**LAPORAN FINAL PROJECT**

**PEMBELAJARAN MESIN B 2023**

**“Implementasi dan Komparasi Model Pembelajaran Mesin terhadap Data UTBK Bidang Saintek Universitas Brawijaya pada Tahun 2019”**



**Disusun oleh Kelompok 2**

Nama Anggota:

1. Kevin Nathanael Halim 5025211140
2. Andika Rahman Teja 5025221022

**INSTITUT TEKNOLOGI SEPULUH NOPEMBER**

**DEPARTEMEN TEKNIK INFORMATIKA**

**2023**

**DAFTAR ISI**

[**1. Pendahuluan 1**](#_uiq7dkzi7daa)

[1.1. Latar Belakang 1](#_1lj9iae8ozan)

[1.2. Rumusan Masalah 2](#_2wjidois6swa)

[1.3. Batasan Masalah 2](#_718t2hc5n69z)

[1.4. Tujuan Penelitian 3](#_z925u8ttcc6v)

[1.5. Manfaat Penelitian 3](#_dg3zj5843m93)

[**2. Metodologi 4**](#_bks4msnszl3u)

[2.1. Metode Uji 4](#_d71pr6gjf3fo)

[2.1.1. Random Forest 4](#_bxl78fu17ey3)

[2.1.2. Support Vector Classifier (SVC) 4](#_wen8vjkryau4)

[2.1.3. eXtreme Gradient Boosting (XGBoost) 4](#_oyvaou1nqgiz)

[2.1.4. Artificial Neural Network (ANN) 5](#_tlq2gsk54mvl)

[2.2. Dataset 5](#_jgb0y0dhul5m)

[2.3. Desain Sistem 12](#_1fm3xe7e3dtz)

[2.3.1. Penjelasan Flowchart Penelitian 12](#_3qj4afjn93mh)

[**3. Hasil dan Pembahasan 16**](#_rzn9zzikrv3s)

[3.1. Skenario dan Hasil Pengujian 16](#_3b2xr6k476ib)

[3.1.1. Perbandingan Data Tanpa Normalisasi dengan Data Normalisasi 16](#_p5jcjnncdb42)

[1. Random Forest 16](#_38xlzh8gtzd9)

[2. SVC 17](#_67iwa1za3plu)

[3. XGBoost 18](#_9xxpcb4ks1xo)

[4. ANN 19](#_8ksdklqmwn8h)

[3.1.2. Perbandingan Data Tanpa Seleksi Fitur dengan Data Seleksi Fitur 20](#_aqg53clbm0y9)

[1. Random Forest 22](#_v2177y1q1xhw)

[a. Tanpa Seleksi Fitur 22](#_bc3l1sryfgk1)

[b. Dengan Seleksi Fitur 22](#_ymzjpqws2kxp)

[2. SVC 23](#_v2be4iv9ic5)

[a. Tanpa Seleksi Fitur 23](#_t3line1rit5j)

[b. Dengan Seleksi Fitur 23](#_fzlsnt1thpkd)

[3. XGBoost 24](#_v8h23bcdwryb)

[a. Tanpa Seleksi Fitur (Accuracy = 95 - 97%) 24](#_xtxwma31pyt1)

[b. Dengan Seleksi Fitur (Accuracy = 97 - 98%) 24](#_vjrojsv651lz)

[4. ANN 25](#_ku0vy9wqw2as)

[a. Tanpa Seleksi Fitur (Accuracy = 85 - 86%; Val\_Accuracy = 84 - 85%) 25](#_y211ij5lu0a7)

[b. Dengan Seleksi Fitur (Accuracy = 84 - 85%; Val\_Accuracy = 83 - 84%) 25](#_dtvzo133jo8m)

[3.1.3. Hyperparameter Tuning 26](#_8hgu6rf96ep7)

[1. Tanpa Hyperparameter Tuning 27](#_r9asg4hu3okp)

[2. Grid Search 28](#_bwmvc022nuol)

[3. Randomized Search 28](#_v7723f8hnho7)

[3.2. Analisis dan Pembahasan 29](#_njhisaq7ozdj)

[**4. Kesimpulan dan Saran 32**](#_qaduu61adekm)

[4.1. Kesimpulan 32](#_onoppiwilsh7)

[4.2. Saran 32](#_ln8s06cze6z4)

[**Daftar Pustaka 33**](#_3mdxrde4esp3)

# **Pendahuluan**

# 

## Latar Belakang

Pendidikan tingkat lanjutan khususnya perkuliahan merupakan hal yang sangat penting dan diperlukan guna membentuk generasi penerus bangsa yang mampu memimpin dan menjalankan bangsa ini di masa depan. Dengan adanya banyak anak muda yang berpendidikan, maka sumber daya manusia yang berkualitas akan tercipta**[1]**. Untuk melanjutkan pendidikan ke tingkat lanjut di universitas negeri yang kualitasnya dijamin oleh negara, seseorang harus melalui sebuah tes berbasis komputer tingkat nasional yang disebut dengan istilah UTBK atau Ujian Tulis Berbasis Komputer. UTBK sendiri merupakan sebuah tes masuk ke perguruan tinggi yang diselenggarakan oleh Lembaga Tes Masuk Perguruan Tinggi (LTMPT) sebagai pihak penanggung jawab tes masuk perguruan tinggi yang terstandarisasi**[2]**.

Sebelum tahun 2023, UTBK merupakan syarat utama untuk mengikuti program Seleksi Bersama Masuk Perguruan Tinggi Negeri (SBMPTN) yang saat ini telah berganti menjadi Seleksi Nasional Berbasis Tes (SNBT). UTBK terdiri dari 2 dua bagian yaitu Tes Potensi Skolastik (TPS) dan Tes Kemampuan Akademik (TKA), dengan TKA terbagi menjadi 2 bidang yaitu Saintek dan Soshum yang bergantung dari jurusan yang ingin diambil**[3]**. Materi untuk TPS terdiri penalaran umum, pengetahuan kuantitatif, pengetahuan dan pemahaman umum, serta pemahaman bacaan dan menulis**[4]**. Sementara itu, untuk materi TKA Saintek terdiri dari matematika, fisika, kimia, dan biologi**[3]**.

Pada tahun 2019, diketahui bahwa jumlah PTN di Indonesia berjumlah 372 instansi dan hanya 40 instansi yang masuk ke kelompok saintek dan mengadakan jalur penerimaan mahasiswa melalui SBMPTN**[5]**. Di tahun 2019, Universitas Brawijaya yang penentuan kelulusannya 100% diambil dari nilai UTBK tanpa pertimbangan lain seperti portofolio atau prestasi itu diketahui memiliki jumlah peminat tertinggi di tahun tersebut dengan angka mencapai 55.871 pendaftar dimana 33.987 di antaranya adalah pendaftar Program Studi Saintek**[6]**. Tentunya penentuan kelulusan menjadi hal yang cukup rumit, terlebih melihat banyaknya jumlah pendaftar yang masing-masing memiliki pilihan 1 dan 2 ketika mendaftarkan diri di SBMPTN. Untuk itu, diperlukan sebuah solusi yang efektif serta efisien guna membantu mempermudah penyelesaian permasalahan tersebut, salah satunya dengan mengimplementasikan Pembelajaran Mesin (*Machine Learning*).

Pembelajaran Mesin didefinisikan sebagai sub-area dari ilmu komputer yang memberikan komputer kemampuan untuk belajar tanpa diprogram secara eksplisit**[7]**. Pembelajaran mesin bekerja dengan mengeksplorasi ilmu dan konstruksi dari algoritma yang dapat belajar membuat prediksi berdasarkan data, terbagi menjadi 2 jenis yakni *Unsupervised Learning* yang menggunakan data tanpa label dan *Supervised Learning* dengan data berlabel yang kemudian dapat dikembangkan lagi menjadi *Deep Learning* yakni pembelajaran mesin yang lebih dalam karena memiliki lebih dari 1 layer tersembunyi**[8]**. Terdapat berbagai metode berbeda yang tersedia untuk mengimplementasikan pembelajaran mesin terkhusus dengan *Supervised Learning* untuk klasifikasi seperti Random Forest, SVC (*Support Vector Classifier*), dan XGBoost. Bahkan, penggunaan *Deep Learning* seperti ANN (*Artificial Neural Network*) juga menjadi salah satu opsi untuk mengatasi permasalahan tersebut. Maka dari itu, perlu dilakukan penelitian lebih lanjut guna mencari metode yang paling sesuai dan efektif untuk mengatasi masalah di atas.

## Rumusan Masalah

Berdasarkan latar belakang masalah yang telah dijelaskan di atas maka rumusan masalah dalam penelitian ini adalah:

1. Bagaimana efektivitas dan performa metode klasifikasi seperti Random Forest, SVC (*Support Vector Classifier*), XGBoost, dan ANN (*Artificial Neural Network*) dalam memahami relasi antara status kelulusan yang terdiri dari lulus di pilihan 1, pilihan 2, atau tidak lulus di keduanya dengan fitur-fitur yang tersedia yakni nilai-nilai UTBK dan program studi pilihan berdasarkan daya tampung?
2. Bagaimana pengaruh normalisasi, seleksi fitur, dan *hyperparameter tuning* dalam meningkatkan kualitas serta ketepatan metode klasifikasi seperti Random Forest, SVC (*Support Vector Classifier*), XGBoost (*eXtreme Gradient Boosting*), dan ANN (*Artificial Neural Network*) dalam menentukan status kelulusan berdasarkan fitur-fitur yang tersedia yakni nilai-nilai UTBK dan program studi pilihan berdasarkan daya tampung?
3. Apa metode klasifikasi yang paling efektif dan sesuai untuk membantu menentukan kelulusan peserta UTBK bidang Saintek di Universitas Brawijaya berdasarkan fitur-fitur yang tersedia yakni nilai-nilai UTBK dan program studi pilihan berdasarkan daya tampung?

## Batasan Masalah

Beberapa batasan masalah untuk menghindari adanya penyimpangan maupun pelebaran pokok masalah agar penelitian tersebut lebih terarah dalam penelitian ini adalah sebagai berikut:

1. Penelitian dilakukan menggunakan data nilai hasil UTBK bidang Saintek pada tahun 2019 dari peserta yang setidaknya memilih program studi di Universitas Brawijaya sebagai salah satu pilihannya.
2. Penelitian dilakukan hanya dengan mengimplementasikan dan membandingkan metode klasifikasi yang telah disebutkan yakni Random Forest, SVC (*Support Vector Classifier*), XGBoost (*eXtreme Gradient Boosting*), dan ANN (*Artificial Neural Network*) tanpa melakukan percobaan terhadap metode lain yang memungkinkan.

## Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk mengimplementasikan, menganalisis, dan membandingkan efektivitas serta kesesuaian dari metode klasifikasi yakni Random Forest, SVC (*Support Vector Classifier*), XGBoost, dan ANN (*Artificial Neural Network*) dalam memprediksi status kelulusan berdasarkan data nilai hasil UTBK dan kedua pilihan program studi serta kampus yang dipilih.

## Manfaat Penelitian

Beberapa manfaat yang dapat diperoleh dari dijalankannya penelitian ini adalah sebagai berikut:

1. Bagi peneliti:
   1. Hasil dari penelitian diharapkan dapat memberi manfaat bagi peneliti untuk menjelaskan terkait perbandingan efektivitas dari setiap metode klasifikasi dalam prediksi status kelulusan UTBK berdasarkan fitur yang digunakan.
   2. Mengimplementasikan pengetahuan serta ilmu yang sudah diperoleh selama mengikuti mata kuliah Pembelajaran Mesin, terutama yang berkaitan dengan *Supervised Learning* dan *Deep Learning*.
2. Bagi pihak lain:
3. Hasil dari penelitian diharapkan dapat digunakan untuk mempermudah penentuan status kelulusan dari UTBK yang dilakukan berdasarkan nilai hasil tes dan program studi serta kampus pilihannya.
4. Diharapkan dapat menciptakan kesempatan untuk penelitian lebih lanjut di masa depan terkait pembuatan sistem yang dapat menentukan kelulusan dari UTBK secara efektif dan efisien.
5. Diharapkan dapat memberi manfaat bagi pembaca untuk mempelajari karakteristik dari setiap metode klasifikasi yang digunakan terhadap permasalahan yang diangkat agar dapat digunakan sebagai referensi untuk penelitian lebih lanjut.

# **Metodologi**

## Metode Uji

### Random Forest

Random Forest adalah sebuah metode *ensemble learning* bersifat non-parametrik yang pertama kali diperkenalkan oleh Breiman untuk permasalahan klasifikasi maupun regresi yang dapat digunakan membangun aturan-aturan prediksi berdasarkan berbagai variabel beragam yang tersedia**[9]**. Metode ini menggunakan *randomization* untuk membuat *decision tree* dalam jumlah yang besar, dimana hasil keluaran dari *tree-tree* tersebut akan dijadikan sebuah keluaran akhir dengan sistem pengambilan suara terbanyak atau mayoritas untuk klasifikasi atau pencarian rata-rata untuk regresi**[10]**.

### Support Vector Classifier (SVC)

Support Vector Classifier atau yang biasa disebut SVC adalah salah satu jenis metode klasifikasi simpel yang berdasarkan Support Vector Machines (SVM) yang pertama kali digagas oleh V. Vapnik, sementara itu SVC sendiri dikemukakan oleh Chih-Jen Lin**[11]**. SVC digunakan untuk mengisolasi dataset *training vectors* yang diberikan menjadi 2 kelas berbeda dengan ruang optimal yang disebut dengan Maximal Margin of Separation (MMS) berdasarkan Optimal Separating Hyperplane (OSH), tujuan akhirnya adalah untuk mendapatkan *trained vectors* dalam jumlah kecil dari setiap kelas untuk dijadikan basis dari SVC guna menyelesaikan berbagai permasalahan klasifikasi dengan akurat**[12]**.

### eXtreme Gradient Boosting (XGBoost)

XGBoost atau yang merupakan singkatan dari eXtreme Gradient Boosting adalah sebuah metode pembelajaran mesin *scalable* yang mengimplementasikan algoritma tree boosting, yakni teknik pembelajaran mesin yang menciptakan *weak learner* di setiap langkah yang nantinya akan digabungkan ke model. Metode ini bekerja dengan melakukan penambahan *tree* di setiap langkah untuk mengkomplemen hasil model yang telah dibuat, berbeda dengan metode Random Forest yang masing-masing *tree*-nya dibangun secara individu**[13]**.

### Artificial Neural Network (ANN)

Artificial Neural Network atau yang biasa disebut dengan ANN adalah sebuah metode *deep learning* berupa sistem komputasi kompleks yang terdiri dari prosesor sederhana yakni Artificial Neurons dalam jumlah yang sangat besar dengan banyak interkoneksi, terinspirasi dari sistem saraf biologis sehingga dianggap dapat bekerja seperti cara pemikiran otak pada manusia karena mencoba menggambarkan cara kerja sistem saraf secara artifisial**[13]**. Metode ini bekerja dengan perulangan pemrosesan massal dari banyak unit yang saling terkoneksi hingga tercapai performa yang cukup baik serta dapat diterima, dengan keluaran dari setiap unit bergantung dari informasi yang ada di unit tersebut, baik yang memang tersimpan secara lokal di dalamnya maupun yang berasal dari unit lain melalui koneksi tersambung**[14]**.

## Dataset

Penelitian ini merupakan sebuah upaya sistematis untuk menggali dan menganalisis informasi dari dataset yang diperoleh dari Kaggle, sebuah platform daring terkemuka dalam berbagai kompetisi data dan sumber daya ilmiah. Dataset yang digunakan dalam penelitian ini berisi data tentang UTBK di Indonesia pada tahun 2019 yang dapat diakses melalui tautan berikut:

## 

<https://www.kaggle.com/datasets/ekojsalim/indonesia-college-entrance-examination-utbk-2019>

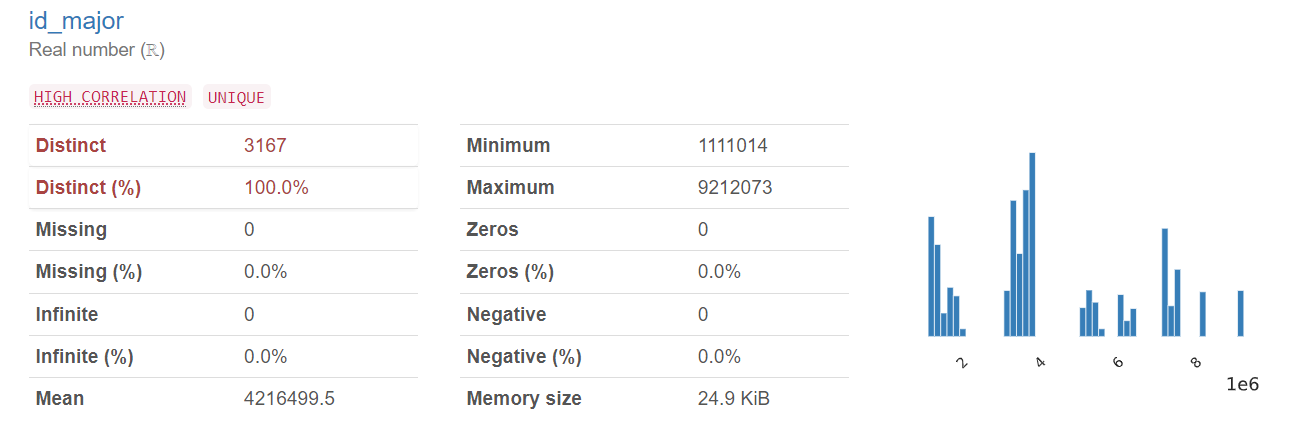
Pada dataset tersebut terdiri atas 4 sub-dataset yang masing-masing memiliki keterkaitan melalui relasi *Primary Key - Foreign Key* yang ada pada atribut di tiap sub-dataset. Pada penelitian ini, dataset ‘score\_humanities’ tidak digunakan karena sesuai dengan batasan penelitian. Adapun atribut dan spesifikasi data pada dataset tersebut adalah sebagai berikut:

* majors.csv

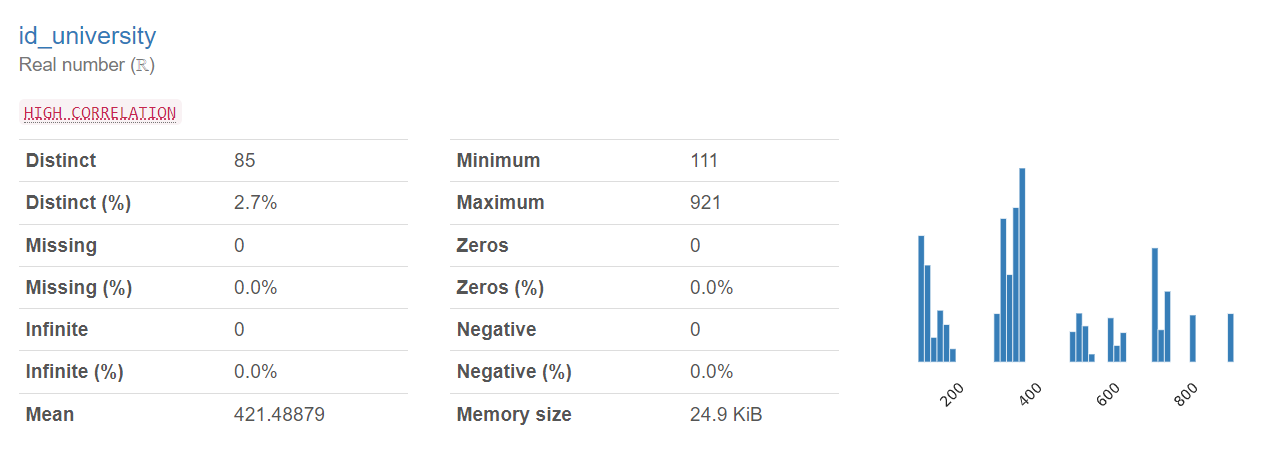
Columns = [‘Index’, ‘id\_major’, ‘id\_university’, ‘type’, ‘major\_name’, ‘capacity’]

Primary Key = **id\_major**

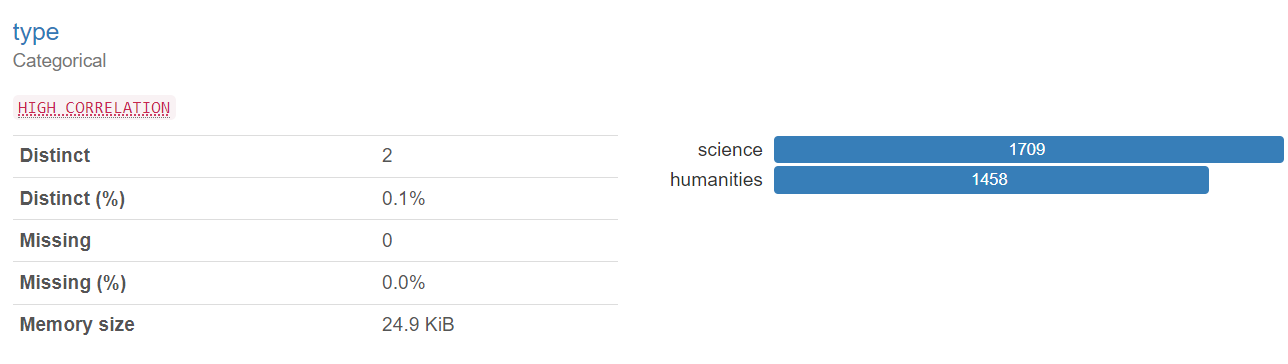
Foreign Key = **id\_university**

****

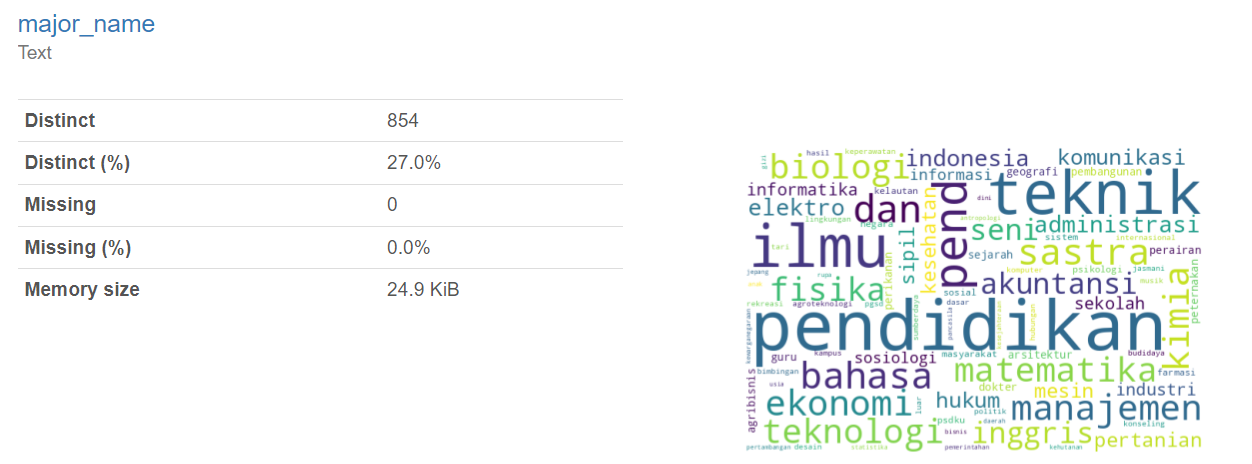
Gambar 2.1 Spesifikasi Data ‘id\_major’

****

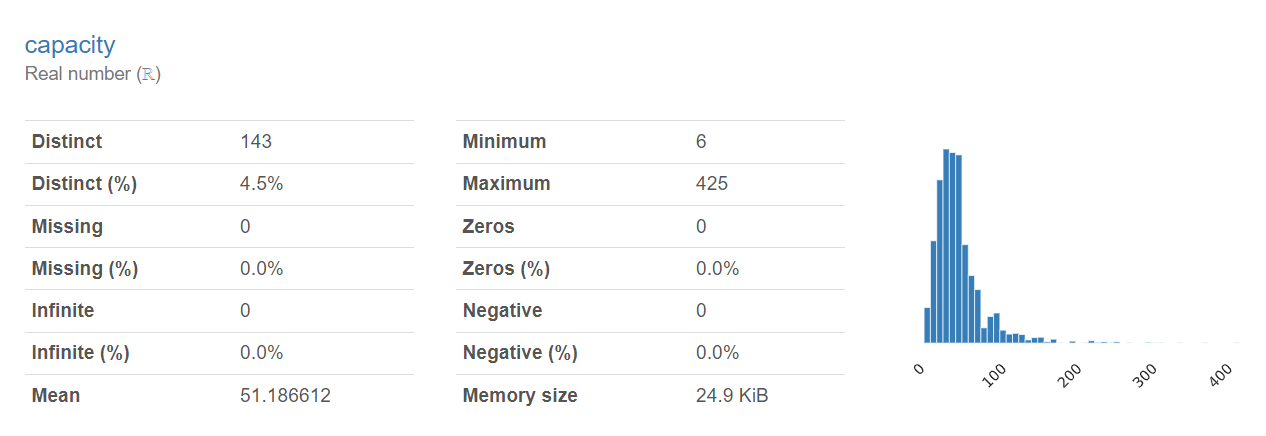
Gambar 2.2 Spesifikasi Data ‘id\_university’

****

Gambar 2.3 Spesifikasi Data ‘type’

****

Gambar 2.4 Spesifikasi Data ‘major\_name’

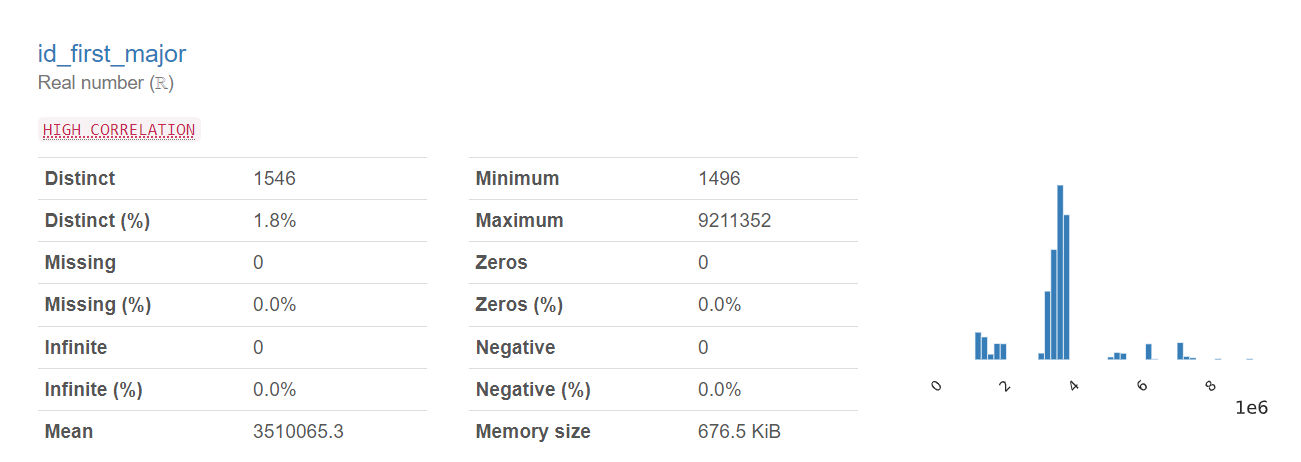
****

Gambar 2.5 Spesifikasi Data ‘capacity’

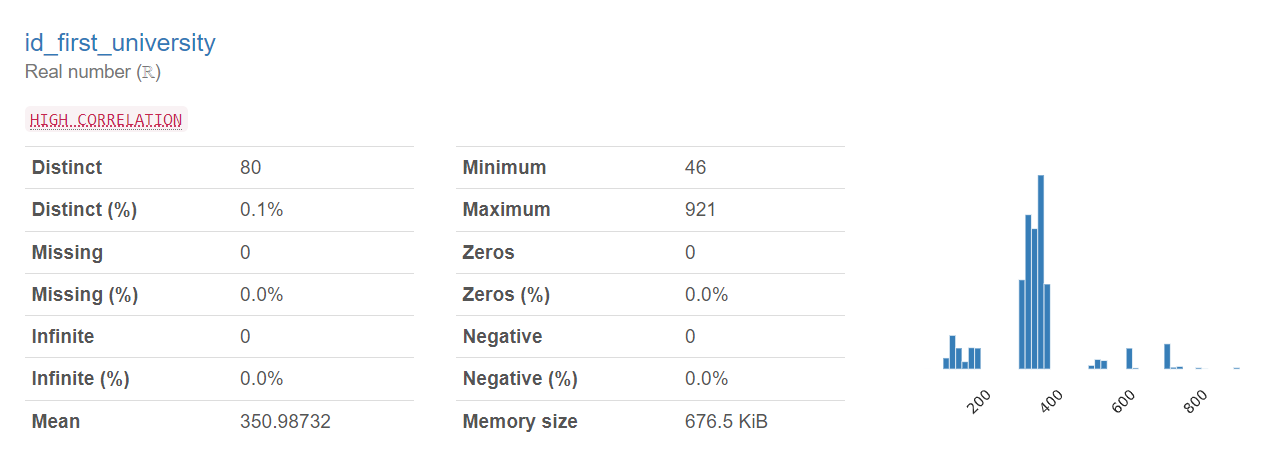
* score\_science.csv

Columns = [‘Index’, ‘id\_first\_major’, ‘id\_first\_university’, ‘id\_second\_major’, ‘id\_second\_university’, ‘id\_user’, ‘score\_bio’, ‘score\_fis’, ‘score\_kim’, ‘score\_kmb’, ‘score\_kpu’, ‘score\_kua’, ‘score\_mat’, ‘score\_ppu’]

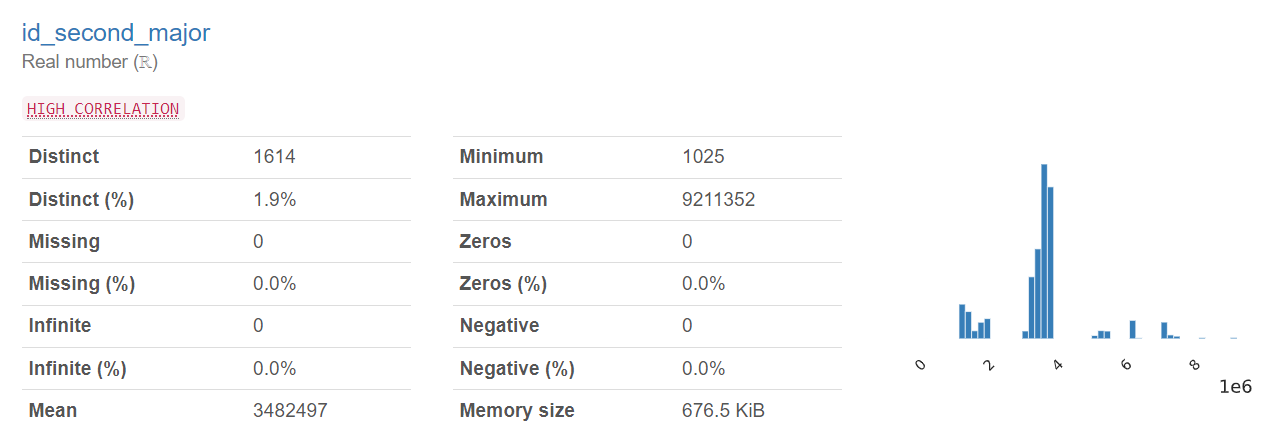
Foreign Key = **id\_first\_major**, **id\_first\_major**, **id\_second\_major**, **id\_second\_university**

****

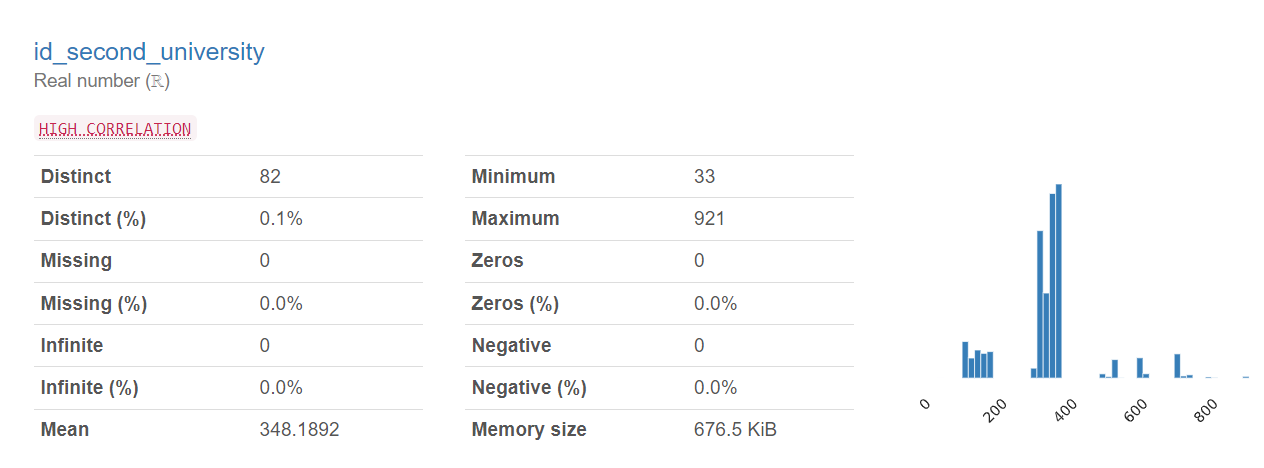
Gambar 2.6 Spesifikasi Data ‘id\_first\_major’

****

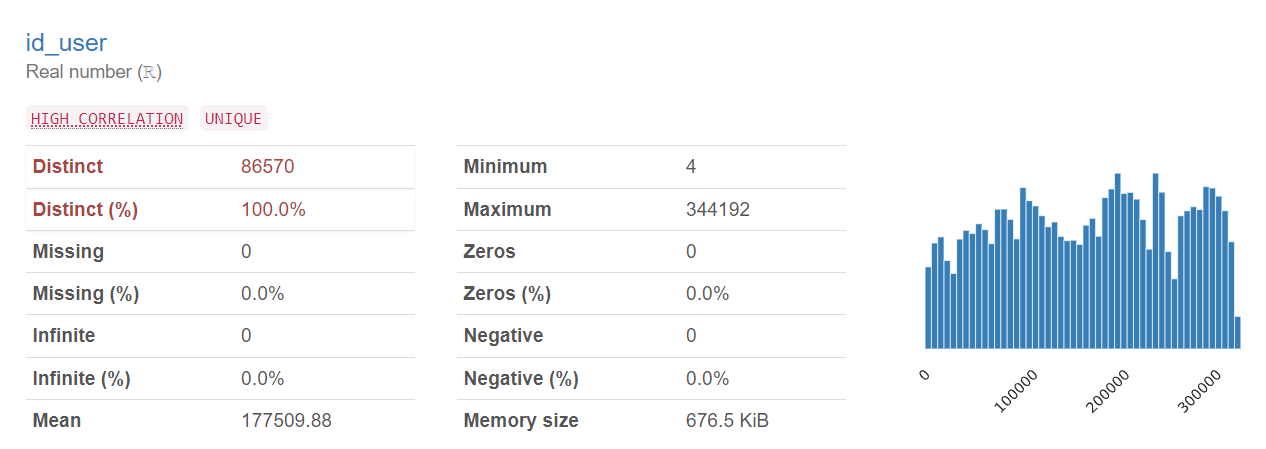
Gambar 2.7 Spesifikasi Data ‘id\_first\_university’

****

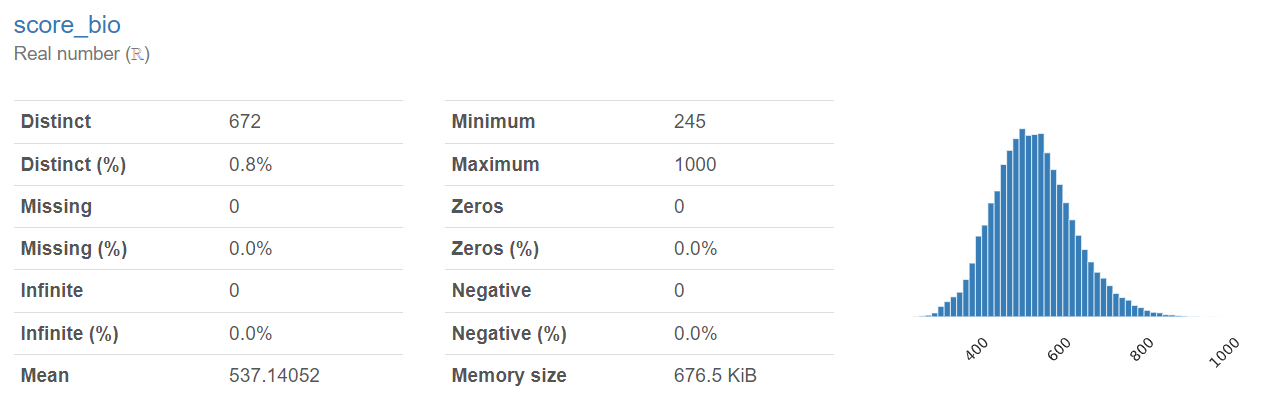
Gambar 2.8 Spesifikasi Data ‘id\_second\_major’

****

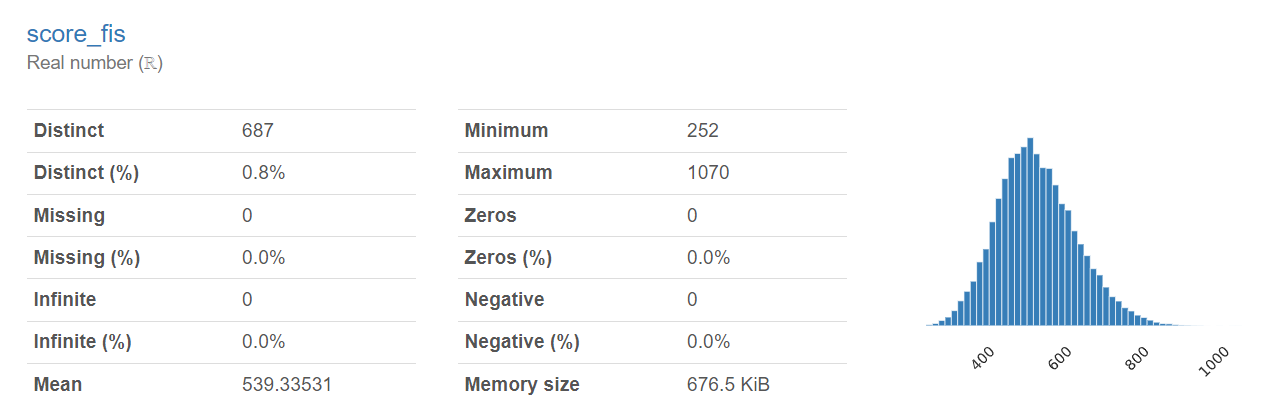
Gambar 2.9 Spesifikasi Data ‘id\_second\_university’

****

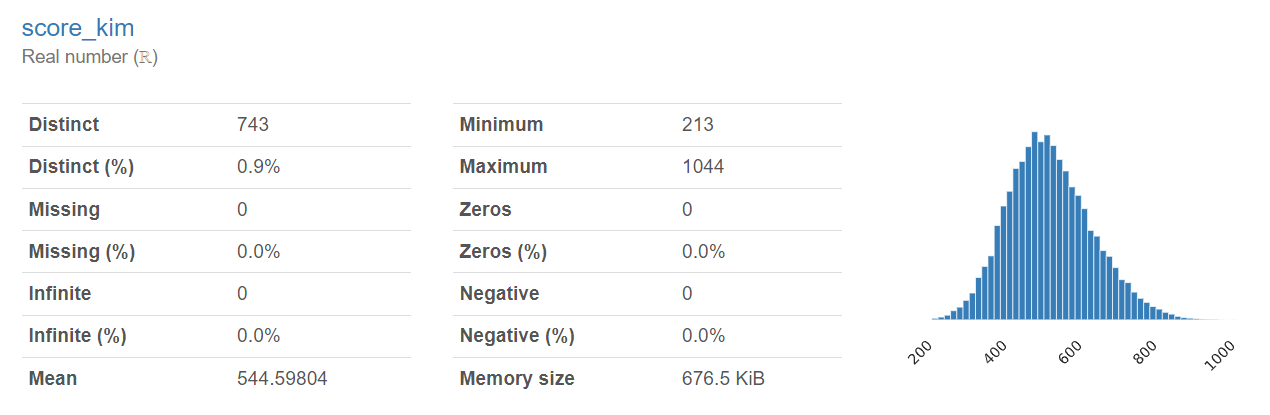
Gambar 2.10 Spesifikasi Data ‘id\_user’

****

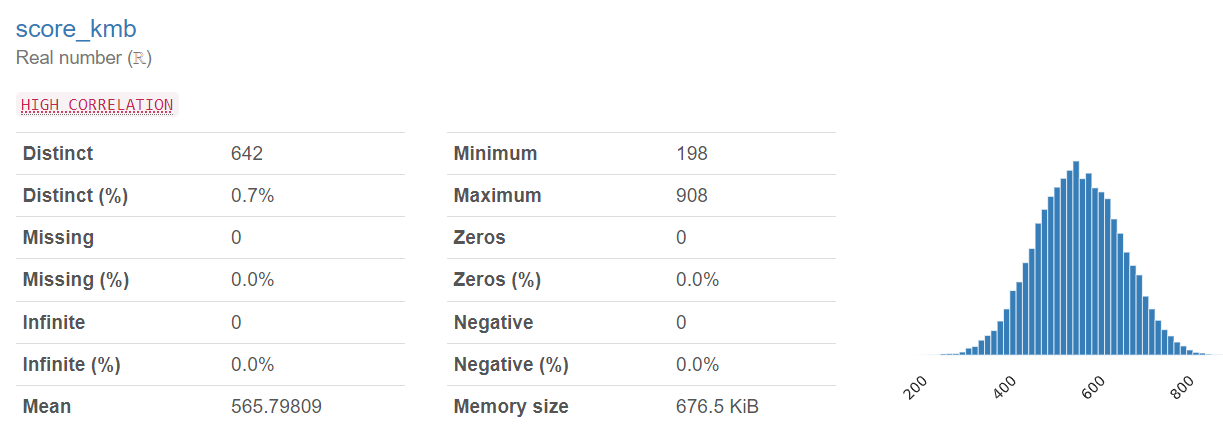
Gambar 2.11 Spesifikasi Data ‘score\_bio’

****

Gambar 2.12 Spesifikasi Data ‘score\_fis’

****

Gambar 2.13 Spesifikasi Data ‘score\_kim’

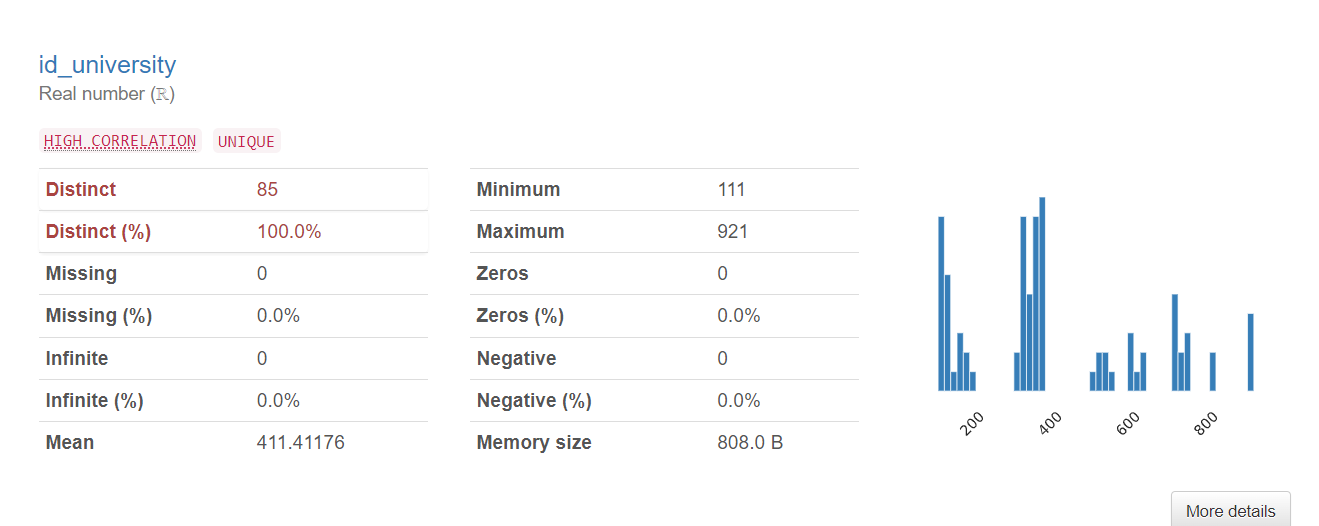
****

Gambar 2.14 Spesifikasi Data ‘score\_kmb’

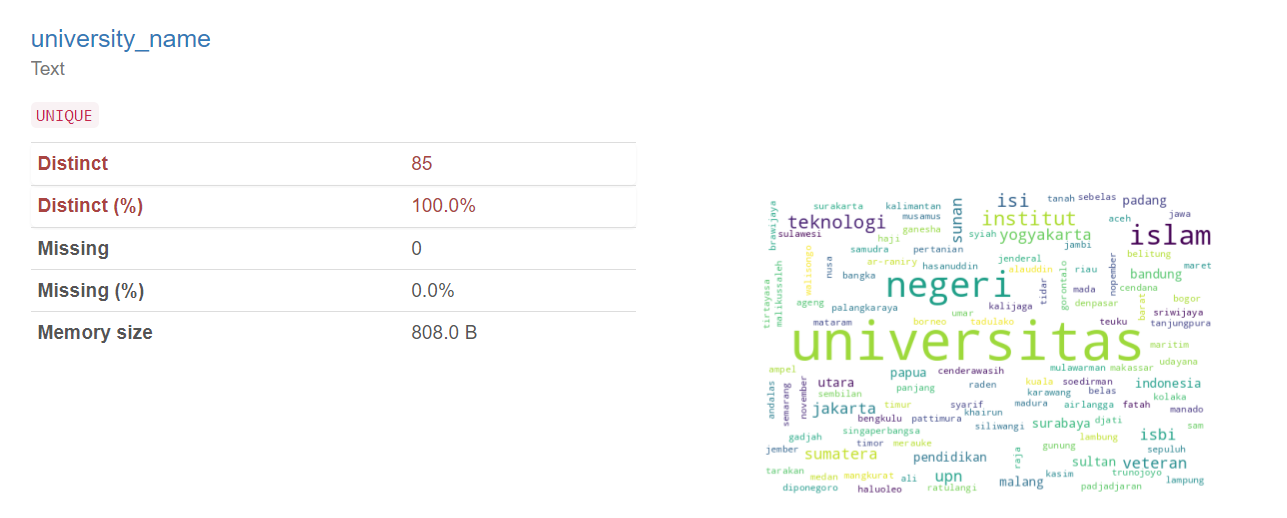
* universities.csv

Columns = [‘Index’, ‘id\_university’, ‘university\_name’]

Primary Key = **id\_university**

****

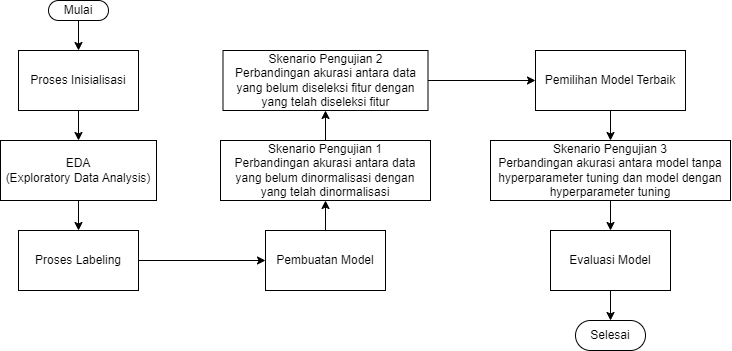
Gambar 2.15 Spesifikasi Data ‘id\_university’

****

Gambar 2.16 Spesifikasi Data ‘university\_name’

## Desain Sistem

Pada tahapan desain sistem ini, dilakukan beberapa langkah kerja yang telah disusun dalam bentuk *flowchart* sebagai berikut:



Gambar 2.17 *Flowchart* Penelitian

## 

### Penjelasan *Flowchart* Penelitian

Berikut ini langkah-langkah yang dilakukan dalam pemecahan masalah pada penelitian:

1. Proses Inisialisasi

Melakukan proses inisialisasi *library* yang akan digunakan dan melakukan proses inisialisasi dataset dengan menggunakan *Pandas.*

1. EDA (*Exploratory Data Analysis*)

Melakukan visualisasi data, melakukan *drop* fitur yang tidak relevandan *drop* baris untuk menghilangkan nilai NaN, menggabungkan sub-dataset menjadi 1 dataset, dan melakukan *filtering* data berdasarkan universitas.

1. Proses *Labeling*

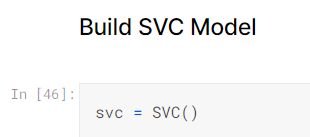
Melakukan penghitungan kapasitas yang tersedia pada tiap jurusan yang ada pada tiap universitas dan memberikan label pada tiap *user* atau peserta UTBK dengan 3 jenis label, yakni '*Accepted First Choice*' (jika peserta diterima di pilihan pertama), '*Accepted Second Choice*' (jika peserta diterima di pilihan kedua), dan '*Failed*' (jika peserta tidak diterima di pilihan manapun).

1. Pembuatan Model

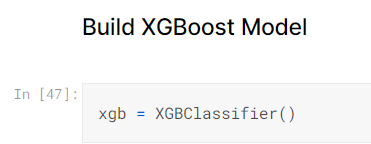
Melakukan deklarasi model pembelajaran mesin dan membuat arsitektur model *deep learning*, yakni ANN.



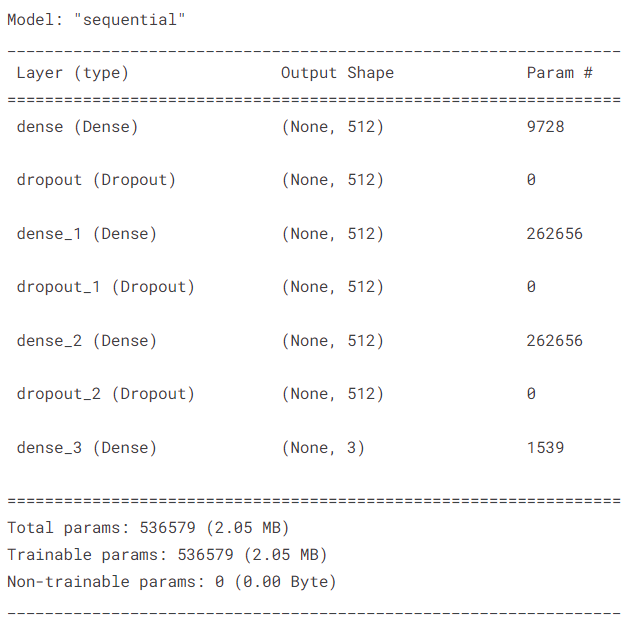
Gambar 2.18 Deklarasi Model Random Forest



Gambar 2.19 Deklarasi Model SVC



Gambar 2.20 Deklarasi Model XGBoost





Gambar 2.21 Arsitektur Model ANN

1. Skenario Pengujian 1

Melakukan perbandingan akurasi antar model pada data yang belum dinormalisasi dengan data yang telah dinormalisasi. Adapun pengujian untuk model pembelajaran mesin (seperti Random Forest, SVC, dan XGBoost) dilakukan dengan melakukan *cross validation*).

1. Skenario Pengujian 2

Melakukan perbandingan akurasi antar model pada data yang belum diseleksi fitur dengan data yang telah diseleksi fitur. Adapun pengujian untuk model pembelajaran mesin (seperti Random Forest, SVC, dan XGBoost) dilakukan dengan melakukan *cross validation*).

1. Pemilihan Model Terbaik

Memilih model yang memiliki akurasi terbaik dari 2 skenario pengujian sebelumnya.

1. Skenario Pengujian 3

Melakukan *hyperparameter tuning* pada model terbaik yang telah dipilih sebelumnya dan membandingkan akurasinya dengan kondisi model tanpa menggunakan *hyperparameter tuning*.

1. Evaluasi Model

Memilih model terbaik setelah ketiga skenario pengujian dilakukan dan melakukan *classification report* pada data hasil prediksi model tersebut terhadap data *test*.

# **Hasil dan Pembahasan**

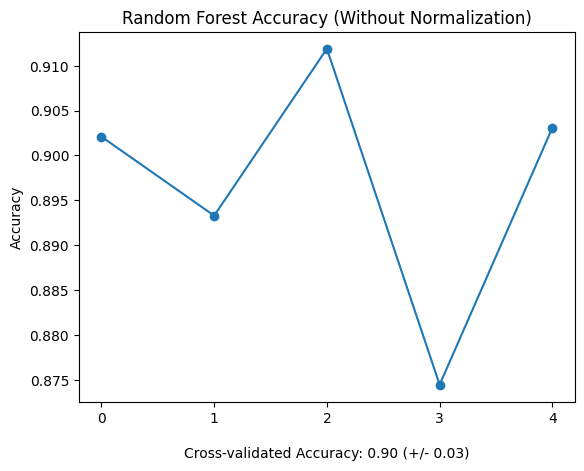
## Skenario dan Hasil Pengujian

### Perbandingan Data Tanpa Normalisasi dengan Data Normalisasi

Proses normalisasi pada kasus ini dilakukan dalam bentuk standarisasi sehingga setiap kolom pada data memiliki mean 0 dan standar deviasi 1 menggunakan StandardScaler, perbandingan hasilnya terhadap 4 metode yang ada sebagai berikut,

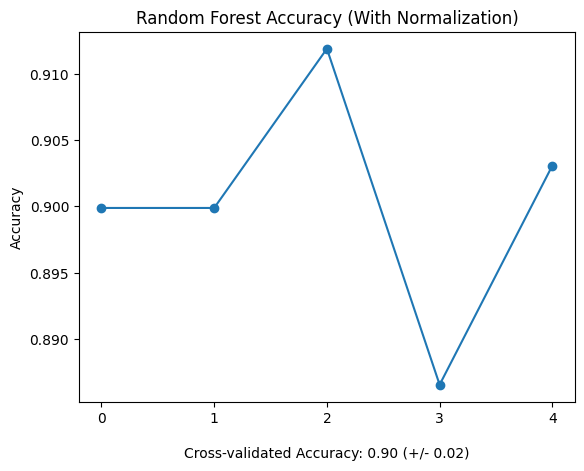
#### Random Forest

* 1. Tanpa Normalisasi (Accuracy = 87 - 91%)



Gambar 3.1 Hasil Akurasi Random Forest Tanpa Normalisasi

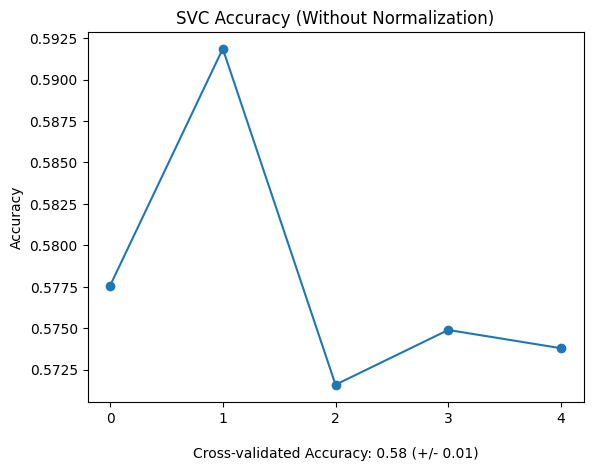
* 1. Dengan Normalisasi (Accuracy = 88 - 91%)



Gambar 3.2 Hasil Akurasi Random Forest dengan Normalisasi

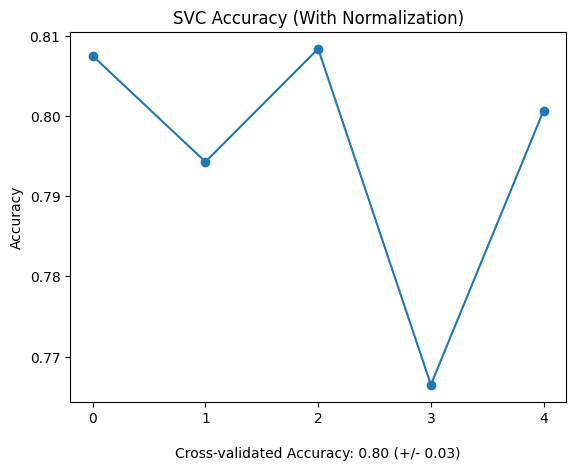
#### SVC

* 1. Tanpa Normalisasi (Accuracy = 57 - 59%)



Gambar 3.3 Hasil Akurasi SVC Tanpa Normalisasi

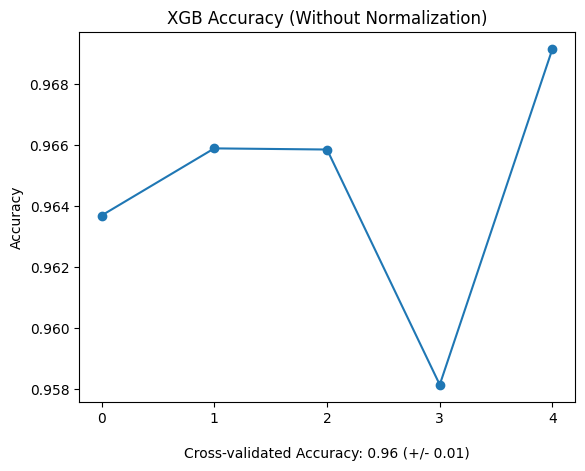
* 1. Dengan Normalisasi (Accuracy = 76 - 81%)



Gambar 3.4 Hasil Akurasi SVC dengan Normalisasi

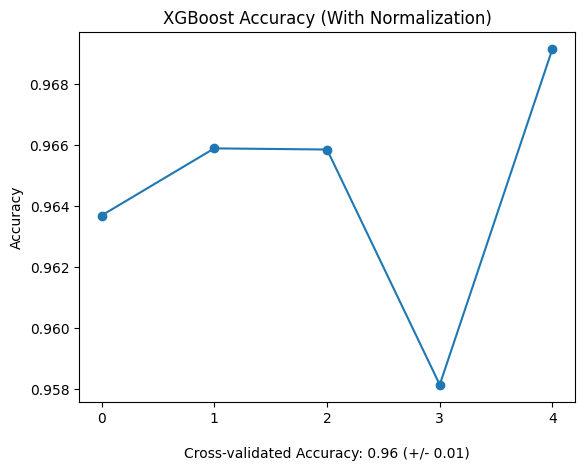
#### XGBoost

* 1. Tanpa Normalisasi (Accuracy = 95 - 97%)



Gambar 3.5 Hasil Akurasi XGBoost Tanpa Normalisasi

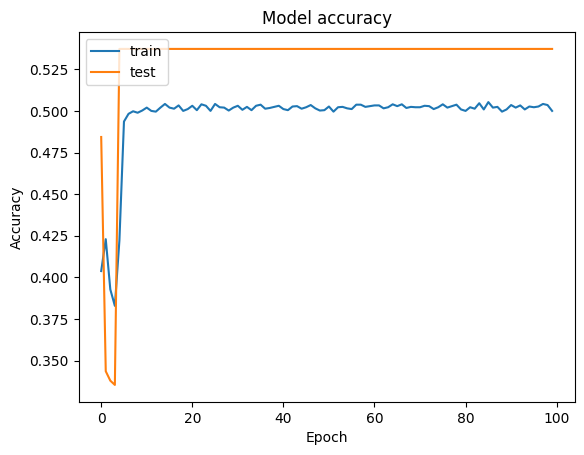
* 1. Dengan Normalisasi (Accuracy = 95 - 97%)



Gambar 3.6 Hasil Akurasi XGBoost dengan Normalisasi

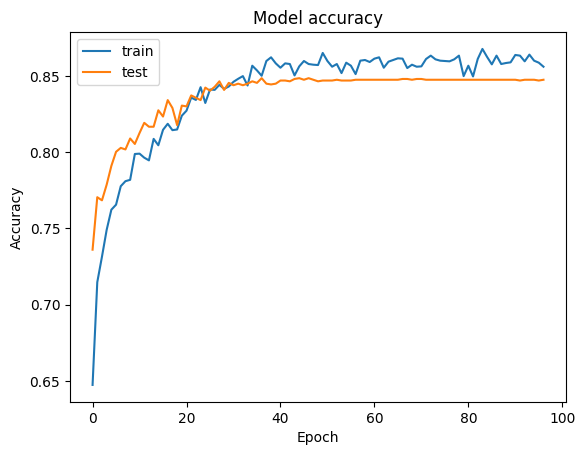
#### ANN

* 1. Tanpa Normalisasi (Accuracy = 49 - 50%; Val\_Accuracy = 53%)



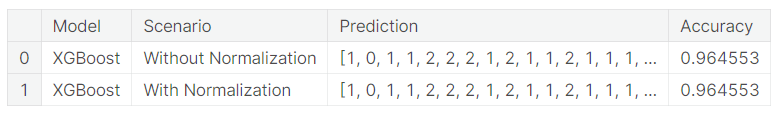
Gambar 3.7 Hasil Akurasi ANN Tanpa Normalisasi

* 1. Dengan Normalisasi (Accuracy = 85 - 86%; Val\_Accuracy = 84 - 85%)



Gambar 3.8 Hasil Akurasi ANN dengan Normalisasi

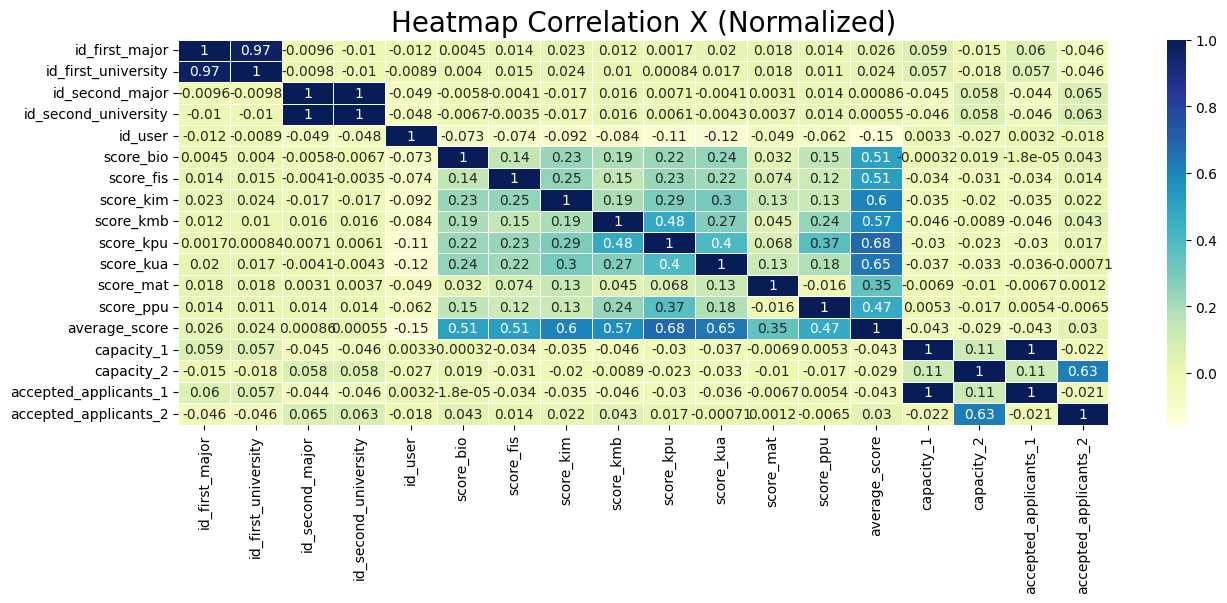
Dari hasil yang didapatkan, dapat dilihat bahwa tidak terjadi perubahan yang signifikan pada model dengan metode Random Forest dan XGBoost dengan nilai akurasi yang hampir tidak berubah. Tetapi, terjadi peningkatan akurasi yang cukup signifikan pada dua model lainnya yakni SVC dan ANN sehingga penggunaan Normalisasi pada kasus ini dianggap dapat membantu meningkatkan performa dan efektivitas model. Selain itu, didapatkan kesimpulan pada skenario ini, model XGBoost adalah model yang paling baik dibandingkan ketiga model lainnya.



Gambar 3.9 Model Terbaik dari Skenario 1

### Perbandingan Data Tanpa Seleksi Fitur dengan Data Seleksi Fitur

Proses seleksi fitur pada kasus ini dilakukan dengan melakukan penghapusan (*drop*) terhadap beberapa fitur dan diwakili oleh fitur lain yang memang dianggap mampu untuk merepresentasikan fitur-fitur tersebut. Guna mencari fitur representasi tersebut dapat dilakukan dengan melakukan pengamatan terhadap *heatmap* korelasi antar setiap atribut yang bentuknya dapat dilihat pada gambar berikut,



Gambar 3.10 *Heatmap Correlation* Antar Fitur-Fitur Pada Dataset

Berdasarkan *heatmap* tersebut, dapat diperoleh pemetaan fitur representasi dan fitur-fitur yang direpresentasikan sebagai berikut,

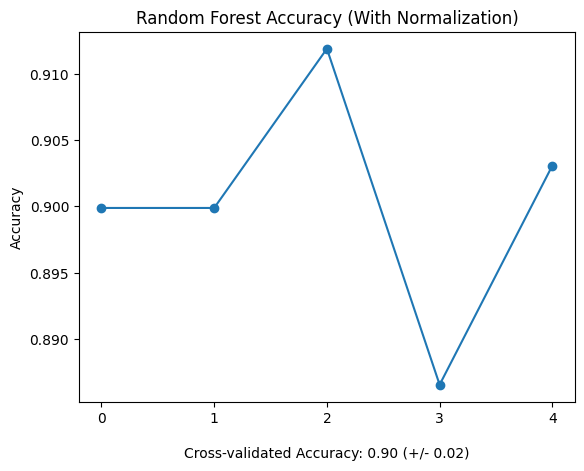
[Fitur Representasi] => [Fitur yang Direpresentasikan]

* id\_first\_major => id\_first\_university
* id\_second\_major => id\_second\_university
* average\_score => score\_bio, score\_fis, score\_kim, score\_kmb, score\_kpu, score\_kua, score\_mat, score\_ppu
* capacity\_1 => accepted\_applicants\_1
* capacity\_2 => accepted\_applicants\_2

Setelah penerapan seleksi fitur di atas, perbandingan hasil dari keempat metode yang digunakan sebagai berikut,

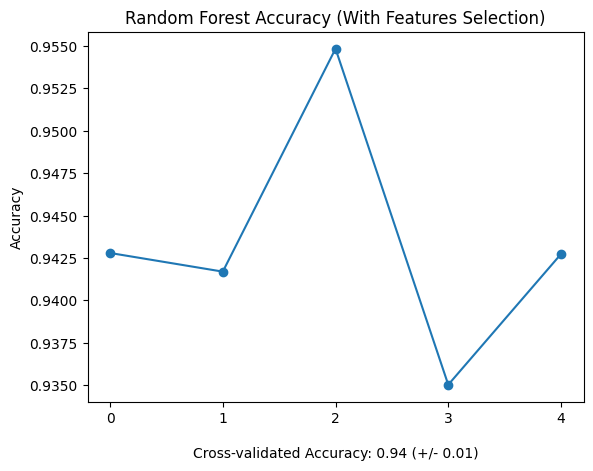
#### Random Forest

##### Tanpa Seleksi Fitur (Accuracy = 88 - 91%)



Gambar 3.11 Hasil Akurasi Random Forest Tanpa Seleksi Fitur

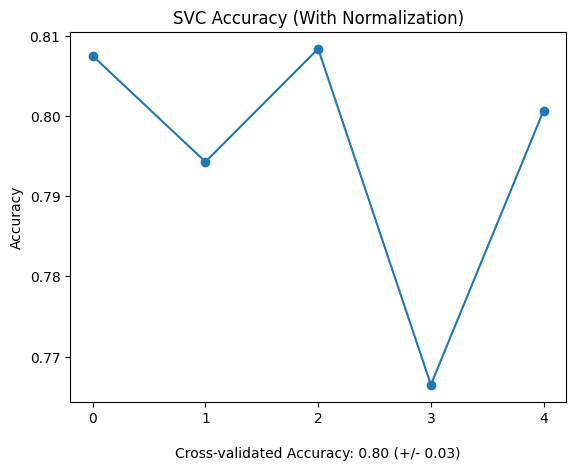
##### Dengan Seleksi Fitur (Accuracy = 93 - 95%)



Gambar 3.12 Hasil Akurasi Random Forest dengan Seleksi Fitur

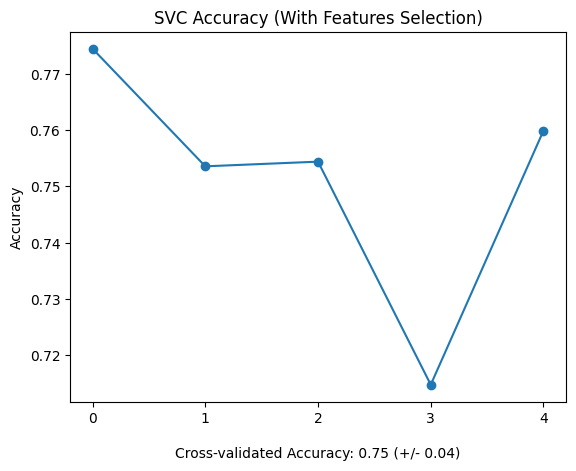
#### SVC

##### Tanpa Seleksi Fitur (Accuracy = 76 - 81%)



Gambar 3.13 Hasil Akurasi SVC Tanpa Seleksi Fitur

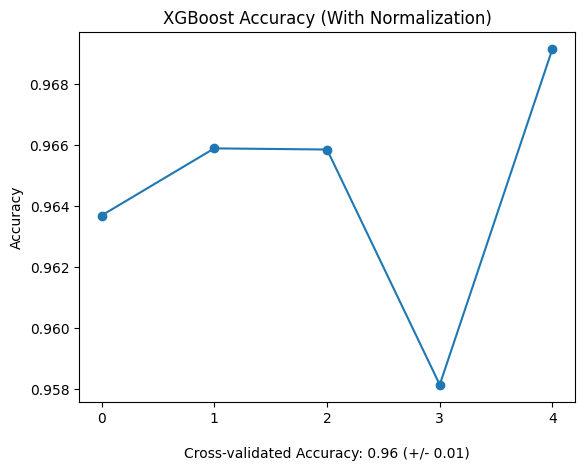
##### Dengan Seleksi Fitur (Accuracy = 71 - 77%)



Gambar 3.14 Hasil Akurasi SVC dengan Seleksi Fitur

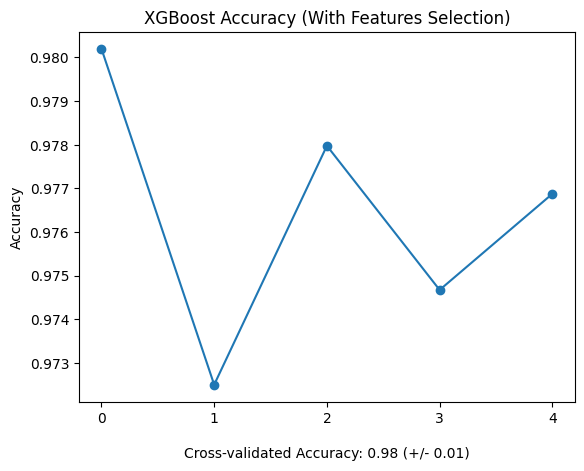
#### XGBoost

##### Tanpa Seleksi Fitur (Accuracy = 95 - 97%)



Gambar 3.15 Hasil Akurasi XGBoost Tanpa Seleksi Fitur

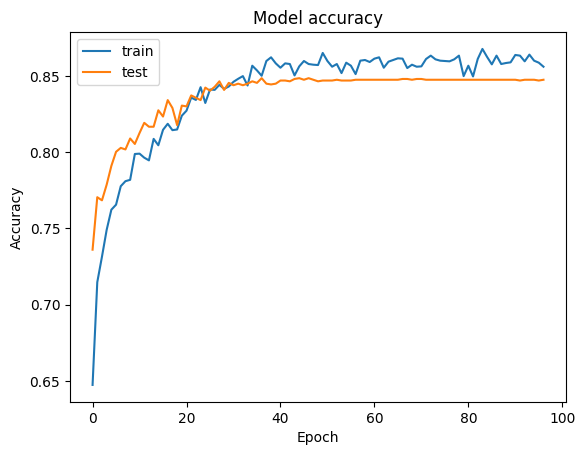
##### Dengan Seleksi Fitur (Accuracy = 97 - 98%)



Gambar 3.16 Hasil Akurasi XGBoost dengan Seleksi Fitur

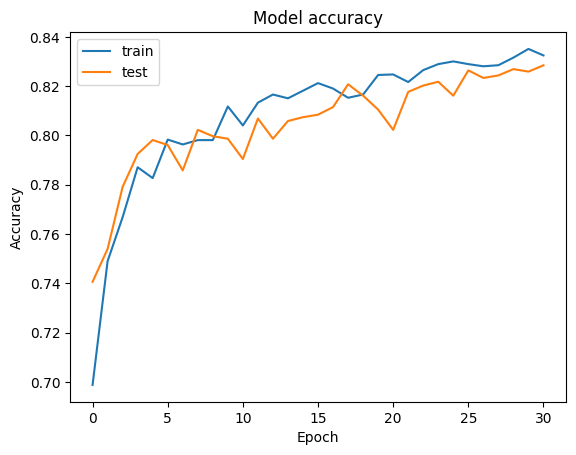
#### ANN

##### Tanpa Seleksi Fitur (Accuracy = 85 - 86%; Val\_Accuracy = 84 - 85%)



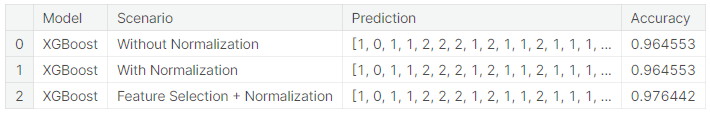
Gambar 3.17 Hasil Akurasi ANN Tanpa Seleksi Fitur

##### Dengan Seleksi Fitur (Accuracy = 84 - 85%; Val\_Accuracy = 83 - 84%)



Gambar 3.18 Hasil Akurasi ANN dengan Seleksi Fitur

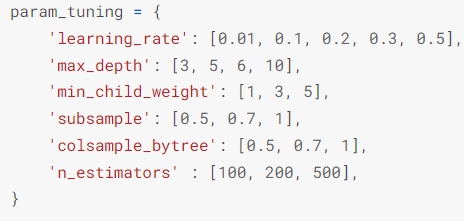
Dari hasil yang didapatkan, dapat dilihat bahwa terjadi kenaikkan akurasi pada 2 model yakni Random Forest dan XGBoost, tetapi di sisi lain 2 model yang tersisa yakni SVM dan ANN justru mengalami penurunan dengan dilakukannya seleksi fitur ini. Akan tetapi, karena di tahap berikutnya kita hanya perlu memilih 1 metode yang terbaik yang dalam hal ini adalah XGBoost dengan akurasi mencapai 97 hingga 98% yang lebih tinggi dibandingkan 3 metode lainnya bila mengaplikasikan seleksi fitur sehingga diputuskan bahwa penggunaan seleksi fitur di kasus ini lebih baik karena dapat meningkatkan performa serta efektivitas dari salah dua metode hingga mencapai akurasi yang sangat tinggi bila dibandingkan dengan penurunan yang disebabkan untuk kedua model lainnya. Adapun perbandingan model terbaik (yakni XGBoost) pada skenario seleksi fitur dengan skenario normalisasi sebelumnya adalah sebegai berikut:



Gambar 3.19 Model Terbaik Setelah 2 Skenario

### *Hyperparameter Tuning*

Proses *hyperparameter tuning* pada kasus ini dilakukan dengan melakukan pencarian parameter yang terbaik bagi kasus ini untuk metode XGBoost yang merupakan metode terbaik hasil dari skenario sebelumnya. Berikut ini parameter yang akan di-*tuning* pada model XGBoost:



Gambar 3.20 Parameter XGBoost

* *Learning\_rate*: mengontrol seberapa besar langkah *learning* yang diambil dalam setiap iterasi.
* *Max\_depth*: menentukan kedalaman maksimum dari setiap *tree*. Kedalaman yang terlalu besar dapat menyebabkan overfitting.
* *Min\_child\_weight*: menentukan jumlah minimum data yang diperlukan di setiap *child* dari *tree*.
* *Subsample*: menentukan fraksi dari data yang akan digunakan dalam setiap iterasi.
* *Colsample\_bytree*: menentukan fraksi dari fitur yang akan digunakan.
* *N\_estimators*: menentukan jumlah *decision tree* yang akan dibangun.

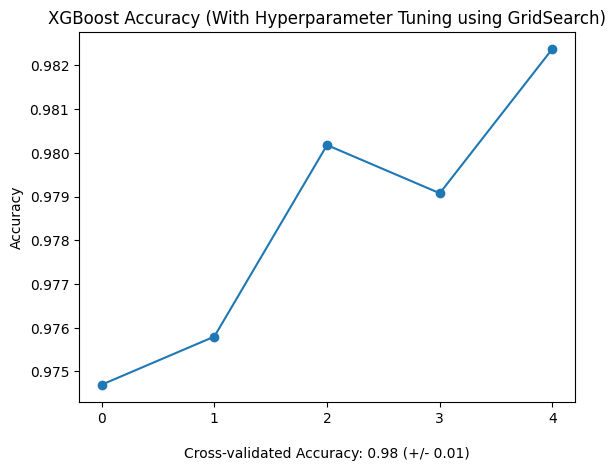
Pencarian parameter terbaik sendiri dilakukan dengan 2 metode yakni *Grid Search* dan *Randomized Search* dengan hasil sebagai berikut,

#### Tanpa *Hyperparameter Tuning*

#### 

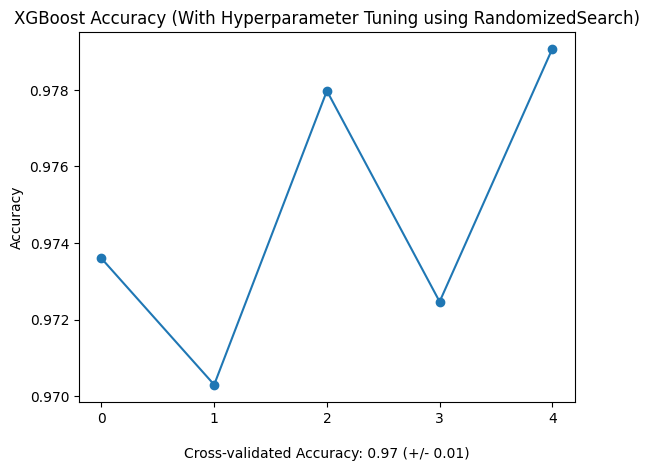
Gambar 3.21 Hasil Akurasi XGBoost Tanpa Hyperparameter Tuning

#### *Grid Search*



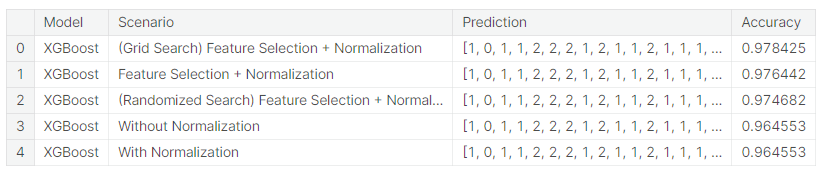
Gambar 3.22 Hasil Akurasi XGBoost dengan *Grid Search*

#### *Randomized Search*



Gambar 3.23 Hasil Akurasi XGBoost dengan *Randomized Search*

Dari hasil yang didapatkan, dapat dilihat bahwa terjadi kenaikkan akurasi walau dalam jumlah kecil pada model dengan *hyperparameter tuning* menggunakan *Grid Search*. Sementara itu, untuk parameter *Randomized Search* sendiri menyebabkan penurunan akurasi. Maka dari itu, penggunaan *Hyperparameter Tuning* dengan *Grid Search* lebih baik di kasus ini karena meningkatkan efektivitas dan performa dari model. Adapun berikut perbandingan akurasi XGBoost terhadap beberapa skenario sebelumnya:

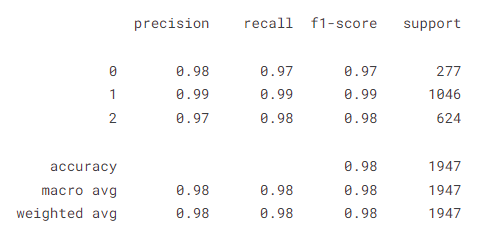


Gambar 3.24 Model XGBoost dari Berbagai Skenario Pengujian

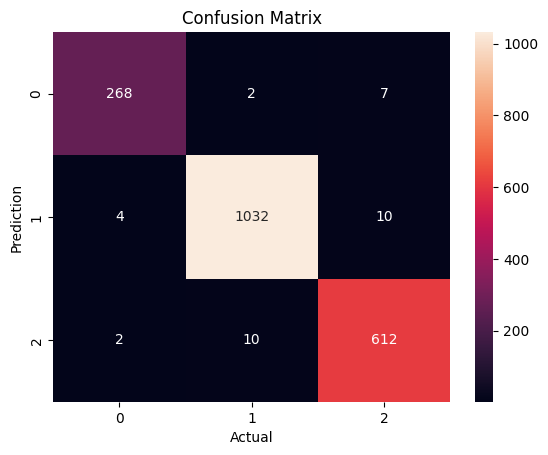
## Analisis dan Pembahasan

Berdasarkan hasil yang telah diperoleh, dapat diketahui bahwa penggunaan Normalisasi sangat dianjurkan untuk permasalahan ini karena adanya kenaikkan akurasi secara signifikan pada 2 metode yakni SVC serta ANN tanpa terjadi penurunan di 2 metode sisanya yaitu Random Forest dan XGBoost yang performanya sudah baik tanpa harus dilakukan *scaling* dengan StandardScaler, kemungkinan disebabkan oleh kemampuan mereka untuk menganalisa data yang sangat mumpuni. Berbeda halnya dengan Seleksi Fitur yang cocok di 2 metode yaitu Random Forest dan XGBoost sehingga menyebabkan peningkatan akurasi, tetapi kurang cocok di 2 metode lainnya yakni SVC serta ANN.

Untuk metode terbaik guna menangani permasalahan ini sendiri bila diamati dari hasil percobaan yang telah dilakukan adalah XGBoost dikarenakan akurasinya yang tinggi. Guna mendapatkan kepastian terkait performa dan efektivitas dari metode XGBoost, berikut ditampilkan juga hasil evaluasi dari testing menggunakan data *test* dalam bentuk *classification report* dan *confusion matrix* sebagai berikut,



Gambar 3.25 *Classification Report* Model XGBoost



Gambar 3.26 *Confusion Matrix* Model XGBoost

Berdasarkan hasil evaluasi proses testing di atas, dapat dilihat bahwa nilai berbagai metriks evaluasi seperti akurasi dan juga f1-scorenya memang tinggi sehingga dapat diketahui pula bahwa metode XGBoost memang benar-benar mumpuni untuk permasalahan ini. Nilai akurasi tinggi yang ditampilkan dari prediksi pada hasil pengujian di atas pun terbukti benar adanya dan bukanlah disebabkan oleh *overfitting* atau penyebab lainnya. Meskipun begitu, performa dari metode-metode lain dari perlakuan-perlakuan sebelumnya juga cukup baik, seperti Random Forest yang bisa mencapai akurasi 97% dan juga SVM serta ANN yang sempat mengalami peningkatan akurasi cukup drastis sehingga kesempatan untuk dilakukannya *hyperparameter tuning* terhadap model-model tersebut sangat mungkin untuk dilakukan.

# 

# **Kesimpulan dan Saran**

## Kesimpulan

Setelah melalui berbagai skenario pengujian dan analisis terhadap hasil yang telah dilakukan, diperoleh kesimpulan bahwa untuk permasalahan prediksi status kelulusan UTBK berdasarkan nilai hasil tes dan juga program studi yang dipilih, metode yang terbaik untuk menanganinya adalah XGBoost. Hal ini terbukti dari nilai evaluasinya yang memang tinggi baik di dataset *training* maupun *testing*. Tentunya dicapainya nilai evaluasi tersebut tidak luput dari berbagai perlakuan seperti Normalisasi, Seleksi Fitur, dan juga *Hyperparameter Tuning* yang juga sekaligus terbukti memiliki dampak yang cukup besar terhadap hasil dari proses pembelajaran mesin. Akan tetapi, dampak yang dibawa bisa ke arah positif yang ditunjukkan dengan peningkatan nilai evaluasi maupun ke arah negatif dengan nilai evaluasi yang justru menurun.

Pada akhirnya, semua tentu bergantung pada masalah yang diangkat, karakteristik dataset yang digunakan, dan juga spesifikasi metode yang dipilih sebagai opsi penyelesaian. Tidak ada metode yang buruk dan tidak ada penanganan data yang salah, bisa dilihat dari performa setiap metode dalam percobaan di atas yang sebenarnya tidak buruk, bahkan bisa menjadi baik bila mendapatkan penanganan yang cukup dan sesuai. Karena dalam pembelajaran mesin sendiri, penanganan yang diberikan bisa berbeda-beda dan harus menyesuaikan dengan permasalahan yang dihadapi serta apa yang harus dilalui guna mencapai penyelesaian dari masalah tersebut agar dapat tercapai hasil yang maksimal. Tidak berhenti di situ, berbagai evaluasi pun harus dilakukan guna memastikan bahwa hasil yang dicapai memang sudah merupakan hasil terbaik ataukah masih ada kesempatan untuk peningkatan kualitas serta performa.

## 

## Saran

Pada penelitian selanjutnya, tentu masih banyak metode klasifikasi yang dapat dieksplorasi untuk topik permasalahan serupa, bahkan bisa juga dilakukan pengembangan terhadap beberapa metode yang sudah disertakan di penelitian kali ini mengingat proses *Hyperparameter Tuning* di akhir hanya dilakukan pada satu metode yang terbaik. Adapun kesempatan meningkatkan nilai evaluasi dari metode-metode di atas juga terbuka lebar.

Di samping itu, dataset yang digunakan juga bisa ditingkatkan levelnya, mungkin dengan dataset berlabel asli yang memang memiliki status kelulusan asli dan bukan hanya bergantung pada nilai semata guna menyamakan dengan kondisi kelulusan di lapangan yang tentunya tidak hanya dipengaruhi oleh nilai. Terlebih lagi mempertimbangkan bahwa setelah tahun 2023 SBMPTN sudah berubah menjadi SNBT, sehingga mungkin perlu dikaji lebih lanjut dan dilakukan beberapa penyesuaian terhadap sistem yang baru.

# **Daftar Pustaka**

[1] T. P. Kiri and A. Atti, “Pengaruh IPK Dan Masa Studi Terhadap Waktu Tunggu Mendapatkan Pekerjaan,” J. Difer., vol. 03, no. 01, pp. 1–12, 2021.

[2] H. Ridwan, I. Darmawati, andM. N. Rahmawati, “Implementasi Kebijakan Physical Distancing dalam Pelaksanaan UTBK SBMPTN di Perguruan Tinggi,” J. Kebijak. Kesehat. Indones. Jkki, vol. 09, no. 04, pp. 197–201, 2020, doi: 10.22146/jkki.59689.

[3] Simarmata, Justin E., et al. "Academic Ability Test Training for 12th Grade Senior High Scool Students for the 2022 Utbk Sbmptn Preparation." Jurnal Abdi Insani Universitas Mataram, vol. 9, no. 2, 2022, pp. 471-479, doi:10.29303/abdiinsani.v9i2.557.

[4] SUTARJO, S. MENINGKATKAN MOTIVASI BELAJAR SISWA MELALUI PENGEMBANGAN TES BAKAT SKOLASTIK DALAM MENGHADAPI UJIAN UTBK DAN SBMPTN DIERA NEW NORMAL COVID-19. LEARNING : Jurnal Inovasi Penelitian Pendidikan Dan Pembelajaran, 2022, pp. 83-94,https://doi.org/10.51878/learning.v2i1.1006.

[5] Aditya, A. and Purwiantono, F.E., The Application of Fuzzy-Analytical Hierarchy Process Method for Majors Selection at Public Universities. Journal Of Informatics And Telecommunication Engineering, 3(2), 2020, pp.240-251.

[6] Midaada, A. (2019) 'Universitas Brawijaya jadi favorit peserta SBMPTN 2019,' https://edukasi.okezone.com/, 27 Juni. https://edukasi.okezone.com/read/2019/06/27/65/2071732/universitas-brawijaya-jadi-favorit-peserta-sbmptn-2019.

[7] Munoz, A., Machine learning and optimization. Courant Institute of Mathematical Sciences, 2014, pp.1-2.

[8] P. Ongsulee, "Artificial intelligence, machine learning and deep learning," 2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE), Bangkok, Thailand, 2017, pp. 1-6, doi: 10.1109/ICTKE.2017.8259629.

[9] Probst, P., Wright, M.N. and Boulesteix, A.L., 2019. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: data mining and knowledge discovery, 9(3), p.e1301.

[10] Rigatti, S.J., 2017. Random forest. Journal of Insurance Medicine, 47(1), pp.31-39.

[11] Cui, L., Hu, C., Zou, Y. and Meng, M.Q.H., 2010, June. Bleeding detection in wireless capsule endoscopy images by support vector classifier. In The 2010 IEEE international conference on information and automation (pp. 1746-1751). IEEE.

[12] Ridha, H.S., 2014. Constructing support vector classifier depending on the golden support vector. Basrah Journal of Agricultural Sciences, 40(2).

[13] Gupta, N., 2013. Artificial neural network. Network and Complex Systems, 3(1), pp.24-28.

[14] Dongare, A.D., Kharde, R.R. and Kachare, A.D., 2012. Introduction to artificial neural network. International Journal of Engineering and Innovative Technology (IJEIT), 2(1), pp.189-194.