 80%, ditemukan bahwa variabel yang berkaitan langsung dengan performa akademik mahasiswa adalah variabel yang berhubungan dengan nilai akademik, seperti unit kurikulum yang disetujui, nilai, dan jumlah SKS, merupakan faktor yang paling signifikan dalam menentukan keberhasilan akademik serta risiko *drop out*. Penelitian ini melibatkan proses *preprocessing*, EDA, dan pengaplikasian beberapa algoritma klasifikasi yang kemudian dipilih model terbaik dengan algoritma *Random Forest*. Model terbaik ini memiliki parameter optimal, yaitu 50 pohon keputusan, 2 sampel minimum, dan tanpa batasan kedalaman maksimum pada setiap pohon.

Berdasarkan hasil penelitian ini, disarankan agar institusi pendidikan tinggi memberikan dukungan tambahan bagi mahasiswa dengan performa akademik rendah, seperti bimbingan intensif atau kelas remedial, untuk membantu mereka memahami materi kuliah dengan lebih baik. Selain itu, model ini juga mengungkapkan bahwa nilai akademik mahasiswa tidak secara langsung dipengaruhi oleh faktor usia maupun keadaan mahasiswa, sehingga fokus intervensi dapat diarahkan pada aspek akademik yang lebih relevan.

I. LATAR BELAKANG

Tingkat keberhasilan akademik mahasiswa merupakan salah satu indikator penting yang mencerminkan kualitas pendidikan tinggi. Pencapaian ini tidak hanya menjadi tolak ukur keberhasilan institusi pendidikan, tetapi juga menggambarkan kesiapan mahasiswa dalam menghadapi tantangan di dunia profesional. Namun, salah satu masalah yang sering dihadapi institusi pendidikan tinggi adalah tingginya risiko mahasiswa mengalami *drop out*. Tingkat *drop out* yang tinggi tidak hanya berdampak pada mahasiswa secara individu, tetapi juga pada reputasi institusi dan efisiensi penggunaan sumber daya pendidikan.



Berbagai faktor diketahui dapat memengaruhi keberhasilan akademik mahasiswa, termasuk latar belakang demografis, kondisi sosial-ekonomi, dan jalur akademik yang mereka tempuh. Oleh karena itu, institusi pendidikan perlu memahami faktor-faktor ini secara menyeluruh untuk dapat mengidentifikasi mahasiswa yang berisiko drop out lebih awal. Pemanfaatan teknologi *machine learning*, seperti algoritma *Random Forest* menjadi pendekatan yang potensial untuk memprediksi risiko tersebut dan memberikan rekomendasi serta saran yang dibutuhkan oleh institusi pendidikan. Dengan menggunakan pendekatan ini, institusi pendidikan diharapkan mampu mengembangkan strategi berbasis data untuk mendukung keberhasilan akademik mahasiswa serta mengurangi risiko *drop out*.

II. DATA DAN METODE

2.1. Data

Penelitian ini menggunakan dataset kuantitatif dan kualitatif berjudul “Predict Students’ Dropout and Academic Success” dimana dataset dibuat berdasarkan institusi pendidikan tinggi yang berkaitan dengan mahasiswa yang terdaftar dalam berbagai program studi sarjana, seperti agronomi, desain, pendidikan, keperawatan, jurnalistik, manajemen, layanan sosial, dan teknologi. Dataset ini mencakup informasi yang diketahui pada saat pendaftaran mahasiswa seperti jalur akademik, demografi, dan faktor sosial-ekonomi serta performa akademik mahasiswa pada akhir semester pertama dan kedua. Data ini digunakan untuk membangun model klasifikasi guna memprediksi dropout dan keberhasilan akademik mahasiswa. Masalah ini diformulasikan sebagai tugas klasifikasi dengan tiga kategori, di mana terdapat ketidakseimbangan yang signifikan pada salah satu kelas.

Berikut adalah cuplikan dari beberapa baris dataset

	Marital Status	Application mode	Application order	Course	Daytime/evening attendance	Previous qualification	Previous qualification (grade)	Nacionality	Mother's qualification	Father's qualification	Mother's occupation	Father's occupation	Admission grade	Displaced
0	1	17	5	171	1	1	122.0	1	19	12	5	9	127.3	1
1	1	15	1	9254	1	1	160.0	1	1	3	3	3	142.5	1
2	1	1	5	9070	1	1	122.0	1	37	37	9	9	124.8	1
3	1	17	2	9773	1	1	122.0	1	38	37	5	3	119.6	1
4	2	39	1	8014	0	1	100.0	1	37	38	9	9	141.5	0

Tuition fees up to date	Gender	Scholarship holder	Age at enrollment	International	Curricular units 1st sem (credited)	Curricular units 1st sem (enrolled)	Curricular units 1st sem (evaluations)	Curricular units 1st sem (approved)	Curricular units 1st sem (grade)	Curricular units 1st sem (without evaluations)	Curricular units 2nd sem (credited)	Curricular units 2nd sem (enrolled)	Curricular units 2nd sem (evaluations)	Curricular units 2nd sem (approved)	Curricular units 2nd sem (grade)
1	1	0	20	0	0	0	0	0	0.000000	0	0	0	0	0	0.000000
0	1	0	19	0	0	6	6	6	14.000000	0	0	6	6	6	13.666667
0	1	0	19	0	0	6	0	0	0.000000	0	0	6	0	0	0.000000
1	0	0	20	0	0	6	8	6	13.428571	0	0	6	10	5	12.400000
1	0	0	45	0	0	6	9	5	12.333333	0	0	6	6	6	13.000000

Curricular units 2nd sem (without evaluations)	Unemployment rate	Inflation rate	GDP	Target
0	10.8	1.4	1.74	Dropout
0	13.9	-0.3	0.79	Graduate
0	10.8	1.4	1.74	Dropout
0	9.4	-0.8	-3.12	Graduate
0	13.9	-0.3	0.79	Graduate

Dataset yang digunakan berisikan 37 kolom yang terdiri atas 19 kolom kategori dan 18 kolom numerik. Kolom-kolom ini merupakan kolom variabel yang menjelaskan variabel target. Berikut adalah penjelasan dari setiap kolom variabel yang ada

Curricular units 2nd sem (approved)	Banyak kurikulum pada semester kedua
Curricular units 2nd sem (grade)	Rata-rata nilai pada semester kedua
Curricular units 2nd sem (without evaluations)	Banyak kurikulum tanpa evaluasi pada semester 1

Educational special	1 (yes), 0 (no)
Debtor	1 (yes), 0 (no)
Tuition fees up to date	1 (yes), 0 (no)

Unemployment rate	Tingkat pengangguran
GDP	GDP
Target	Variabel target berupa kategori keberhasilan akademik (*drop out*, *enrolled*, *graduated*)
International	1 (yes) 0 (no)
Curricular units 1st sem (credited)	Banyak kurikulum kredit semester 1
Curricular units 1st sem (enrolled)	Banyak kurikulum enrolled semester 1
Curricular units 1st sem (evaluations)	Banyak kurikulum evaluasi semester 1
Curricular units 1st sem (approved)	Banyak kurikulum yang disetujui semester 1
Curricular units 1st sem (grade)	Rata-rata nilai semester 1
Curricular units 1st sem (without evaluations)	Banyak kurikulum tanpa evaluasi semester 1
Curricular units 2nd sem (credited)	Banyak kurikulum kredit semester 2
Curricular units 2nd sem (enrolled)	Banyak kurikulum enrolled semester 2
Curricular units 2nd sem (evaluations)	Banyak kurikulum evaluasi semester 2
Mother Occupation	Pekerjaan Ibu
Father Occupation	Pekerjaan Ayah

Gender	1 (yes), 0 (no)
Scholarship holder	1 (yes), 0 (no)
Age at enrollment	Umur saat masuk
Marital status	Status menikah
Application mode	Jalur masuk
Application order	Penerimaan
Course	Program Studi
Daytime/evening attendance	1 (daytime), 0 (evening)
Previous qualification	Kualifikasi terakhir
Previous qualification (grade)	Nilai terakhir
Nationality	Kebangsaan
Mother qualification	Kualifikasi Ibu
Father qualification	Kualifikasi Ayah
Admission grade	Nilai saat penerimaan
Displaced	1 (yes) , 0 (no)

2.2. Metode Random Forest

Random Forest (RF) pertama kali diperkenalkan oleh Leo Breiman (2001). RF merupakan salah satu metode yang dapat meningkatkan hasil akurasi dalam membangkitkan atribut untuk setiap node yang dilakukan secara acak. RF terdiri dari sekumpulan *decision tree*, di mana kumpulan pohon keputusan ini digunakan untuk mengklasifikasi data ke suatu kelas. Pohon keputusan dibuat dengan

menentukan *node* akar dan berakhir dengan beberapa *node* daun untuk mendapatkan hasil akhir (Debby &Rahman, 2020).

Membentuk pohon keputusan pada metode RF sama dengan proses pada *Classification and Regression Tree* (CART), hanya saja pada RF tidak dilakukan *pruning* (pemangkasan). Indeks Gini digunakan untuk memilih fitur di setiap simpul internal dari pohon keputusan. Nilai Indeks Gini dapat dihitung sebagai berikut:

$$\text{Gini}(S_i) = 1 - \sum_{i=0}^{c-1} p_i^2$$

dengan *pi* merupakan frekuensi relatif kelas *Ci* di dalam set.

Ci merupakan kelas untuk *i* = 1, ...,c-1, dan *c* adalah jumlah kelas yang telah ditentukan.

Kualitas *split* pada fitur *k* kedalam subset *Si* merupakan jumlah sampel milik kelas *Ci*, kemudian dihitung sebagai jumlah pertimbangan indikasi Gini dari subset yang dihasilkan. Data dapat dihitung dengan rumus sebagai berikut:

$$\text{Gini}_{split} = \sum_{i=0}^{k-1} \left(\frac{n_i}{n}\right) \text{Gini}(S_i)$$

dimana *ni* merupakan jumlah sampel dalam subset *Si* setelah di split dan *n* merupakan jumlah sampel di *node* yang diberikan.

Misalkan {*h(x, Θk)*, *k = 1*, ...} dimana {*Θk*} merupakan vektor *random* ayang *independent identically distributed* (iid) dan tiap pohon memilih kelas yang paling banyak dari rata-rata (*majority vote*). Untuk RF, batas atas dapat diturunkan untuk kesalahan generalisasi dalam hal dua parameter yang mengukur seberapa kuat pengklasifikasian individu dan ketergantungan diantara keduanya (Breiman, 2001):

Fungsi margin untuk RF adalah

$$\text{mr}(X, Y) = P_\Theta(h(X, \Theta) = Y) - \max_{j \neq Y} P_\Theta(h(X, \Theta) = j)$$

dan kekuatan himpunan pengklasifikasi {h(X,Θ)}adalah

$$s = E_{XY} mr(X, Y)$$

Dengan asumsi s ≥ 0, ketidaksamaan Chebyshev serta penurunan variansi *mr* dari fungsi margin untuk metode RF, akan didapatkan persamaan batas atas kesalahan generalisasi sebagai berikut:

$$PE \leq \frac{\bar{p}(1-s^2)}{s^2}$$

Dimana $\bar{p}$ adalah nilai rata-rata korelasi, yaitu:

$$\bar{\rho} = \frac{E_{\theta,\theta'}\left(\rho(\theta,\theta')sd(\theta)sd(\theta')\right)}{E_{\theta,\theta'}\left(sd(\theta)sd(\theta')\right)}$$

2.3. Confusion Matrix

Confusion matrix adalah tabel yang digunakan untuk melihat akurasi serta seberapa baik algoritma yang dihasilkan dari klasifikasi yang sudah dibuat untuk mengklasifikasi dan memprediksi atribut dari *data testing*. Metode ini dikembangkan sebagai penilaian algoritma *machine learning* yang diterapkan dalam menyelesaikan masalah klasifikasi. Dalam *confusion matrix* terdapat *False Negative* (FN), *False Positive* (FP), *True Negative* (TN), dan *True Positive* (TP). Berikut merupakan tabel dari *confusion matrix*.

Kelas Prediksi	Kelas Aktual	
	Positive	*Negative*
Positive	TP	FP
Negative	FN	TN

TP adalah kondisi dimana baik prediksi maupun nilai aktualnya benar; FN adalah kasus dimana nilai prediksi tidak benar tetapi nilai aktualnya benar; FP adalah kasus dimana nilai prediksi benar tapi nilai aktualnya tidak benar. Dalam mengevaluasi kinerja model, ada berbagai macam performa diantaranya akurasi, *recall*, dan presisi. Nilai akurasi, *recall*, dan presisi dapat diperoleh dengan menggunakan persamaan berikut.

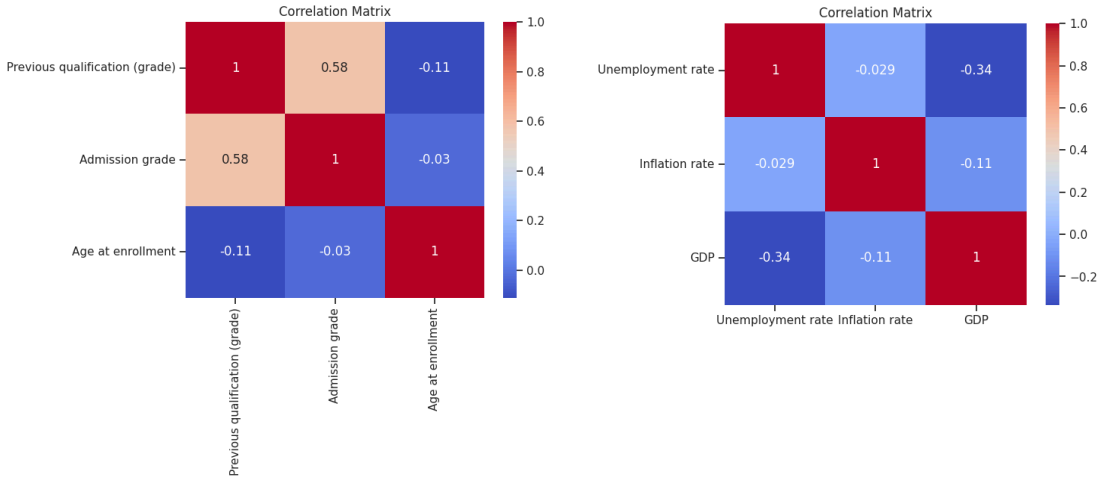
Performance Metrics	Rumus
Akurasi	$\frac{TP+TN}{TP+TN+FP+FN} \times 100\%$
Recall	$\frac{TP}{TP+FN} \times 100\%$
Presisi	$\frac{TP}{TP+FP} \times 100\%$

Akurasi merupakan rasio prediksi benar dengan keseluruhan data. *Recall* adalah rasio prediksi benar positif dibandingkan dengan keseluruhan data aktual positif. Presisi merupakan rasio prediksi benar positif dibandingkan dengan seluruh data yang diprediksi positif.

III. HASIL PENGOLAHAN DAN ANALISIS DATA

3.1. Eksplorasi dan Visualisasi Data

3.1.1 Correlation Matrix

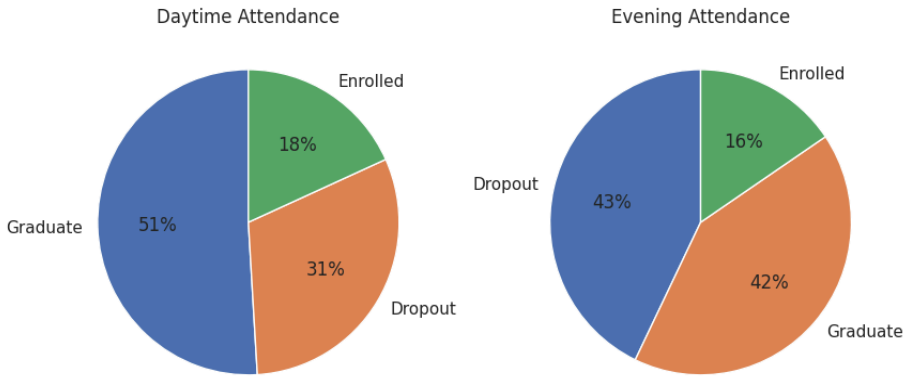


Correlation Matrix (Kiri: Age & Grade, Kanan: Wilayah Ekonomi)

Berdasarkan hasil analisis *heatmap*, ditemukan beberapa hubungan menarik antara variabel-variabel yang dianalisis. Pada kelompok variabel *Age*, *Qualification Grade*, dan *Admission Grade*, terdapat hubungan linier positif yang sedang antara nilai kualifikasi dan nilai pendaftaran (0.58). Hal ini menunjukkan bahwa siswa dengan nilai kualifikasi yang lebih tinggi cenderung memiliki nilai pendaftaran yang juga lebih tinggi, meskipun hubungan ini tidak berlaku untuk seluruh siswa. Namun, hubungan antara umur dengan nilai kualifikasi maupun nilai pendaftaran cenderung tidak signifikan, sehingga dapat disimpulkan bahwa umur tidak memengaruhi nilai akademik.

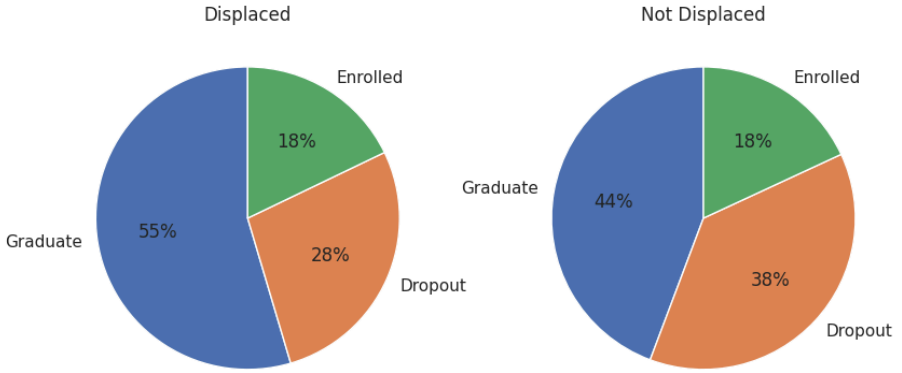
Pada kelompok variabel Wilayah Ekonomi, ditemukan hubungan linier negatif yang lemah hingga sedang antara GDP dan rate pengangguran (-0.34). Temuan ini mengindikasikan bahwa ketika tingkat pengangguran menurun di suatu wilayah, GDP cenderung meningkat. Di sisi lain, hubungan antara GDP dan *rate* inflasi sangat lemah (-0.11), menunjukkan bahwa GDP hampir tidak memengaruhi tingkat inflasi, begitu pula sebaliknya. Selain itu, hubungan antara *rate* inflasi dan *rate* pengangguran hampir tidak ada (-0.029), yang menyiratkan bahwa kedua variabel ini tidak memiliki keterkaitan yang signifikan.

3.1.2 Kehadiran Kelas



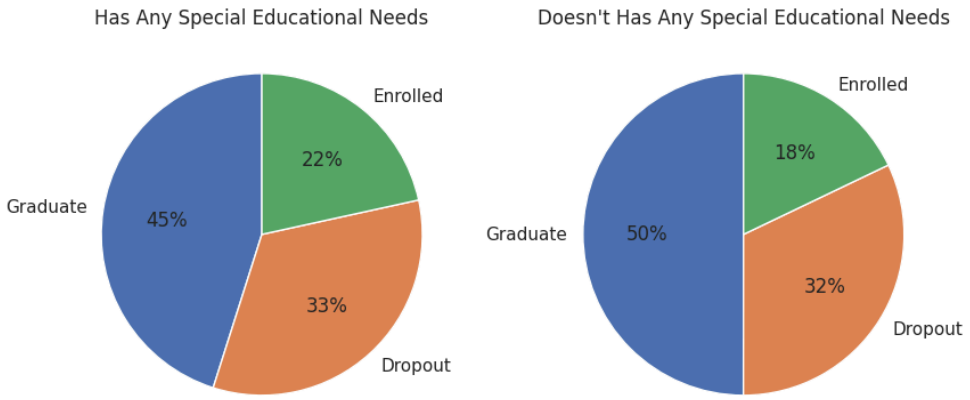
Mahasiswa yang menghadiri *daytime class* memiliki proporsi *graduate* yang lebih besar dibandingkan dengan mahasiswa yang menghadiri *evening class*. Sebaliknya, mahasiswa yang menghadiri *evening class* memiliki proporsi *drop out* yang lebih besar dibandingkan dengan *daytime class*. Hal ini menunjukkan bahwa mahasiswa yang mengambil *evening class* lebih berisiko mengalami *drop out* dibandingkan dengan mereka yang mengikuti *daytime class*.

3.1.3 Displaced Situation



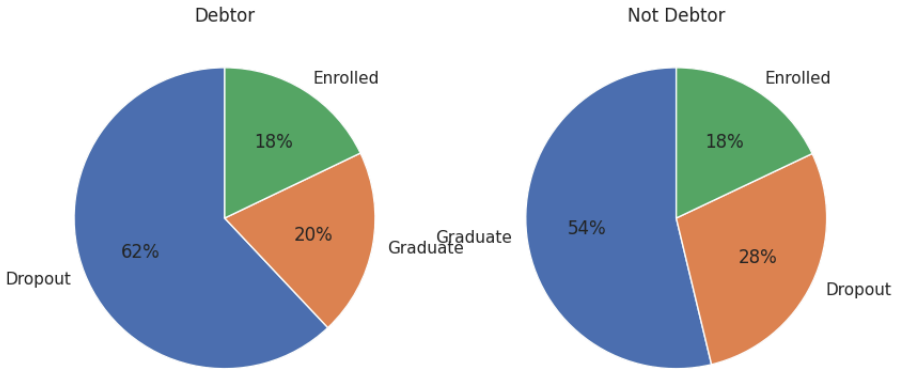
Displaced merujuk pada individu yang terpaksa meninggalkan rumah atau tanah air mereka karena alasan seperti perang, kelaparan, bencana alam, atau faktor lainnya, tetapi tetap berada di dalam negara tersebut. Berdasarkan analisis, status sebagai *displaced person* tidak memengaruhi dominasi proporsi mahasiswa yang lulus, karena baik mahasiswa dengan status *displaced* maupun *non-displaced* menunjukkan proporsi kelulusan yang dominan. Namun, pada mahasiswa yang tidak memiliki status *displaced*, terdapat proporsi *drop out* yang lebih besar dibandingkan dengan mahasiswa yang berstatus *displaced*.

3.1.4 Special Needs



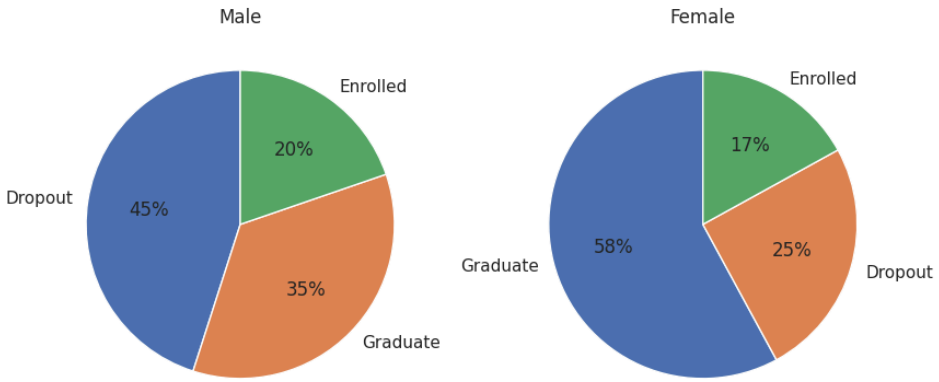
Menurut *pie chart* kebutuhan khusus didapatkan baik mahasiswa berkebutuhan/tidak memiliki proporsi kelulusan sebagai status yang dominan. Berarti perbedaan keberhasilan akademik tidak dipengaruhi oleh status berkebutuhan khusus/tidaknya seseorang.

3.1.5 Hutang



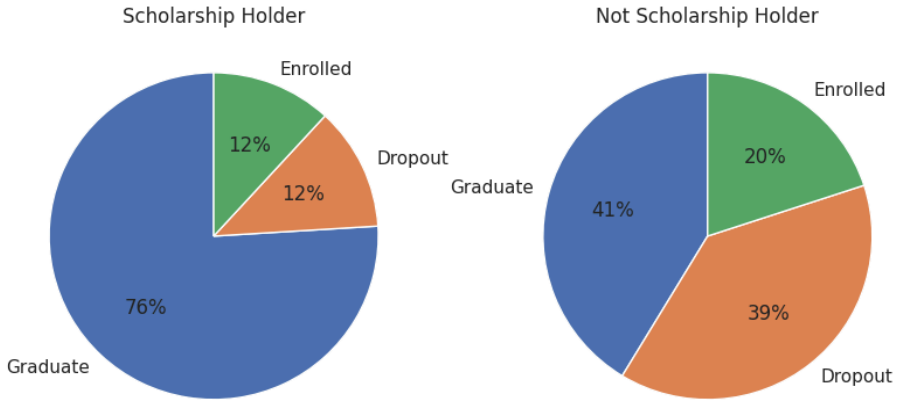
Mahasiswa dengan hutang memiliki proporsi *drop out* yang lebih tinggi, yang menunjukkan bahwa keberadaan hutang memengaruhi keberhasilan mahasiswa untuk lulus. Sebaliknya, mahasiswa tanpa hutang memiliki tingkat proporsi kelulusan yang lebih tinggi dibandingkan dengan mereka yang memiliki hutang. Hal ini mengindikasikan bahwa hutang berpengaruh signifikan terhadap keberhasilan akademik mahasiswa.

3.1.6 Jenis Kelamin



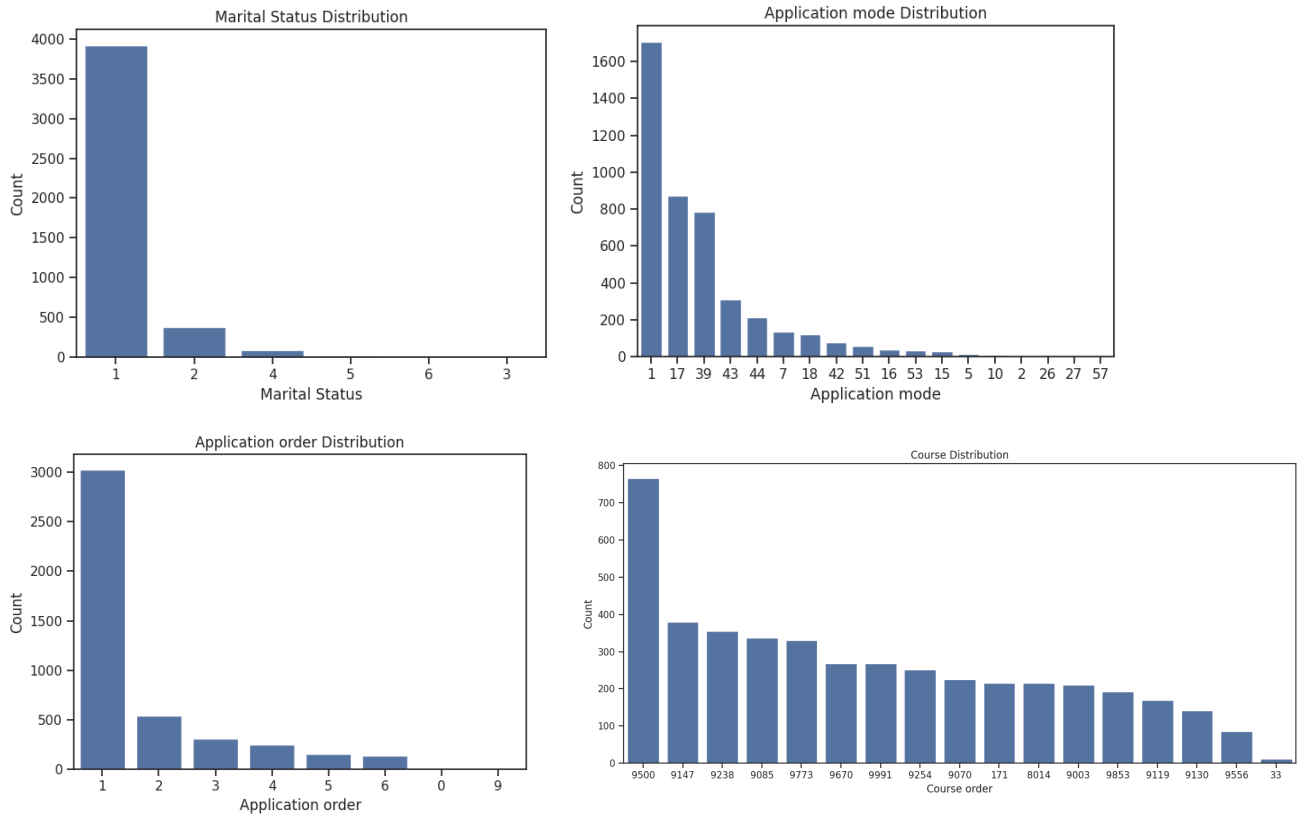
Mahasiswa laki-laki memiliki proporsi *drop out* yang lebih tinggi, yang menunjukkan bahwa mereka memiliki risiko yang lebih besar untuk mengalami dropout dibandingkan dengan mahasiswa perempuan. Sebaliknya, mahasiswa perempuan memiliki tingkat proporsi kelulusan yang lebih tinggi, yang berarti risiko *drop out* pada mahasiswa perempuan lebih kecil. Hal ini mengindikasikan bahwa gender berpengaruh terhadap keberhasilan akademik mahasiswa.

3.1.7 Beasiswa



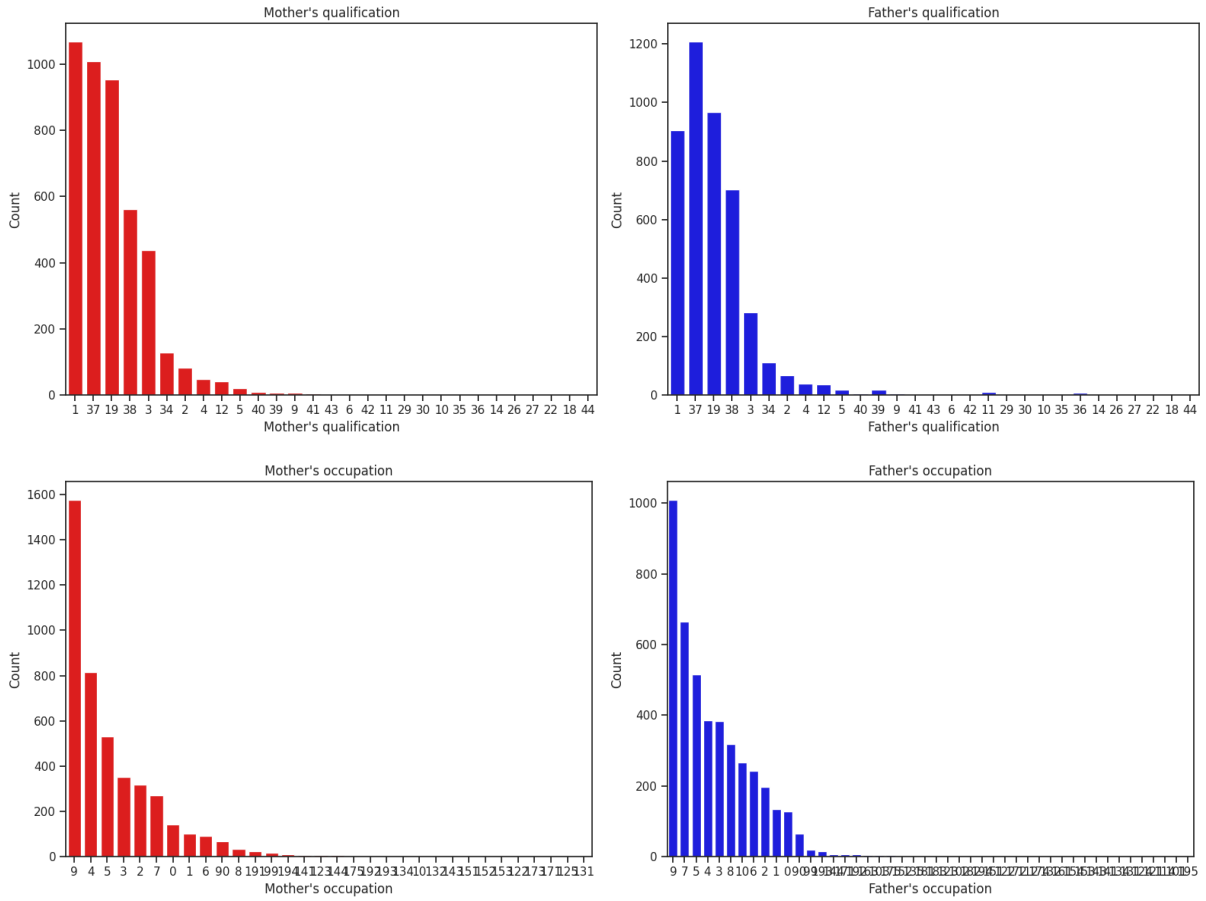
Mahasiswa dengan beasiswa memiliki proporsi *graduate* yang lebih tinggi dibandingkan dengan mahasiswa tanpa beasiswa. Sebaliknya, mahasiswa tanpa beasiswa memiliki proporsi *drop out* yang lebih besar dibandingkan dengan mahasiswa penerima beasiswa. Hal ini menunjukkan bahwa mahasiswa dengan beasiswa memiliki risiko *drop out* yang lebih kecil. Hal ini dapat dimengerti karena untuk memperoleh beasiswa diperlukan kemampuan akademis yang mumpuni, sehingga mahasiswa penerima beasiswa cenderung memiliki performa akademik yang lebih baik.

3.1.8 Status, Jurusan, dan Pendaftaran Mahasiswa



Mayoritas mahasiswa memiliki status belum menikah (1), meskipun terdapat mahasiswa dengan status menikah (2) dan sudah pernah menikah (4). Dari segi jalur pendaftaran, mahasiswa yang diterima didominasi oleh mereka yang lulus seleksi pertama (1), diikuti oleh seleksi kedua (17), jalur khusus untuk usia di atas 23 tahun (39), mahasiswa yang pindah jurusan (43), dan mahasiswa dari program diploma (44). Mahasiswa yang diterima melalui pendaftaran pertama mencapai jumlah kurang lebih 3.000, sedangkan pada pendaftaran kedua jumlahnya menurun signifikan hingga sekitar 500 mahasiswa, atau hanya sekitar 1/6 dari jumlah pendaftar pertama. Berdasarkan jurusan, mahasiswa dari program keperawatan (*nursing*) memiliki jumlah terbanyak, yaitu lebih dari 750 mahasiswa, sementara jurusan dengan jumlah mahasiswa paling sedikit adalah *biofuel*. Selain itu, terdapat perbedaan kualifikasi sebelumnya yang signifikan, di mana mayoritas mahasiswa memiliki pendidikan menengah (*secondary education*, 1), dengan jumlah yang jauh lebih banyak dibandingkan mahasiswa dengan spesialisasi teknis (*tech specialization course*, 39), pendidikan dasar (*basic education*, 19), dan pendidikan tinggi (*higher education*, 3).

3.1.9 Faktor Orang Tua



Kualifikasi ibu didominasi oleh s*econdary education* (1), diikuti oleh *basic education* (37 dan 19). Hanya sebagian kecil ibu yang memiliki tingkat pendidikan tinggi, seperti *bachelor* (2) dan doktor (5). Hal serupa juga terjadi pada kualifikasi ayah, yang didominasi oleh *secondary education* (1) dan diikuti oleh *basic education* (37 dan 19). Namun, proporsi pendidikan ayah cenderung sedikit lebih besar dibandingkan ibu pada tingkat pendidikan tinggi.

Untuk pekerjaan, mayoritas ibu bekerja sebagai pekerja tidak terampil (9) yang tidak memerlukan keterampilan khusus, diikuti oleh pekerjaan di sektor *personal services* (5) dan administratif (4). Pekerjaan seperti *teacher* (123) dan *technicians as legal, social, sport* (134) memiliki frekuensi yang sangat rendah hingga hampir tidak ada. Sementara itu, pekerjaan ayah juga didominasi oleh pekerja tidak terampil (9), diikuti oleh sektor *personal services* (5) dan *skilled workers* (7). Namun, pekerjaan ayah cenderung lebih beragam dan melibatkan keterampilan dibandingkan ibu.

3.2. Featured Selection

Featured Selection dilakukan untuk mengetahui variabel mana saja yang berpengaruh secara signifikan dalam menjelaskan variabel target. Dilakukan *featured selection* dengan fungsi ANOVA menggunakan hipotesis berikut

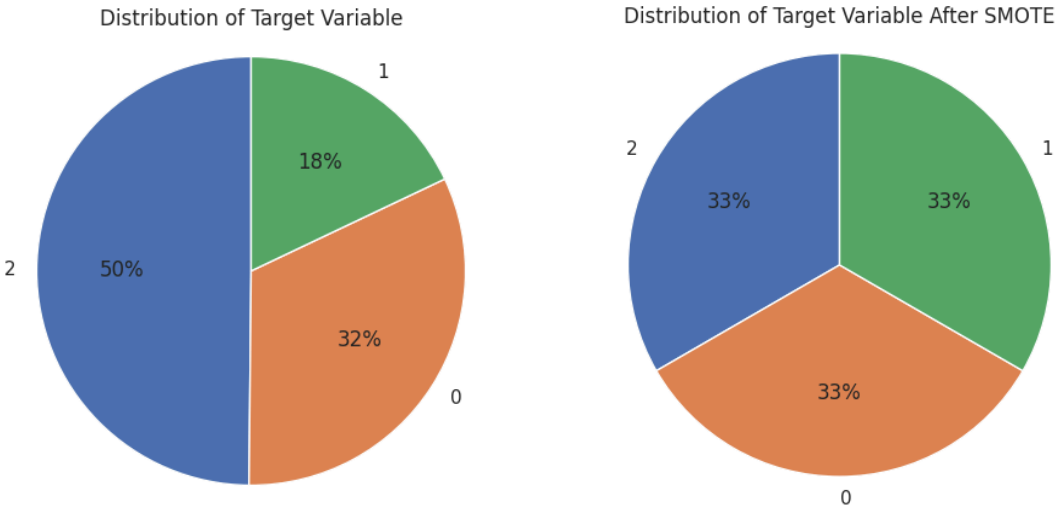
H_0 : Not Significant Features

H_1 : Significant Features

```
Non-Significant Features (p-value >= 0.05):
                       Feature     Score   p-Value
34             Inflation rate  1.741990  0.175292
3                      Course  0.796673  0.450891
20              International  0.639709  0.527494
7                 Nacionality  0.509016  0.601122
14  Educational special needs  0.320854  0.725546
```

Berdasarkan hasil *non significant features* yang tertera di atas maka variabel 'Inflation rate', 'Course', 'International', 'Nacionality', dan 'Educational special needs' tidak berpengaruh dalam menjelaskan variabel target dan akan dibuang dari *database*.

3.3. Penyelesaian Imbalanced Data



0 (*drop out),* 1 (*enrolled*), dan 2 (*graduated*)

Berdasarkan nilai distribusi variabel target didapatkan bahwa pembagian variabel target tidaklah seimbang. Untuk menyeimbangkan distribusi tersebut maka akan digunakan SMOTE agar data lebih seimbang dan tidak bias pada suatu kategori target tertentu. Setelah dilakukan SMOTE didapatkan masing-masing dataset di *resample* menjadi 1764 setiap kategori pada variabel target.

3.4. Evaluasi Model

Data akan dimodeling menggunakan model regresi logistik, *decision tree*, dan *random forest* dimana data sudah melalui proses *preprocessing*, standarisasi, dan *pre modeling* (*labeling, feature selection, data split,* dan SMOTE)

3.4.1 Regresi Logistik

```
[[200  44  37]
 [ 29  96  34]
 [ 18  65 362]]
              precision    recall  f1-score   support

           0       0.81      0.71      0.76       281
           1       0.47      0.60      0.53       159
           2       0.84      0.81      0.82       445

    accuracy                           0.74       885
   macro avg       0.70      0.71      0.70       885
weighted avg       0.76      0.74      0.75       885
```

(confusion matrix dan classification report regresi logistik)

Logistic Regression adalah model linear yang sederhana dan sering digunakan sebagai baseline dalam analisis data. Model ini bekerja dengan baik pada dataset yang memiliki hubungan linear antarfitur. Pada dataset ini, *Logistic Regression* menunjukkan performa yang stabil dengan akurasi mencapai 74% pada data uji, menjadikannya alat yang andal untuk menangkap pola *linear* dalam data, meskipun memiliki keterbatasan pada hubungan *non-linear.*

Dalam analisis per kelas, model lebih andal mengenali mahasiswa yang *Graduate* dan *Drop out* dibandingkan dengan mahasiswa yang *Enrolled*. Untuk kelas 0 (*Drop out*), model memiliki *recall* sebesar 71%, yang berarti 71% dari mahasiswa yang sebenarnya *drop out* berhasil dikenali, dan *precision* mencapai 81%, menunjukkan tingkat keandalan yang baik meskipun masih terdapat beberapa kesalahan klasifikasi (*false negatives*). Pada kelas 2 (*Graduate*), recall mencapai 81% dan *precision* 84%, mengindikasikan bahwa model sangat andal dalam memprediksi mahasiswa yang lulus. Namun, pada kelas 1 (*Enrolled*), performa model masih lemah, dengan *recall* sebesar 60% dan *precision* hanya 47%, menunjukkan bahwa lebih dari setengah prediksi untuk kelas ini salah (*false positives*).

Kelemahan model pada kelas *Enrolled* ini mengindikasikan perlunya penyesuaian lebih lanjut untuk meningkatkan *precision* dan *f1-score* pada kelas tersebut. Meskipun *Logistic Regression* sederhana, performanya yang stabil tetap menjadikannya pilihan awal yang baik untuk memahami pola *linear* dalam dataset.

3.4.2 Decision Tree

```
[[196  49  36]
 [ 47  73  39]
 [ 35  64 346]]
              precision    recall  f1-score   support

           0       0.71      0.70      0.70       281
           1       0.39      0.46      0.42       159
           2       0.82      0.78      0.80       445

    accuracy                           0.69       885
   macro avg       0.64      0.64      0.64       885
weighted avg       0.71      0.69      0.70       885
```

(confusion matrix dan classification report decision tree)

Decision Tree adalah model yang bersifat interpretatif dan dirancang untuk menangkap hubungan *non-linear* dalam data. Model ini sering digunakan bersama model *ensemble* seperti *Random Forest* karena hasilnya yang berbeda dapat memperkaya analisis. Pada dataset ini, *Decision Tree* memberikan akurasi sebesar 69%, yang lebih rendah dibandingkan *Logistic Regression* (74%), namun tetap menunjukkan kemampuan dalam menangkap pola *non-linear*.

Dalam analisis per kelas, performa *Decision Tree* menunjukkan variasi yang signifikan. Pada kelas 0 (*Drop out*), model memiliki recall sebesar 70%, yang berarti 70% dari mahasiswa yang sebenarnya *dropout* berhasil dikenali, dan *precision* sebesar 71%, menunjukkan keandalan yang cukup baik meskipun masih terdapat beberapa kesalahan klasifikasi (*false negatives*). Pada kelas 2 (*Graduate*), model menunjukkan performa terbaik dengan recall sebesar 78% dan precision sebesar 82%, mengindikasikan bahwa sebagian besar mahasiswa *graduate* dapat dikenali dengan benar dan prediksi untuk kelas ini sangat andal. Namun, performa model pada kelas 1 (*Enrolled*) masih lemah, dengan *recall* hanya 46% dan *precision* 39%, yang menunjukkan bahwa banyak prediksi untuk kelas ini yang salah (*false positives*).

Secara keseluruhan, *Decision Tree* memiliki performa yang lebih baik dalam mengenali mahasiswa *Graduate* dan *Dropout* dibandingkan dengan mahasiswa *Enrolled*. Meskipun akurasi dan stabilitasnya lebih rendah dibandingkan *Logistic Regression*, *Decision Tree* tetap memberikan wawasan tambahan yang berharga, terutama dalam pola *non-linear*. Penyesuaian lebih lanjut pada model diperlukan untuk meningkatkan performa pada kelas dengan hasil yang lemah.

3.4.3 Random Forest

```
[[204  38  39]
 [ 31  85  43]
 [ 13  42 390]]
              precision    recall  f1-score   support

           0       0.82      0.73      0.77       281
           1       0.52      0.53      0.52       159
           2       0.83      0.88      0.85       445

    accuracy                           0.77       885
   macro avg       0.72      0.71      0.72       885
weighted avg       0.77      0.77      0.77       885
```

Random Forest adalah model *non-linear* berbasis *ensemble* yang sangat kuat. Dengan menggabungkan banyak *decision tree*, *Random Forest* mampu mengurangi *overfitting* dan menangani hubungan *non-linear* dalam data dengan baik. Dalam eksperimen ini, *Random Forest* menjadi model terbaik dengan akurasi mencapai (77%) lebih tinggi dibandingkan *Logistic Regression* (74%) dan *Decision Tree* (69%).

Dalam analisis per kelas, performa *Random Forest* menunjukkan keandalan yang tinggi pada kelas 0 (*Drop out*) dan kelas 2 (*Graduate*). Untuk kelas 0 (*Dropout*), *recall* sebesar 73% menunjukkan bahwa 73% mahasiswa yang sebenarnya *dropout* berhasil dikenali, dan *precision* sebesar 82% mengindikasikan tingkat keandalan prediksi yang sangat baik, meskipun masih terdapat beberapa kesalahan klasifikasi (*false negatives*). Pada kelas 2 (*Graduate*), model menunjukkan performa terbaik dengan *recall* mencapai 88% dan *precision* sebesar 83%, menunjukkan bahwa sebagian besar mahasiswa *graduate* dapat dikenali dengan benar dan prediksi untuk kelas ini sangat andal.

Namun, pada kelas 1 (*Enrolled*), performa model masih lebih lemah dibandingkan kelas lainnya. Dengan *recall* sebesar 53%, lebih dari setengah mahasiswa yang sebenarnya *enrolled* berhasil dikenali, dan *precision* sebesar 52% menunjukkan bahwa setengah dari prediksi untuk kelas ini benar, sementara sisanya salah (*false positives*). Hal ini mengindikasikan perlunya penyesuaian model untuk meningkatkan kinerja pada kelas ini.

Secara keseluruhan, *Random Forest* menunjukkan kinerja yang unggul dalam menangkap pola *non-linear*, terutama dalam mengenali mahasiswa *Graduate* dan *Dropout*. Meskipun performa pada kelas *Enrolled* masih lemah, model ini tetap memberikan hasil yang lebih baik dibandingkan *Logistic Regression* dan *Decision Tree*, menjadikannya pilihan terbaik dalam eksperimen ini.

3.5. Best Model Random Forest

```
Summary of Results (Cross-Validation Only):
                Model  CV Mean Accuracy  CV Std Dev
0  LogisticRegression          0.755310    0.041548
1        DecisionTree          0.731886    0.058380
2        RandomForest          0.832791    0.052068
```

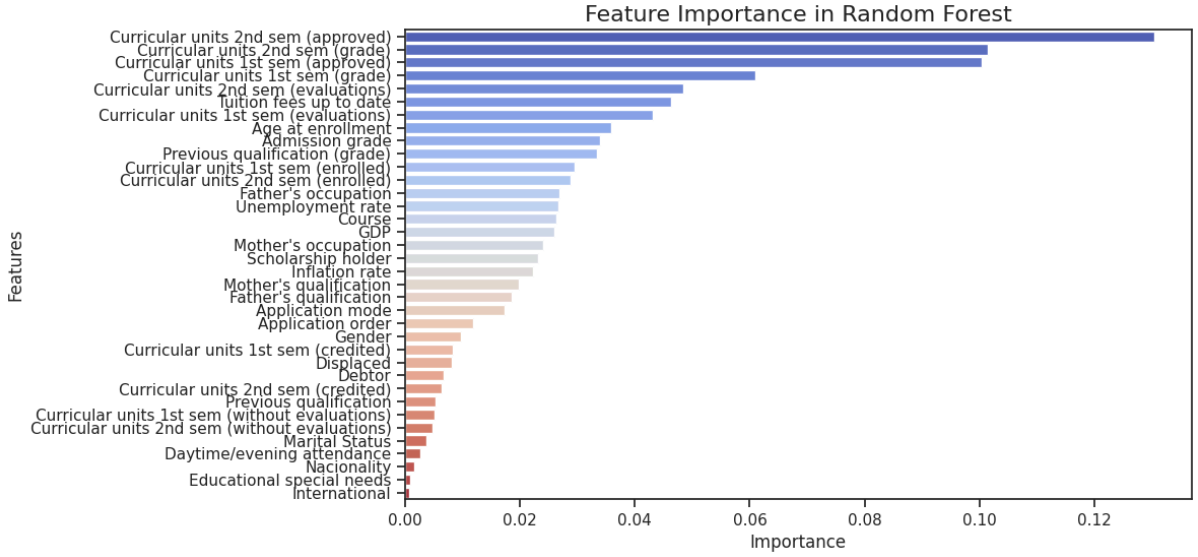
(cross validation models)

Berdasarkan performa model, akurasi, dan akurasi rata-rata didapatkan bahwa *Random Forest* merupakan model terbaik dengan variasi hasil yang stabil. Untuk mengoptimalkan performa dari model akan dilakukan *hyperparameter tuning* dengan menggunakan nilai akurasinya. Alasan dipilihnya "accuracy" dibandingkan "recall" ataupun "precision: dalam penilaian *cross validation* pada model adalah sudah dilakukannya handling SMOTE sehingga dataset sudah *balanced*. *Hyperparameter tuning* dilakukan dengan tujuan untuk mengukur akurasi dalam mengklasifikasikan kelas yang ada dan kelebihan dalam interpretasi.

```
Accuracy: 0.7672316384180791
Best Parameters: {'random_state': 13, 'n_estimators': 50, 'min_samples_split': 2, 'max_depth'
Best Score: 0.8034769463340892
Tuning time: 26.25 seconds
```

Didapatkan bahwa *Random Forest* memiliki akurasi 76.7% *data testing* benar yang di mana model dapat mempelajari pola data training dan mengklasifikasikan data dengan baik. Kombinasi *best* parameter pada model ini adalah menggunakan 50 *tree* yang akan dibagi minimal berisikan 2 sampel tanpa batasan kedalaman maksimum pada setiap pohon. Skor terbaik atau akurasi terbaik yang didapat selama *cross validation* adalah 80.34%.

Dengan pengaplikasian *Random Forest* juga dapat dilakukan *feature importance* pada model yang memberikan visualisasi mengenai variabel apa saja yang memiliki pengaruh terbesar dalam memberikan hasil keputusan.



Ditemukan bahwa faktor utama yang memengaruhi prediksi keberhasilan mahasiswa adalah performa akademik mereka, terutama jumlah unit kurikulum yang disetujui di semester pertama dan kedua. *Curricular units* 2nd sem (*approved*) muncul sebagai variabel paling berpengaruh, diikuti oleh nilai pada unit semester kedua dan jumlah unit semester pertama yang disetujui. Faktor-faktor lain seperti evaluasi unit semester kedua dan pembayaran biaya kuliah tepat waktu juga memainkan peran penting, mencerminkan tingkat komitmen mahasiswa terhadap studi mereka. Usia saat pendaftaran juga memiliki dampak moderat, yang mungkin terkait dengan kedewasaan atau kesiapan mahasiswa. Selain itu, nilai saat masuk universitas dan kualifikasi sebelumnya menjadi indikator penting, menguatkan pentingnya performa akademik masa lalu. Di sisi lain, kondisi sosial ekonomi, seperti pekerjaan orang tua, tingkat pengangguran, dan GDP, memberikan pengaruh sedang terhadap hasil prediksi, sementara faktor demografis seperti status pernikahan, kehadiran siang atau malam, serta kewarganegaraan memiliki pengaruh yang lebih kecil. Fitur dengan dampak paling rendah, seperti kebutuhan pendidikan khusus dan status internasional, hampir tidak memberikan pengaruh dalam model ini. Secara keseluruhan, prediksi keberhasilan mahasiswa dalam model ini sangat bergantung pada faktor akademik langsung, sementara kondisi sosial ekonomi dan demografi menjadi faktor pendukung yang kurang signifikan.

IV. SARAN DAN KESIMPULAN

Random Forest sebagai model terbaik memiliki akurasi mencapai 80.34% setelah penerapan *hyperparameter tuning*, melampaui *Logistic Regression* (74%) dan *Decision Tree* (69%). Dengan kemampuan menangkap pola *non-linear*, *Random Forest* memberikan hasil yang paling stabil dan andal dibandingkan model lainnya.

Hasil penelitian menunjukkan bahwa faktor utama yang memengaruhi prediksi keberhasilan mahasiswa adalah performa akademik mereka. Variabel seperti jumlah unit kurikulum yang disetujui di semester pertama dan kedua, nilai akademik, serta evaluasi unit kurikulum memiliki pengaruh paling besar. Faktor lain, seperti kepemilikan hutang memberikan risiko *drop out* bagi mahasiswa, sementara usia saat pendaftaran memberikan dampak moderat yang berkaitan dengan kedewasaan atau kesiapan mahasiswa. Nilai saat masuk universitas dan kualifikasi sebelumnya juga menjadi indikator penting keberhasilan akademik.

Di sisi lain, kondisi sosial ekonomi, seperti pekerjaan orang tua, tingkat pengangguran, dan GDP, memberikan pengaruh sedang terhadap prediksi, sedangkan faktor demografis seperti status pernikahan, kehadiran siang atau malam, serta kewarganegaraan memiliki dampak yang lebih kecil. Faktor dengan pengaruh paling rendah, seperti kebutuhan pendidikan khusus dan status internasional, hampir tidak berkontribusi pada model ini.

Penelitian ini merekomendasikan agar institusi pendidikan tinggi memberikan perhatian lebih pada variabel akademik utama untuk meningkatkan keberhasilan mahasiswa. Beberapa langkah strategis yang dapat diambil meliputi:

1. Memberikan kesempatan **remedial** atau tugas tambahan kepada mahasiswa dengan nilai rendah untuk membantu meningkatkan pemahaman mereka.
2. Menempatkan **asisten dosen** pada setiap mata kuliah untuk memberikan bimbingan tambahan, sehingga mahasiswa lebih mudah memahami materi kuliah.
3. Memberikan **pelayanan tanggap hutang** kepada mahasiswa kurang mampu dan memiliki hutang, sehingga kendala finansial tidak menjadi alasan mahasiswa *drop out*.
4. Memberikan ***reward*** kepada mahasiswa dengan nilai atau IP semester tertinggi sebagai motivasi untuk meningkatkan performa akademik.

Dengan pendekatan berbasis data dan strategi yang terencana, institusi pendidikan dapat mengidentifikasi mahasiswa berisiko secara lebih akurat, memberikan dukungan yang relevan, dan meningkatkan keberhasilan akademik secara menyeluruh.

Daftar Pustaka:

1. Amaliah, S., Nusrang, M., & Aswi, A. (2022). Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng. *VARIANSI: Journal of Statistics and Its application on Teaching and Research*, *4*(3), 121-127.
2. Baykara, Batuhan. (2015). Impact of Evaluation Methods on Decision Tree Accuracy. Tampere: University of Tampere.
3. Breiman, L., Friedman, Jerome H., Olshen, Richard A., dan Stone, Charles J. (1984). Classification and Regression Trees. New York: Chapman and Hall.
4. Realinho, V., Vieira Martins, M., Machado, J., & Baptista, L. (2021). Predict Students' Dropout and Academic Success [Dataset]. UCI Machine Learning Repository.

Lampiran:

Berikut adalah lampiran *codes* yang digunakan DATMIN UAS - Colab

Berikut adalah lampiran sumber data yang digunakan Data