**ANALISIS UNBALANCE DENGAN TEKNIK UNDERSAMPLING DAN OVERSAMPLING UNTUK AKURASI NAÏVE BAYES**

**SKRIPSI**

Diajukan sebagai Salah Satu Syarat untuk Mendapatkan

Gelar Sarjana Komputer (S.Kom) Program Studi Informatika



# RENY RAHAYU. S 105841107320

**PROGRAM STUDI INFORMATIKA**

**FAKULTAS TEKNIK**

**UNIVERSITAS MUHAMMADIYAH MAKASSAR 2024**

**ABSTRAK**

**RENY RAHAYU.S.** Analisis Unbalance dengan Teknik Undersampling dan Oversampling untuk Akurasi Naïve Bayes (dibimbing Titin Wahyuni S.Pd.,M.T dan Fahrim Irhamna Rachman S.Kom.,M.T.).

Ketidakseimbangan dataset adalah kondisi dimana distribusi kelas dalam dataset tidak seimbang, dengan kelas mayoritas memiliki jumlah data yang jauh lebih banyak dibandingkan dengan kelas minoritas, yang dapat menyebabkan kesalahan dalam hasil klasifikasi dimana data kelas minoritas sering diklasifikasikan sebagai kelas mayoritas. Tujuan dari penelitian ini adalah untuk memahami dampak ketidakseimbangan data terhadap akurasi model Naïve Bayes. Model Naïve Bayes yang dimaksud dalam penelitian ini adalah salah satu algoritma klasifikasi berbasis probabilitas yang sederhana namun efektif, yang didasarkan pada Teorema Bayes. Algoritma ini mengasumsikan bahwa fitur-fitur dalam dataset bersifat independen satu sama lain, sebuah asumsi yang jarang benar dalam kenyataan tetapi sering kali menghasilkan hasil yang cukup baik dalam praktik. Hasil penelitian menunjukkan bahwa keseimbangan pada dataset sangat mempengaruhi akurasi model Naïve Bayes. Teknik undersampling menghasilkan akurasi tertinggi sebesar 96,6%, sementara teknik oversampling mencapai akurasi maksimal 92%, dan penggunaan SMOTE (Synthetic Minority Over-sampling Technique) memberikan hasil yang baik dengan akurasi tertinggi 91,8%. Kesimpulan dari penelitian ini adalah bahwa teknik undersampling cenderung mengurangi akurasi dengan menghilangkan informasi penting dari kelas mayoritas, sedangkan teknik oversampling berisiko menyebabkan overfitting. Oleh karena itu, pemilihan metode penyeimbangan harus dilakukan dengan hati-hati, menyesuaikan sifat data dan tujuan analisis, serta dievaluasi dengan cermat untuk mencapai performa model yang optimal.

**Kata Kunci** : Analisis Unbalance, Naïve Bayes, Oversampling, SMOTE, Undersampling.

***ABSTRACT***

**RENY RAHAYU.S.** *Unbalance Analysis with Undersampling and Oversampling Techniques for Naïve Bayes Accuracy (supervised by Titin Wahyuni ​​S.Pd., M.T. and Fahrim Irhamna Rachman S.Kom., M.T.).*

*Dataset imbalance is a condition where the distribution of classes in a dataset is unbalanced, with the majority class having a much larger amount of data compared to the minority class, which can cause errors in classification results where minority class data is often classified as the majority class. The aim of this research is to understand the impact of data imbalance on the accuracy of the Naïve Bayes model. The Naïve Bayes model referred to in this research is a simple but effective probability-based classification algorithm, which is based on Bayes' Theorem. This algorithm assumes that the features in the dataset are independent of each other, an assumption that is rarely true in reality but often produces quite good results in practice. The research results show that balance in the dataset greatly influences the accuracy of the Naïve Bayes model. The undersampling technique produces the highest accuracy of 96.6%, while the oversampling technique achieves a maximum accuracy of 92%, and the use of SMOTE (Synthetic Minority Over-sampling Technique) provides good results with the highest accuracy of 91.8%. The conclusion of this research is that undersampling techniques tend to reduce accuracy by removing important information from the majority class, while oversampling techniques risk causing overfitting. Therefore, the selection of balancing methods must be done carefully, adapting to the nature of the data and analysis objectives, and carefully evaluated to achieve optimal model performance.*

***Keywords :*** *Unbalance Analysis, Naïve Bayes, Oversampling, SMOTE, Undersampling.*

# KATA PENGANTAR

Alhamdulillah puji dan syukur kepada Allah SWT yang senantiasa melimpahkan rahmat dan hidayah-Nya kepada penulis karena telah memberikan kesehatan,kesempatan, kemudahan serta melapangkan pemikiran sehingga penulis dapat menyelesaikan proposal skripsi yang berjudul **“ ANALISIS UNBALANCE DENGAN TEKNIK UNDERSAMPLING DAN OVERSAMPLING UNTUK**

**AKURASI NAÏVE BAYES’’.** Shalawat beserta salam senantiasa tercurahkan kepada kekasih Allah SWT panutan seluruh umat yakni baginda Rasulullah SAW, yang telah memperbaiki akhlak dan budi pekerti manusia seperti yang kita rasakan sekarang ini.

Penulis dapat menyelesaikan Proposal Skripsi ini yang merupakan salah satu syarat agar bisa mendapatkan gelar Sarjana (S1) di Universitas Muhammadiyah Makassar. Proses pengerjaan proposal skripsi ini tidak lepas dari bimbingan dan bantuan dari berbagai pihak, penulis menyadari bahwa di dalam penulisan proposal skripsi ini masih jauh dari kata sempurna baik dari segi isi, Bahasa maupun dari segi penulisannya.Untuk itu penulis sangat mengharapkan kritik dan saran yang membangun untuk kemajuan masa yang akan datang.

Oleh karena itu, pada kesempatan kali ini penulis ingin mengucapkan banyak terima kasih kepada segenap pihak yang membantu khususnya kepada :

1. Superhero,Panutanku dan cinta pertamaku, Ayahanda Suardi Dg. Sijaya terima kasih sudah berjuang untuk penulis walaupun itu penulis dan beliau tidak tinggal Bersama lagi. Beliau memang tidak sempat merasakan Pendidikan sampai bangku perkuliahan, Namun beliau mampu mendidik penulis, memberikan dukungan.
2. Untuk Ayahanda tercinta, Susandi Samma terima kasih telah mendidik penulis sangat baik dan memberikan dukungan serta motivasi sampai tibanya penulis akhirnya mampu menyelsaikan studinya sampai sarjana.
3. Pintu surgaku,wanita Tangguh dan hebatku, Ibunda Nirwana, terima kasih atas perjuangan dan pengorbanan yang tak ternilai ibunda kepada penulis yang tidak henti-hentinya memberikan kasih sayang dengan penuh cinta dan selalu memberikan motivasi serta doa yang selalu diarahkan kepada Allah untuk

kesuksesan penulis hingga penulis mampu menyelesaikan studinya sampai sarjana.

1. Bapak Prof. Dr.. H. Ambo Asse, M.Ag., sebagai Rektor Perguruan Tinggi Universitas Muhammadiyah Makassar.
2. Ibu Dr. Hj. Ir. Nurnawaty ST.,MT Selaku Dekan Fakultas Teknik Universitas Muhammadiyah Makassar.
3. Bapak Muhyidin A.M.Hayat S.Kom.,MT Selaku Ketua Prodi Informatika Fakultas Teknik Universitas Muhammadiyah Makassar.
4. Ibu Titin Wahyuni S.Pd.,MT Selaku Dosen Pembimbing I Dan Bapak Fahrim Irhamna Rachman, S.Kom., MT Selaku Dosen Pembimbing II yang senantiasa meluangkan waktunya,pikirannya untuk memberikan bimbingan serta mengarahkan penulis selama penyusunan proposal skripsi ini .
5. Segenap Bapak/Ibu Dosen Prodi Informatika Fakultas Teknik Universitas Muhammadiyah Makassar yang telah banyak memberikan bakat dan ilmu pengetahuan serta mendidik penulis selama proses belajar mengajar di Universitas Muhammadiyah Makassar.
6. Rekan-rekan mahasiswa utamanya dari Program studi Informatika Angkatan 2020 Fakultas Teknik Universitas Muhammadiyah Makassar terima kasih atas dukungan dan kerjasamanya selama menempuh Pendidikan serta penyelesaian penyususan proposal skripsi ini.
7. Terakhir, terima kasih untuk diri sendiri, karena telah mampu berusaha keras dan berjuang sampai sejauh ini.

Semoga semua pihak yang membantu semoga mendapatkan pahala yang berlipat ganda disisi Allah SWT dan Proposal Skripsi yang sederhana ini dapat bermanfaat bagi penulis, rekan- rekan. Masyarakat serta Bangsa dan Negara. Aamiin**.**

Billahi Fisabilhaq, Fastabiqul Khairat. Wassalamualaikum Wr. Wb

Makassar , 15 Maret 2024 Reny Rahayu. S

**DAFTAR ISI**

ABSTRAK ii

ABSTRACT iii

KATA PENGANTAR iv

DAFTAR ISI vi

DAFTAR GAMBAR viii

DAFTAR TABEL ix

DAFTAR LAMPIRAN x

DAFTAR ISTILAH xi

BAB I PENDAHULUAN 1

1. Latar Belakang 1
2. Rumusan Masalah 2
3. Tujuan Penelitian 2
4. Manfaat Penelitian 2
5. Ruang Lingkup Penelitian 3
6. Sistematika Penulisan 3

BAB II TINJAUAN PUSTAKA 5

1. Landasan Teori 5
2. Penelitian Terkait 14
3. Kerangka Pikir 20

BAB III METODE PENELITIAN 21

1. Tempat dan Waktu Penelitian 21
2. Alat dan Bahan 21
3. Perancangan Sistem 21
4. Teknik Pengujian Sistem 25
5. Teknik Analisis Data 25

BAB IV HASIL DAN PEMBAHASAN 27

1. Pengambilan Data 27
2. Normalisasi Data 34
3. Pelabelan Data Mahasiswa 35
4. Penerapan Naive Bayes 37
5. Hasil Pengujian Data Tanpa Menggunakan Smote 45
6. Hasil Pengujian Data Dengan Menggunakan Smote 48
7. Perhitungan Manual 49
8. Perbandingan Pengujian Menggunakan SMOTE dan Tanpa Mengg unakan SMOTE 52

BAB V KESIMPULAN DAN SARAN 55

1. Kesimpulan 55
2. Saran 55

DAFTAR PUSTAKA 56

LAMPIRAN 59

# DAFTAR GAMBAR

Gambar 1. Kerangka Pikir 22

Gambar 2. Perancangan Sistem 25

Gambar 3. Diagram Alur Penelitian 26

# DAFTAR TABEL

Tabel 1. Data Mahasiswa Angakatan 2018 27

Tabel 2. Data Mahasiswa Angakatan 2019 28

Tabel 3. Data Mahasiswa Angakatan 2020 30

Tabel 4. Data Mahasiswa Angakatan 2021 31

Tabel 5. Data Mahasiswa Angakatan 2022 32

Tabel 6. Data Mahasiswa Angakatan 2023 32

Tabel 7. Min – Max 34

Tabel 8. Dataset 35

Tabel 9. Undersampling Data 200 46

Tabel 10. Undersampling Data 300 46

Tabel 11. Undersampling Data 400 46

Tabel 12. Oversampling Data 500 47

Tabel 13. Oversampling Data 600 47

Tabel 14. Oversampling Data 700 48

Tabel 15. Hasil Pengujian Data Dengan Menggunakan Smote 49

Tabel 16. Perhitungan Manual Undersampling 200 49

Tabel 17. Perhitungan Manual Undersampling 300 50

Tabel 18. Perhitungan Manual Undersampling 400 50

Tabel 19. Perhitungan Manual Oversampling 500 51

Tabel 20. Perhitungan Manual Oversampling 600 51

Tabel 21. Perhitungan Manual Oversampling 700 51

Tabel 22. Perhitungan Manual Menggunakan Smote 52

Tabel 22. Hasil Perbandingan 53

# DAFTAR LAMPIRAN

Lampiran 1. Data Mentah Angakatan 2018 59

Lampiran 2. Data Mentah Angakatan 2019 60

Lampiran 3. Data Mentah Angakatan 2020 62

Lampiran 4. Data Mentah Angakatan 2021 63

Lampiran 5. Data Mentah Angakatan 2022 64

Lampiran 6. Data Mentah Angakatan 2023 65

Lampiran 7. Data Mahasiswa Setelah Dilakukan Pelabelan Data 66

Lampiran 8. Source Code Naïve Bayes Menggunakan SMOTE 67

Lampiran 9. Hasil Akurasi Menggunakan SMOTE 68

Lampiran 10. Permohonan Penelitian Kepada Kaprodi Informatika 69

Lampiran 11. Pengantar Penelitian Kepada Ketua Umum LP3M

Unismuh Makassar 70

Lampiran 12. Hasil Scan Plagiasi Per Bab 71

# DAFTAR ISTILAH

*Machine Learning* Adalah cabang dari kecerdasan buatan yang

memungkinkan computer untuk belajar dari data tanpa harus diprogram secara eksplisit.

*Flowchart* Adalah representasi grafis dari algoritma atau proses. Flowchart menggunakan simbol-simbol seperti kotak, oval, panah, dan berbagai bentuk geometris lainnya untuk menunjukkan langkah-langkah yang harus diambil dalam suatu proses atau algoritma.

Klasifikasi Adalah tugas untuk mengelompokkan instance (data) ke dalam kelas atau kategori yang telah ditentukan berdasarkan atribut-atributnya.

*Naive Bayes Classifier* Adalah jenis algoritma klasifikasi yang berdasarkan

teorema Bayes dengan asumsi independensi antara fitur-fitur yang diberikan. Meskipun sederhana, Naive Bayes sering kali efektif dalam banyak aplikasi klasifikasi.

*Python* Adalah bahasa pemrograman tingkat tinggi yang sangat populer dan sering digunakan dalam pengembangan aplikasi web, analisis data, machine learning, dan berbagai bidang lainnya.

*Undersampling* Adalah teknik untuk mengurangi jumlah sampel dari

kelas mayoritas dalam data yang tidak seimbang sehingga setiap kelas memiliki jumlah sampel yang lebih serupa.

*Oversampling* Adalah teknik untuk meningkatkan jumlah sampel

dari kelas minoritas dalam data yang tidak seimbang

dengan membuat salinan data yang sudah ada atau dengan menambahkan data sintetis.

*Unbalance* Unggul dalam hal jumlah, atau lebih banyak, seperti kelas mayoritas dalam masalah klasifikasi yang tidak seimbang.

*Mayoritas* Kelas atau kategori yang memiliki jumlah instance yang lebih besar dalam dataset, terutama dalam konteks data yang tidak seimbang.

*Minoritas* Kelas atau kategori yang memiliki jumlah instance yang lebih kecil dalam dataset, terutama dalam konteks data yang tidak seimbang.

*SMOTE* Adalah teknik oversampling yang populer untuk menangani ketidakseimbangan kelas dengan membuat sampel sintetis berdasarkan sampel minoritas yang ada.

*Overfitting* Terjadi ketika model machine learning terlalu kompleks sehingga dapat "menghafal" data pelatihan dengan sangat baik, tetapi tidak mampu melakukan generalisasi dengan baik pada data baru yang belum pernah dilihat sebelumnya.

*Supervised Learning* Adalah tipe pembelajaran mesin di mana model

belajar dari contoh data yang berlabel. Model ini diberi input-output pairs, dan tujuannya adalah untuk mempelajari fungsi pemetaan dari input ke output.

*Imblearn* Adalah pustaka Python yang berfokus pada teknik- teknik untuk menangani masalah klasifikasi dengan data yang tidak seimbang, seperti undersampling, oversampling, dan SMOTE

# BAB I PENDAHULUAN

# Latar Belakang

Ketidakseimbangan dataset adalah keadaan dimana distribusi kelas didalam dataset tidak seimbang. Sebuah kelas dikatakan tidakseimbang apabila ada suatu kelas yang memiliki data yang lebih banyak dibandingkan dengan kelas lainnya. Kelompok kelas dengan jumlah data yang banyak disebut dengan kelas mayoritas, sedangkan kelompok kelas dengan jumlah yang sedikit disebut dengan kelas minoritas. Dataset yang tidak seimbang menyebabkan misleading atau kesesatan dalam hasil klasifikasi dimana data kelas minoritas sering diklasifikasikan sebagaikelas mayoritas.Penerapan algoritma klasifikasi tanpa memperhatikan keseimbangan kelas mengakibatkan prediksi yang baik bagi kelas mayoritas dan kelas minoritas diabaikan.Apabila algoritma klasifikasi di implementasikan langsung terhadap dataset yang imbalance maka akan mengalami penurunan performa (Sulistiyono et al., 2021).

Masalah Ketidakseimbangan ini bisa menjadi masalah dalam pembelajaran mesin karena model yang dibuat mungkin cenderung lebih baik dalam memprediksi kelas mayoritas, sementara kelas minoritas diabaikan (Technology, 2023). Dalam hal ini Masalah ketidakseimbangan kelas dalam dataset dapat menjadi kendala serius dalam pengembangan model. Model Machine Learning yang digunakan yaitu Algoritma Naïve bayes. Naïve Bayes merupakan pengklasifikasi probabilitas sederhana untuk menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari kumpulan data yang diberikan (Ericha Apriliyani & Salim, 2022).

Dalam penelitian ini, peneliti menggunakan Teknik Under Sampling dan Oversampling dalam menangani Unbalance atau Ketidakseimbangan dataset sebelum melatih model Naïve Bayes. Pada Teknik Under Sampling, data yang mayoritas dikurangi sehingga jumlahnya sama dengan kelas minoritas, sedangkan oversampling mengenerate data baru untuk kelas minoritas sehingga

jumlahnya seimbang dengan kelas mayoritas (Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti et al., 2017).

Berdasarkan pernyataan diatas penelitian ini diharapkan dapat memberikan solusi dalam mengatasi ketidakseimbangan kelas dalam dataset dengan menerapkan teknik under sampling dan over sampling guna meningkatkan akurasi model Naïve Bayes.

# Rumusan Masalah

Berdasarkan pada latar belakang masalah di atas, maka rumusan masalah pada penelitian ini yaitu “Bagaimana ketidakseimbangan (unbalance) pada dataset mempengaruhi akurasi model Naïve Bayes?”

# Tujuan Penelitian

Berdasarkan pada latar belakang masalah, rumusan masalah di atas, maka Adapun tujuan yang ingin dicapai dari penelitian ini yaitu : Adalah untuk memahami ketidakseimbangan data pada akurasi model Naïve Bayes?

# Manfaat Penelitian

Berdasarkan pada uraian latar belakang , rumusan masalah dan tujuan penelitian di atas, maka manfaat penelitian ini dibagi menjadi bebera aspek sebagai berikut :

* 1. Aspek Teoritis

Memberikan pemahaman yang lebih mendalam tentang konsep dasar Naïve Bayes dalam konteks klasifikasi data..

* 1. Aspek Praktis
     1. Bagi Peneliti
        1. Dapat menghasilkan publikasi ilmiah yang dapat meningkatkan reputasi peneliti dalam komunitas akademis.
        2. Akan memperluas pemahaman peneliti tentang analisis data, model klasifikasi, dan teknik penanganan ketidakseimbangan data.
        3. Dapat memperluas jaringan profesionalnya dengan berkolaborasi dengan rekan peneliti lainnya yang tertarik dalam bidang analisis data dan machine learning.
     2. Bagi Universitas
        1. Universitas dapat membantu mengembangkan keterampilan dan pengetahuan mereka dalam bidang analisis data dan machine learning.
        2. Memberikan kontribusi pada pengetahuan dan pemahaman dalam bidang analisis data, klasifikasi, dan penanganan ketidakseimbangan data, yang merupakan aset bagi universitas dalam mendukung pengajaran, penelitian, dan layanan masyarakat.

# Ruang Lingkup Penelitian

Dari analisis rumusan masalah di atas dapat dirumuskan beberapa batasan masalah yaitu :

* 1. Fokus pada Model Naïve Bayes
  2. Penggunaan Teknik Under Sampling dan Oversampling
  3. Evaluasi pada Akurasi Model, pada penelitian ini akan membatasi analisis pada pengukuran akurasi model Naïve Bayes sebelum dan setelah penerapan teknik under sampling dan oversampling.
  4. Dataset, dalam penelitian ini akan menggunakan dataset tertentu yang relevan dengan tujuan analisis dan memiliki masalah ketidakseimbangan kelas yang signifikan.
  5. Data yang digunakan diambil dari data mahasiswa Fakultas Teknik Prodi Pengairan Universitas Muhammadiyah Makassar, dan penelitian ini membutuhkan data mengenai mahasiswa, seperti Jenis Kelamin, IPK, SKS, Lama Studi dan Mata kuliah wajib maupun tambahan yang berkaitan dengan Tingkat kelulusan mahasiswa. Data ini diperoleh dari Simak Fakultas.
  6. Tidak Memperhitungkan Teknik Lain.
  7. Tidak Memperhitungkan Variasi dalam Teknik

# Sistematika Penulisan

Untuk memberikan gambaran umum dari seluruh penulisan ini, Adapun sistematika penulisan yaitu :

BAB I PENDAHULUAN

Bab ini menerangkan secara singkat dan jelas mengenai latar belakang penulisan penelitian tugas akhir, rumusan masalah, tujuan dan manfaat, Batasan permasalahan, metodologi yang digunakan dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Pada bab ini membahas tentang teori – teori yang melandasi penulis dalam melaksanakan skripsi.

BAB III METODE PENELITIAN

Membahas tentang metode penelitian dan alat yang digunakan untuk pembuatan system.

BAB IV ANALISA DAN PENGUJIAN

Bab ini berisikan hasil desain system serta pembahasan terhadap desain tersebut.

BAB V KESIMPULAN DAN SARAN

Bab ini atau bab terakhir akan memuat Kesimpulan isi dari kesulurahan uraian bab sebelumnya dan saran – saran dari hasil yang telah diperoleh serta yang diharapkan daoat bermanfaat dalam pengembangan selanjutnya

# BAB II

**TINJAUAN PUSTAKA**

# Landasan Teori

* 1. Pengertian Analisis

Analisa berasal dari kata Yunani Kuno “analusis” yang berarti melepaskan. Analusis terbentuk dari dua suku kata yaitu “ana” yang berarti kembali dan “luein” yang berarti melepas. Sehingga pengertian analisa yaitu suatu usaha dalam mengamati secara detail pada suatu hal atau benda dengan cara menguraikan komponen-komponen pembentuknya atau menyusun komponen tersebut untuk dikaji lebih lanjut. Analisa adalah sebuah proses untuk memecahkan sesuatu ke dalam bagian-bagian yang saling berkaitan satu sama lainnya. merupakan suatu kegiatan berfikir untuk menguraikan suatu keseluruhan menjadi komponen sehingga dapat mengenal tanda-tanda dari setiap komponen, hubungan satu sama lain dan fungsi masing-masing dalam suatu keseluruhan yang terpadu.(Wicaksana & Rachman, 2018) Pengertian analisis adalah memecahkan atau menguraikan sesuatu unit menjadi unit terkecil.

Dari pendapat diatas dapat ditarik kesimpulan bahwa analisis merupakan suatu kegiatan berfikir untuk menguraikan atau memecahkan suatu permaslaahan dari unit menjadi unit terkecil (Septiani et al., 2020).

* 1. Teknik Undersampling

Teknik undersampling merupakan proses sampling yang dilakukan dengan mengurangi atau mengeliminasi sebagian data pada kelas mayoritas pada data. Proses eliminasi tersebut dapat dilakukan secara random (paling sederhana) sehingga biasa disebut dengan random undersampling. Selain itu, undersampling juga dapat dilakukan dengan menggunakan perhitungan statistik yang biasa disebut dengan informed undersampling. Pada teknik ini, metode iterasi dan teknik data cleaning juga diterapkan untuk menyaring data pada kelas mayoritas lebih lanjut. Metode undersampling menyebabkan beberapa informasi pada kelas negatif terhapus dan proporsi jumlah data pada kelas mayoritas serta minoritas lebih berimbang. Teknik eliminasi data yang diusulkan pada metode undersampling sangat bermacam-macam. Terdapat metode yang focus menghapus pada area kelas mayoritas saja ,ada pula yang berfokus pada kedua kelas namun hanya di area border saja, dan lain sebagainya (Choirunnisa, 2019).

* 1. Teknik Oversampling

Berkebalikan dengan Teknik Undersampling, Teknik Oversampling justru merupakan metode sampling dengan menambahkan jumlah data pada kelas minoritas sehingga dapat mengimbangi atau mendekati jumlah data pada kelas mayoritas. Konsep penambahan data pada oversampling dibagi menjadi dua yaitu: oversampling menggunakan data asli, seperti metode Random Oversampling dan yang kedua yaitu metode penambahan menggunakan data sintetik seperti *Synthetic Minority Over-sampling Technique* (SMOTE) SMOTE Borderline, Safe Level SMOTE, Adaptive *Semi-unsupervised Weighted Oversampling* (A-SUWO) Namun pada penelitian ini, penulis hanya berfokus pada oversampling menggunakan metode penambahan data secara sintetik dimana metode state-of-the-art yang dipilih ialah metode A-SUWO.

Permasalahan yang umum pada oversampling adalah terjadinya overfitting dikarenakan penambahan data secara berulang menyebabkan decision boundary menjadi lebih ketat. Oleh karena itu, pada perkembangannya, metode oversampling bukan lagi mengopi data yang sama tetapi membuat data baru yang mirip (Choirunnisa, 2019)

* 1. Naïve Bayes

Naive Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma mengunakan teorema Bayes dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas. Definisi lain mengatakan Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya.

Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Keuntungan penggunaan Naive Bayes adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (Training Data) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. Naive Bayes sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan.Berikut Persamaan Metode Naive Bayes.

Persamaan dari teorema Bayes adalah :

P(H|X) = (1)

Di mana :

X : Data dengan class yang belum diketahui

H : Hipotesis data merupakan suatu class spesifik

P(H|X) : Probabilitas hipotesis H berdasar kondisi X (posteriori probabilitas)

P(H) : Probabilitas hipotesis H (prior probabilitas)

P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H P(X) Probabilitas X

Untuk menjelaskan metode Naive Bayes, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut.

Penerapan Metode Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Keuntungan penggunaan Naive Bayes adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (Training Data) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. Klasifikasi Naive Bayes merupakan klasifikasi yang bersifat supervised learning karena memiliki supervisor (manusia melakukan klasifikasi secara manual pada data yang digunakan dalam pelatihan) selaku pengajar dalam proses belajar atau learning . Selain itu, performansi Naive Bayes memiliki waktu klasifikasi yang singkat sehingga mempercepat proses sistem analisis sentimen. Pada penelitian ini pengujian digunakan dengan menggunakan) 3 kelas (negatif, netral, dan positif).

Dan dalam metode Naive Bayes data String yang bersifat konstan dibedakan dengan data numerik yang bersifat kontinyu, perbedaan ini akan terlihat pada saat menentukan nilai probabilitas setiap kriteria baik itu kriteria dengan nilai data string maupun kriteria dengan nilai data numerik.(Kawani, 2019)

* 1. Klasifikasi

Klasifikasi merupakan proses menemukan model atau fungsi yang menggambarkan, membedakan class data atau konsep dengan tujuan agar bisa digunakan untuk class prediction dari objek yang label class tidak diketahui. Klasifikasi banyak digunakan untuk mendeteksi fraud atau penipuan, target pemasaran, prediksi kinerja, manufaktur, dan mendiagnosa Kesehatan. Tahapan klasifikasi data terdiri dari 2 langkah. Pertama yaitu tahap learning atau fase pembelajaran, dimana algoritma klasifikasi untuk menganalisis data training atau data latih lalu direpresentasikan ke bentuk model klasifikasi. Kedua adalah proses tahapan klasifikasi, dimana datatesting digunakan untuk memprediksi nilai accuracy dari model klasifikasi. Jika nilai akurasi acceptable atau dapat diterima, maka rule bisa diterapkan pada klasifikasi tupel data baru (Heranova, 2019).

Klasifikasi adalah teknik mempelajari kumpulan data sehingga menghasilkan aturan yang dapat mengenali data baru yang belum dipelajari, Klasifikasi dapat didefinisikan sebagai proses pemetaan objek data menjadi kategori dari salah satu class yang telah ditentukan sebelumnya (Mutmainah, 2021).

* 1. Supervised Learning

*Supervised learning* merupakan salah satu metode untuk mengklasifikasikan masing masing objek dalam data ke beberapa kelas. Pada supervised learning setiap objek pada suatu data memiliki fitur, yaitu ciri-ciri yang ada pada masing-masing objek. Setiap objek dalam suatu data memiliki jumlah fitur yang sama. Fitur digunakan sebagai input untuk menentukan kelas pada objek. Dalam supervised learning, kelas dari masing-masing objek sudah diketahui. Oleh karena itu, permasalahan yang dihadapi dalam supervised learning adalah bagaimana memetakan objek ke dalam kelas yang tepat menggunakan fitur-fitur yang dimiliki oleh setiap objek.(Hari et al., 2020)

*Supervised Learning* merupakan suatu pendekatan dimana sistem dilatih terlebih dahulu agar dapat melakukan prediksi atau melakukan klasifikasi Supervised Learning merupakan metode yang memiliki data latih yang terdiri dari pasangan input dan output yang diinginkan dan bertujuan mempelajari pemetaan antara ruang input dan output. Supervised Learning bertujuan untuk menemukan pola baru dalam data dengan cara menghubungkan pola data yang sudah ada dengan data yang baru.metode klasifikasi atau biasa disebut sebagai supervised learning adalah teknik pengumpulan data sepenuhnya diberikan label untuk mengklasifikasikan kelas yang tidak dikenal.

Ada banyak algoritma-algoritma dalam supervised learning, beberapa diantaranya adalah C4.5, *K-Nearest Neighbor* *(KNN),* *Naive Bayes Classifier (NBC)*, *Artificial Neural Network (ANN), Probabilistic Neural Network (PNN) dan Self-Organizing Map (SOM).(*Prihandari, 2022)

* 1. Imblearn Library Undersampling

*Imbalanced-learn*, atau yang lebih dikenal sebagai imblearn, adalah pustaka Python yang digunakan untuk menangani masalah ketidakseimbangan kelas dalam data. Salah satu teknik yang diterapkan dalam imblearn adalah undersampling, yang bertujuan untuk mengurangi jumlah sampel dari kelas mayoritas sehingga seimbang dengan kelas minoritas.

Undersampling adalah pendekatan yang digunakan untuk menyeimbangkan dataset dengan mengurangi jumlah sampel dari kelas mayoritas sehingga sebanding dengan jumlah sampel dari kelas minoritas. Ini membantu mencegah model pembelajaran mesin dari menjadi bias terhadap kelas mayoritas dan meningkatkan kinerja prediktif pada kelas minoritas yang sering kali lebih penting. Salah satu cara menangani dataset yang tidak seimbang adalah dengan mengurangi jumlah observasi dari semua kelas kecuali kelas minoritas. Kelas minoritas adalah kelas yang jumlah observasinya paling sedikit. Algoritme yang paling terkenal dalam kelompok ini adalah random undersampling, di mana sampel dari kelas target diambil secara acak.

Berikut adalah penjelasan langkah-langkah umum dalam melakukan undersampling menggunakan imblearn:

* + 1. Identifikasi Kelas Mayoritas dan Kelas Minoritas: Pertama, perlu mengidentifikasi kelas mayoritas (kelas dengan jumlah sampel yang lebih besar) dan kelas minoritas (kelas dengan jumlah sampel yang lebih kecil).
    2. Pilih Teknik Undersampling**:** imblearn menyediakan beberapa teknik undersampling yang dapat dipilih, seperti RandomUnderSampler, NearMiss, dan lain-lain. Setiap teknik memiliki cara tersendiri untuk memilih sampel mana yang akan dihapus dari kelas mayoritas.
    3. Terapkan Teknik Undersampling: Setelah memilih teknik yang sesuai, selanjutnya dapat menerapkannya pada dataset dan menggunakan fungsi yang disediakan oleh imblearn. Biasanya, sehingga perlu menyediakan parameter seperti proporsi sampel yang ingin dipertahankan atau jumlah absolut dari kelas minoritas yang ingin dicapai.

Berikut adalah contoh kode sederhana menggunakan imblearn untuk melakukan undersampling dengan RandomUnderSampler:

from imblearn.under\_sampling import

RandomUnderSampler

# Membuat objek RandomUnderSampler undersampler = RandomUnderSampler()

# Melakukan undersampling pada dataset

X\_resampled, y\_resampled = undersampler.fit\_resample(X, y)

Dalam contoh ini, X adalah matriks fitur dan y adalah vektor label dari dataset Anda. Setelah proses undersampling, X\_resampled dan y\_resampled akan berisi dataset yang telah diundersample.

* + 1. Evaluasi Hasil: Setelah melakukan undersampling, penting untuk mengevaluasi kembali dataset terlebih dahulu. Yang dimana ingin memastikan bahwa kelas sekarang seimbang dan tidak ada informasi penting yang hilang dalam proses undersampling. sehingga dapat menggunakan metrik evaluasi seperti akurasi, presisi, recall, atau F1- score untuk mengevaluasi kinerja model pada dataset yang telah diundersample.
  1. Imblearn Library Oversampling

*Imbalanced-learn* (atau *imblearn*) adalah pustaka Python yang berguna untuk menangani masalah ketidakseimbangan kelas dalam dataset. Salah satu fitur utamanya adalah kemampuannya untuk melakukan oversampling, yaitu teknik yang digunakan untuk menyeimbangkan kembali distribusi kelas dengan meningkatkan jumlah sampel dari kelas minoritas. Salah satu cara untuk mengatasi masalah ini adalah dengan menghasilkan sampel baru di kelas-kelas yang kurang terwakili. Strategi yang paling naif adalah menghasilkan sampel baru dengan pengambilan sampel secara acak.

Berikut adalah beberapa poin penting yang perlu dipahami tentang penggunaan oversampling dengan imbalanced-learn:

* + 1. Penanganan Masalah Ketidakseimbangan: Masalah ketidakseimbangan kelas terjadi ketika jumlah sampel dalam satu atau beberapa kelas dalam dataset secara signifikan lebih rendah daripada kelas lainnya. Hal ini dapat menyebabkan model pembelajaran mesin cenderung memprediksi kelas mayoritas dengan lebih baik daripada kelas minoritas.
    2. Oversampling: Oversampling adalah pendekatan untuk menyeimbangkan kembali dataset dengan meningkatkan jumlah sampel dari kelas minoritas. Dengan cara ini, proporsi antara kelas minoritas dan mayoritas menjadi lebih seimbang.

Berbagai Metode Oversampling: imbalanced-learn menyediakan berbagai metode oversampling yang dapat dipilih sesuai dengan kebutuhan:

* + - 1. RandomOverSampler: Menggandakan atau mengulang sampel- sampel dari kelas minoritas secara acak.
      2. SMOTE *(Synthetic Minority Over-sampling Technique)*: Membangkitkan sampel sintetis baru untuk kelas minoritas dengan mengambil titik tengah antara sampel-sampel yang sudah ada.
      3. ADASYN *(Adaptive Synthetic Sampling*): Mirip dengan SMOTE, tetapi menimbang kelas minoritas berdasarkan kompleksitas lokal, sehingga lebih banyak sampel sintetis dibangkitkan di daerah yang kurang diwakili.
    1. Menerapkan Teknik Oversampling: Setelah memilih teknik yang sesuai, sehingga dapat membuat objek dari kelas tersebut dan menerapkannya pada dataset,selanjutnya menggunakan metode fit\_resample(). Sebagai contoh:

from imblearn.over\_sampling import RandomOverSampler oversampler = RandomOverSampler()

X\_resampled, y\_resampled = oversampler.fit\_resample(X, y)

Di sini, X adalah matriks fitur dan y adalah vektor label dari dataset Anda. Setelah proses oversampling, X\_resampled dan y\_resampled akan berisi dataset yang telah dioversample.

* + 1. Evaluasi Hasil: Setelah melakukan oversampling, sangat penting untuk mengevaluasi kembali dataset Anda untuk memastikan bahwa keseimbangan antara kelas telah dicapai tanpa menimbulkan overfitting.

Anda juga harus melakukan pengujian untuk memastikan bahwa model Anda memberikan hasil yang baik pada dataset yang telah dioversample.

* 1. Kelulusan Mahasiswa Tepat Waktu

Ketepatan waktu kelulusan mahasiswa merupakan salah satu tujuan untuk menyelesaikan pendidikan mahasiswa di perguruan tinggi dengan status tepat waktu atau tamat. Butuh waktu ≤ 4 tahun untuk mencapai tujuan kelulusan tepat waktu dengan gelar sarjana, namun pada kenyataannya masih terdapat kasus mahasiswa yang tidak lulus tepat waktu

Salah satu peran yang berguna dalam mendukung kemajuan suatu negara adalah bidang pendidikan. Pendidikan yang berkualitas akan menghasilkan generasi yang berkualitas. Salah satu jenjang pendidikan yang harus mengutamakan mutu adalah pendidikan tinggi. Setiap perguruan tinggi akan berupaya meningkatkan mutu mutu yang disebut mutu lulusan. Salah satu cara untuk mencapai kualitas lulusan yang baik adalah dengan lebih cepat meluluskan mahasiswa. Sesuai dengan Keputusan Menteri Pendidikan Nomor 232/U/2000 Pasal 5 bahwa beban studi untuk program sarjana direncanakan selama 8 semester (4 tahun) dan dapat ditempuh paling lama 14 semester, yaitu ketika mahasiswa menyelesaikan program sarjana dengan pembulatan lebih 4 tahun sampai dengan ≤ 7 tahun, berarti lulus tetapi tidak tepat waktu karena melebihi waktu yang direncanakan. Setiap mahasiswa tentunya berharap dapat mencapai waktu lulus yang ditentukan dalam arti memperoleh gelar sarjana tepat waktu. Namun, masih banyak kasus mahasiswa yang belum dapat menyelesaikan kuliahnya dengan tepat waktu.(Ningsi & Arofah, 2021)

Lulusan adalah status yang dicapai mahasiswa setelah menyelesaikan proses pendidikan sesuai dengan persyaratan kelulusan yang ditetapkan oleh program studi. Sebagai salah satu keluaran langsung dari proses pendidikan yang dilakukan oleh program studi, lulusan yang bermutu memiliki ciri penguasaan kompetensi akademik termasuk hard skills dan soft skills sebagaimana dinyatakan dalam sasaran mutu serta dibuktikan dengan kinerja lulusan di masyarakat sesuai dengan profesi dan bidang ilmu. Program studi yang bermutu memiliki sistem pengelolaan lulusan yang baik sehingga mampu menjadikannya sebagai human capital bagi progam studi yang bersangkutan.

Perguruan tinggi seharusnya memiliki dan menerapkan kebijakan untuk mengelola dan mengevaluasi standar tersebut yaitu kelulusan mahasiswa yang mencerminkan kinerja perguruan tinggi dalam peningkatan mutu. Salah satu upaya yang dilakukan untuk mengelola kelulusan mahasiswa adalah tercapainya masa studi mahasiswa sesuai dengan lama studi yang telah terjadwal. Namun beberapa perguruan tinggi mengalami kesulitan bagaimana mahasiswa dapat mencapai masa studi tepat waktu sesuai dengan jadwal yang telah ditetapkan perguruan tinggi.

Naïve Bayes merupakan algoritma yang dapat digunakan untuk prediksi masa studi mahasiswa yang bisa dikategorikan tepat waktu, terlambat dan tidak lulus. Dengan menggunakan data mining khususnya klasifikasi untuk prediksi dengan algoritma naïve bayes dapat dilakukan prediksi terhadap ketepatan waktu studi dari mahasiswa berdasarkan data training yang ada.(Rahmatullah et al., 2019)

Dengan demikian dibutuhkan suatu metode Klasifikasi dengan Algoritma Naive Bayes yang dapat memprediksi dengan akurat tingkat kelulusan tepat waktu pada mahasiswa, maka dibuat suatu system pendeteksi kelulusan pada perguruan tinggi menggunakan metode Naïve Bayes. Beberapa penelitian sejenis tentang metode klasifikasi Naïve Bayes diantaranya adalah Perkiraan Waktu Studi Mahasiswa Menggunakan Metode Klasifikasi dengan Algoritma Naive Bayes dengan menggunakan metode Data Mining dengan algoritma Naïve Bayes menyebutkan bahwa sistem yang dirancang adalah sistem prediksi tingkat kelulusan mahasiswa menggunakan metode Data Mining dengan algoritma Naïve Bayes .Kriteria yang digunakan dalam memprediksi tingkat kelulusan adalah NIM, IPK, SKS, kategori kelulusan (tepat, terlambat).Dalam penelitian Prediksi Kelulusan Mahasiswa.(Yustira et al., 2021)

* 1. Lama Studi Mahasiswa

Lama masa studi yang ditempuh oleh mahasiswa merupakan salah satu standar yang termasuk ke dalam standar penilaian pada Standar Nasional Pendidikan Tinggi atau SN-DIKTI, masa studi untuk program sarjana maksimal tujuh tahun akademik dengan beban belajar mahasiswa paling sedikit 144 SKS. Mahasiswa harus berjuang melewati lebih dari 100 SKS dengan ketentuan IPK minimal 2.00

Evaluasi hasil studi pada akhir jenjang studi Strata 1 (S-1) menyebutkan bahwa mahasiswa yang mencapai Indeks Predikat Kumulatif (IPK) minimal 2,00, tidak ada nilai E. Artinya bahwa mahasiswa bisa menempuh perkuliahan hanya dengan 3,5 tahun bila mencapai syarat yang telah ditentukan.(Hendrawan et al., 2021)

# Penelitian Terkait

* 1. Gusti Ngurah Ady Kusuma, Made Pradipta, Made Ari Santosa dan Komang Dharmendra ( 2023 )

Pada penelitian yang dilakukan oleh I Gusti Ngurah Ady Kusuma,I Made Pradipta ,I Made Ari Santosa I Komang Dharmendra dengan judul “Penanganan Ketidakseimbangan Data Pada Klasifikasi Pengaduan Masyarakat” Penelitian ini bertujuan untuk untuk mengatasi ketidakseimbangan data dalam klasifikasi pengaduan masyarakat menggunakan teknik oversampling SMOTE dan ADASYN, serta untuk mengevaluasi pengaruh penggunaan model sampling terhadap kinerja algoritma klasifikasi seperti Naive Bayes Classifier (NBC), Support Vector Machine (SVM), dan Random Forest.

Metode yang digunakan dalam penelitian ini meliputi pengumpulan data pengaduan masyarakat dari halaman publik, vektorisasi teks menggunakan TF-IDF, pembagian dataset menjadi data latih dan data uji, serta penggunaan teknik oversampling SMOTE dan ADASYN untuk menangani ketidakseimbangan data. Selain itu, penelitian ini juga melibatkan pembangunan model klasifikasi menggunakan algoritma Naive Bayes Classifier (NBC), Support Vector Machine (SVM), dan Random Forest, serta evaluasi kinerja model menggunakan matriks evaluasi seperti akurasi, presisi, recall, dan F1-Score. Pengumpulan data dilakukan dengan mengambil data pengaduan masyarakat secara publik dari halaman https://pengaduan.denpasarkota.go.id. Data yang digunakan terdiri dari 10306 data pengaduan yang terbagi menjadi empat kelas, yaitu Keluhan, Usul/Saran, Pertanyaan, dan Informasi.

Hasil penelitian menunjukkan bahwa penggunaan teknik oversampling SMOTE dan ADASYN dapat meningkatkan nilai akurasi pada algoritma SVM dan Random Forest dalam klasifikasi pengaduan masyarakat. Namun, pada algoritma Naive Bayes Classifier (NBC), penggunaan model sampling justru menurunkan akurasi. Waktu proses juga menjadi faktor penting dalam pemilihan algoritma, di mana SVM memiliki waktu proses paling lama, NBC memiliki waktu proses paling pendek, dan Random Forest berada di antara keduanya.

* 1. Cindy Magnolia, Ade Nurhopipah, dan Bagus Adhi Kusuma ( 2022 )

Pada penelitian yang dilakukan oleh Cindy Magnolia, Ade Nurhopipah, dan Bagus Adhi Kusuma dengan judul “Penanganan Imbalanced Dataset untuk Klasifikasi Komentar Program Kampus Merdeka Pada Aplikasi Twitter” Tujuan dari penelitian ini adalah untuk membahas dan menguji berbagai metode penanganan data tidak seimbang (imbalanced data) dalam klasifikasi komentar terkait Program Kampus Merdeka di Twitter. Penelitian ini bertujuan untuk mengevaluasi performa berbagai metode balancing data seperti Undersampling, SMOTE, ADASYN, dan Random Combination Sampling, serta untuk membandingkan hasil klasifikasi menggunakan algoritma SVM dengan pembobotan kata TF-IDF. Penelitian ini penting untuk membantu dalam pemilahan dan pengklasifikasian komentar dari masyarakat terhadap program Kampus Merdeka.

Metode yang digunakan dalam penelitian ini adalah Undersampling, SMOTE, ADASYN, dan Random Combination Sampling untuk menangani data tidak seimbang dalam klasifikasi komentar terkait Program Kampus Merdeka di Twitter. Selain itu, untuk pembobotan kata, penelitian menggunakan metode Term Frequency-Inverse Document Frequency (TF- IDF), dan untuk klasifikasi data, digunakan algoritma Support Vector Machine (SVM)

Adapun Data yang digunakan dalam penelitian ini berasal dari platform Twitter dengan periode Juni hingga Agustus 2022. Pengumpulan data dilakukan dengan menggunakan Twitter API. Data yang diperoleh dalam proses scraping sejumlah 16946 baris.

Hasil dari penelitian menunjukkan bahwa metode ADASYN merupakan yang terbaik dalam menangani data yang tidak seimbang dalam klasifikasi komentar terkait Program Kampus Merdeka di Twitter. Metode ini memberikan performa terbaik dengan nilai akurasi dan F1-score sebesar 0,9. Sementara itu, metode Undersampling tidak memberikan hasil yang baik dalam penelitian tersebut. Penggunaan max\_features pada pembobotan TF-IDF tidak menunjukkan perbedaan signifikan, namun nilai max\_features=5000 cenderung memberikan hasil klasifikasi yang lebih tinggi. Algoritma SVM dipilih karena performanya yang baik dalam mengolah data tidak seimbang. Evaluasi dilakukan untuk melihat F1-score dari model klasifikasi, dan hasilnya menunjukkan bahwa metode ADASYN memberikan performa terbaik. Penelitian ini penting untuk membantu dalam pemilahan dan pengklasifikasian komentar dari masyarakat terhadap program Kampus Merdeka.

* 1. Nana Suryana, Pratiwi dan Rizki Tri Prasetio ( 2021 )

Pada penelitian yang dilakukan oleh Nana Suryana, Pratiwi, Rizki Tri Prasetio dengan judul “Penanganan Ketidakseimbangan Data pada Prediksi Customer Churn Menggunakan Kombinasi SMOTE dan Boosting” Penelitian bertujuan untuk memprediksi customer churn dalam industri telekomunikasi dan mengatasi ketidakseimbangan data dengan menggabungkan teknik SMOTE dan Boosting. Penelitian bertujuan untuk meningkatkan akurasi algoritma klasifikasi dalam memprediksi customer churn dengan menggunakan kombinasi optimalisasi data dan algoritma.

Metode yang digunakan dalam penelitian ini adalah kombinasi teknik sampling SMOTE (Synthetic Minority Over-sampling Technique) dan teknik ensemble Boosting. Teknik resampling digunakan untuk mengatasi ketidakseimbangan data pada customer churn, sedangkan Boosting digunakan untuk meningkatkan ketelitian dalam proses klasifikasi dan prediksi . Penelitian ini juga mengkombinasikan optimasi level data menggunakan SMOTE dan optimasi level algoritma menggunakan AdaBoost untuk mengatasi ketidakseimbangan data pada dataset customer churn.

Data pada penelitian ini dikumpulkan dari industri telekomunikasi yang mencakup informasi tentang pelanggan, layanan yang digunakan, dan apakah pelanggan tersebut melakukan churn atau tidak. Data ini kemudian diolah dan dipersiapkan untuk analisis menggunakan teknik sampling SMOTE dan ensemble Boosting.

Adapun hasil penelitian menunjukkan bahwa dengan mengkombinasikan teknik SMOTE untuk mengatasi ketidakseimbangan data dan teknik Boosting menggunakan algoritma AdaBoost, dapat meningkatkan akurasi algoritma klasifikasi dalam memprediksi customer churn hingga 11%. Algoritma yang memberikan akurasi tertinggi setelah dioptimasi adalah Random Forest dengan akurasi mencapai 89,19%. Penelitian ini juga menunjukkan bahwa penggunaan teknik ensemble seperti Boosting dapat meningkatkan ketelitian dalam proses klasifikasi dan prediksi, serta memberikan hasil yang lebih optimal dalam memprediksi customer churn dalam industri telekomunikasi. Metode yang diusulkan dalam penelitian ini dapat diterapkan dalam berbagai bidang lainnya, seperti prediksi kerusakan tulang belakang dan deteksi cacat perangkat lunak.

* 1. Sabiq Sofyan dan Achmad Prasetyo (2021)

Pada penelitian yang dilakukan oleh Windyaning Ustyannie, Suprapto dengan judul “ Penerapan Synthetic Minority Oversampling Technique (SMOTE) Terhadap Data Tidak Seimbang Pada Tingkat Pendapatan Pekerja Informal Di Provinsi D.I. Yogyakarta Tahun 2019”

Penelitian ini bertujuan mengatasi masalah ketidakseimbangan kelas dalam dataset klasifikasi dengan menggunakan algoritma Synthetic Minority Over-sampling Technique (SMOTE) dan menguji pengaruhnya terhadap kinerja berbagai algoritma klasifikasi seperti Naïve Bayes, SVM, K-NN, dan Decision Tree. Penelitian ini bertujuan untuk meningkatkan akurasi dan g-mean pada data yang tidak seimbang serta mengevaluasi efektivitas penggunaan SMOTE dalam menangani masalah ketidakseimbangan kelas dalam data mining klasifikasi.

Metode yang digunakan dalam penelitian ini adalah Synthetic Minority Over-sampling Technique (SMOTE) untuk menangani ketidakseimbangan kelas dalam dataset klasifikasi. SMOTE digunakan untuk menghasilkan data sintesis pada kelas minoritas agar seimbang dengan kelas mayoritas . Algoritma klasifikasi yang digunakan dalam penelitian ini meliputi C45, Naïve Bayes, K-NN, dan SVM.

Data yang digunakan dalam penelitian ini berasal dari dataset klasifikasi ecoli dengan rasio ketidakseimbangan kelas yang berbeda, yaitu IR 3,3, IR 5,4, IR 8,6, dan IR 15,4. Setiap dataset ecoli memiliki jumlah instance dan distribusi kelas mayoritas dan minoritas yang berbeda . Data ini kemudian dibagi menjadi data training (80%) dan data testing (20%) untuk keperluan eksperimen.

Dan Adapun hasil dari penelitian ini menunjukkan bahwa penerapan algoritma Synthetic Minority Over-sampling Technique (SMOTE) dapat meningkatkan akurasi dan kinerja berbagai algoritma klasifikasi seperti Naïve Bayes, SVM, K-NN, dan Decision Tree pada dataset klasifikasi dengan ketidakseimbangan kelas. Penggunaan SMOTE dapat meningkatkan rata-rata G-Mean dan F-Measure dari dataset yang tidak seimbang, sehingga dapat membantu dalam menangani masalah ketidakseimbangan kelas dalam data mining klasifikasi.

* 1. Windyaning Ustyannie, Suprapto ( 2020 )

Pada penelitian yang dilakukan oleh Windyaning Ustyannie, Suprapto dengan judul “ Oversampling Method To Handling Imbalanced Datasets Problem In Binary Logistic Regression Algorithm” Penelitian ini bertujuan untuk untuk menangani dataset yang tidak seimbang dalam regresi logistik biner menggunakan metode RWO-Sampling dengan pendekatan replikasi acak. Studi ini bertujuan untuk meningkatkan akurasi dengan menghasilkan data sintetis pada atribut diskrit. Hasil penelitian menunjukkan bahwa pendekatan ini lebih unggul dibandingkan metode lain dan dapat menangani ketidakseimbangan kelas secara efektif. Regresi logistik digunakan sebagai algoritma klasifikasi, dan kinerjanya diukur menggunakan akurasi, AUC, f-measure, dan g-mean. Studi ini mendemonstrasikan efektivitas metode yang diusulkan dalam meningkatkan akurasi dan menangani underfitting dibandingkan dengan metode tanpa oversampling.

Metode yang digunakan dalam penelitian ini adalah metode RWO- Sampling dengan pendekatan replikasi acak untuk menangani dataset yang tidak seimbang dalam regresi logistik biner. Pendekatan ini bertujuan untuk meningkatkan akurasi dengan menghasilkan data sintetis pada atribut diskrit. Hasil penelitian menunjukkan efektivitas metode ini dalam menangani ketidakseimbangan kelas dan meningkatkan kinerja regresi logistik.

Data dalam penelitian ini dikumpulkan dari dataset dunia nyata yang digunakan untuk menguji metode RWO-Sampling dalam penanganan dataset yang tidak seimbang dalam regresi logistik biner. Dataset tersebut meliputi NASA MDP dan dataset UCI Repository. Data ini kemudian dikelompokkan berdasarkan kelas yang ada untuk menghasilkan data mayoritas dan minoritas. Fokus penelitian ini adalah pada data minoritas.

Setelah itu, data minoritas didefinisikan dalam atribut diskrit atau atribut kontinu.

Hasil penelitian menunjukkan bahwa metode RWO-Sampling dengan pendekatan replikasi acak efektif dalam menangani dataset yang tidak seimbang dalam regresi logistik biner. Pendekatan ini mampu meningkatkan akurasi dengan menghasilkan data sintetis pada atribut diskrit. Studi ini menunjukkan bahwa metode ini lebih unggul dibandingkan dengan metode lain dalam penanganan ketidakseimbangan kelas dan dapat mengatasi underfitting secara efektif. Regresi logistik digunakan sebagai algoritma klasifikasi, dan kinerjanya diukur menggunakan berbagai metrik seperti akurasi, AUC, f-measure, dan g-mean. Hasil penelitian ini mendemonstrasikan keefektifan metode RWO-Sampling dalam meningkatkan akurasi dan kinerja regresi logistik dalam penanganan dataset yang tidak seimbang.

# Kerangka Pikir

|  |  |  |
| --- | --- | --- |
|  | | Bagaimana keseimbangan (unbalance) pada dataset mempengaruhi akurasi model Naïve Bayes |
| **Masalah** |  |
|  |
|  | |

|  |  |  |
| --- | --- | --- |
|  | | Dengan menggunakan Teknik undersampling dan oversampling pada dataset dengan model naïve bayes |
| **Solusi** |  |
|  |
|  |

|  |  |  |
| --- | --- | --- |
| **Judul** |  | Analisis Unbalance Dengan Teknik Under Sampling Dan Oversampling Untuk Akurasi  Naïve Bayes |
|  |
|  | |

|  |  |  |
| --- | --- | --- |
|  | | Untuk Mengetahui Ketidakseimbangan (Unbalance) dengan menggunakan Teknik Undersampling Dan Oversampling melalui  akurasi pada Naïve Bayes |
| **Hasil** |  |
|  |
|  | |

Gambar 1. Kerangka Pikir

# BAB III

**METODE PENELITIAN**

# Tempat dan Waktu Penelitian

Penelitian ini dilakukan di Fakultas Teknik Universitas Muhammadiyah Makassar, dengan pelaksanaan yang dimulai pada bulan Januari 2024 dan berlangsung hingga seluruh proses pengumpulan data selesai.

# Alat dan Bahan

Adapun alat dan bahan yang akan digunakan dalam penelitian ini adalah sebagai berikut :

* 1. Kebutuhan *Hardware* ( Perangkat Keras )
     1. Laptop Lenovo IdeaPad 1-11ADA05
     2. System operasi Windows 11
  2. Kebutuhan *Software* ( Perangkat Lunak )
     1. Exel
     2. Phyton
     3. Data
     4. Google colabolatory / Google colab

# Perancangan Sistem

Perancangan sistem sangat penting dalam pembangunan suatu system karena menguraikan bagaimana suatu sistem dibangun dari tahap perencanaan hingga tahap pembuatan fungsi-fungsi yang diperlukan untuk pengoperasian sistem. Tujuan dari perancangan sistem adalah untuk menentukan apakah sistem yang akan dikembangkan akan mengahsilkan hasil yang di inginkan.

* 1. Studi Literatur

Langkah pertama adalah melakukan studi literatur untuk memahami masalah ketidakseimbangan kelas dalam dataset, teknik-teknik yang digunakan untuk menangani masalah tersebut, serta konsep dasar dari metode klasifikasi Naïve Bayes.

* 1. Pengumpulan Data

Tahap ini melibatkan pengumpulan dataset yang akan digunakan dalam penelitian. Dataset ini harus mencakup kelas yang tidak seimbang untuk memungkinkan analisis dan evaluasi yang tepat terhadap teknik under sampling dan oversampling.

* 1. Pengolahan Data

Data yang dikumpulkan perlu diproses untuk mempersiapkannya untuk analisis lebih lanjut. Ini termasuk langkah-langkah seperti membersihkan data, menghapus nilai-nilai yang hilang atau tidak valid, dan melakukan normalisasi jika diperlukan.

* 1. Perancangan Sistem

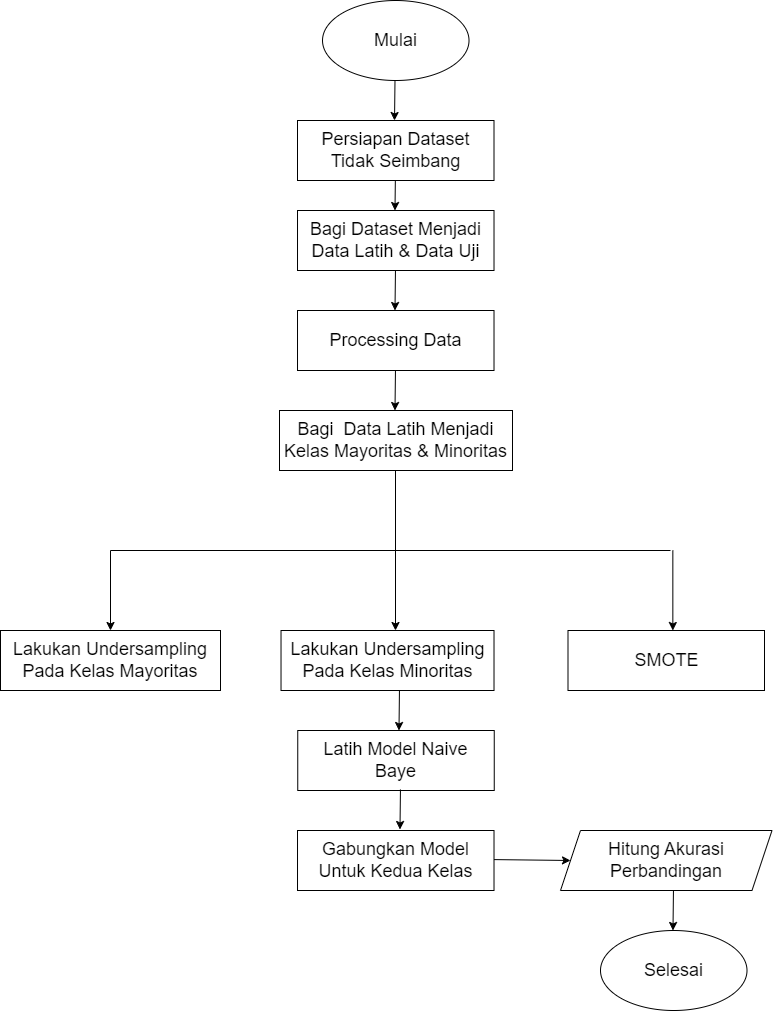
Perancangan sistem dalam konteks penelitian ini akan melibatkan implementasi teknik under sampling dan oversampling untuk menangani ketidakseimbangan kelas dalam dataset. Selain itu, perancangan akan melibatkan implementasi metode klasifikasi Naïve Bayes sebagai model prediksi.

* 1. Pengujian Sistem

Setelah sistem dirancang dan diimplementasikan, tahap pengujian diperlukan untuk mengevaluasi kinerja sistem. Ini melibatkan penggunaan dataset yang telah diproses untuk melihat seberapa baik teknik under sampling dan oversampling meningkatkan kinerja Naïve Bayes dalam mengklasifikasikan data yang tidak seimbang.

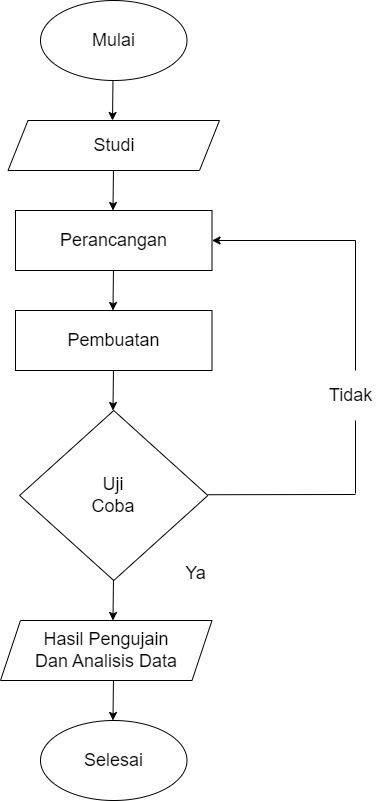
* 1. Penarikan Kesimpulan

Setelah pengujian selesai, kesimpulan dapat ditarik berdasarkan hasil evaluasi. Ini mencakup menganalisis efektivitas teknik under sampling dan oversampling dalam meningkatkan akurasi Naïve Bayes pada dataset yang tidak seimbang dan mengidentifikasi temuan penting serta implikasi dari penelitian tersebut.



Gambar 2. Perancangan Sistem

Flowchart di atas menggambarkan bahwa proses dimulai dengan mempersiapkan dataset tidak seimbang, kemudian membaginya menjadi data latih dan data uji. Data latih diproses dan dibagi lagi menjadi kelas mayoritas dan minoritas. Pada kelas mayoritas, dilakukan undersampling untuk mengurangi jumlah instance, sedangkan pada kelas minoritas dilakukan oversampling menggunakan teknik *SMOTE* untuk membuat contoh sintetis baru. Setelah data seimbang, model *Naïve Bayes* dilatih menggunakan data latih yang telah diimbangi. Hasilnya akan dibandingkan untuk melihat peningkatan kinerja model setelah penyeimbangan data. Dalam perancangan sistem atau diagram system yang akan dibuat yaitu sebagai berikut :



Gambar 3. Diagram Alur Penelitian

Pada diagram di atas dapat dijelaskan bahwa Langkah awal dimulai dengan studi untuk memahami masalah yang akan diselesaikan atau sistem yang akan dikembangkan, diikuti dengan perancangan sistem untuk merencanakan solusi yang sesuai dengan kebutuhan yang telah diidentifikasi. Tahap berikutnya adalah pembuatan sistem berdasarkan rencana yang telah dirancang sebelumnya, yang kemudian diikuti oleh uji coba sistem untuk mengvaluasi kinerjanya. Jika hasil uji coba memenuhi kriteria yang ditetapkan, maka proses akan dilanjutkan, tetapi jika tidak,sistem akan Kembali ke tahap perancangan untuk penyesuaian. Setelah itu, hasil pengujian dan analisis data dilakukan untuk mengevaluasi kinerja sistem dan mengidentifikasi masalah atau kekurangan. Akhirnya, setelah iterasi dan revisi yang sesuai, sistem dianggap selesai dan siap untuk digunakan atau diimplementasikan.

# Teknik Pengujian Sistem

Teknik pengujian sistem yang akan digunakan pada pengujian ini adalah dimana dataset akan dibagi menjadi set latihan dan set pengujian. Kemudian, Teknik under sampling dan oversampling akan diterapkan pada set latihan untuk menangani masalah ketidakseimbangan kelas. Setelah itu, model Naïve Bayes akan dilatih menggunakan dataset yang telah diubah. Pengujian dilakukan dengan menggunakan set pengujian yang tidak diubah untuk mengevaluasi kinerja model. Hasil perbandingan kemudian dianalisis untuk menilai apakah Teknik Undersampling dan oversampling berhasil meningkatkan akurasi Naïve Bayes dalam mengatasi ketidakseimbangan kelas dalam dataset. Kesimpulan dari analisis ini akan memberikan wawasan tentang afektivitas Teknik-teknik tersebut dalam meningkatan kinerja model klasifikasi.

Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%

𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖

# Teknik Analisis Data

Analisis data merupakan upaya yang dilakukan untuk mengklasifikasi dan mengelompokkan data. Pada tahap ini dilakukan upaya mengelompokkan, menyamakan data yang sama dan membedakan data yang memang berbeda, serta menyisihkan pada kelompok lain data yang serupa, tetapi tidak sama.(Sutriani & Octaviani, 2019) Proses analisis data dilakukan pada penelitian ini adalah sebagai berikut :

* 1. Reduksi Data *( Data Reduction )*

Proses reduksi data adalah langkah krusial dalam menghadapi volume besar data lapangan. Ini menuntut peneliti untuk mencatat setiap detail dengan teliti, dengan tujuan memberikan gambaran yang lebih jelas dan memudahkan proses pengumpulan dan penelusuran data lebih lanjut. Bagi peneliti kualitatif, fokusnya terletak pada hasil, sehingga reduksi data menjadi kunci untuk mencapai pemahaman yang mendalam. Namun, dalam proses ini, peneliti harus berhati-hati terhadap outliers, data yang tidak diketahui, dan kualifikasi yang tidak terstruktur.

* 1. Penyajian Data *( Display Data )*

Hasil penelitian akan disajikan secara eksklusif untuk setiap pola, kategori, fokus, dan tema yang ingin dipahami oleh peneliti. Penggunaan display data membantu peneliti melihat gambaran keseluruhan atau bagian- bagian tertentu berdasarkan output penelitian. Dalam penelitian kualitatif, penyajian data bisa berupa uraian singkat, diagram, interaksi antar kategori, dan bentuk visual lainnya. Tekstual deskripsi seringkali digunakan untuk menyajikan data dalam konteks penelitian kualitatif.

* 1. Penarikan Kesimpulan *( Conclusing Drawing Verivication )*

Langkah ketiga dalam analisis data melibatkan penarikan kesimpulan dan verifikasi. Kesimpulan awal yang diajukan bersifat sementara dan dapat berubah sesuai dengan kebutuhan bukti tambahan melalui pengumpulan data lanjutan. Oleh karena itu, kesimpulan dalam penelitian mungkin dapat menjawab pertanyaan penelitian yang sudah dirumuskan awalnya. Ini disebabkan oleh sifat sementara dari perumusan masalah dalam penelitian kualitatif, yang dapat berkembang seiring dengan proses penelitian lapangan.

# BAB IV

**HASIL DAN PEMBAHASAN**

# Pengambilan Data

Tahap pengumpulan data yang digunakan dalam penelitian ini diperoleh dari Universitas SIMAK Muhammadiyah Makassar. SIMAK merupakan sumber data resmi yang mencatat informasi akademik mahasiswa, antara lain riwayat akademik, kualifikasi mata kuliah, dan status akademik lainnya. Data yang dikumpulkan adalah mahasiswa kurikulum teknik pengairan Fakultas Teknik yang masuk pada tahun 2017 hingga 2023. Melibatkan beberapa kelompok mahasiswa. Data yang dikumpulkan meliputi informasi lengkap tentang siswa seperti nama, nomor induk siswa, waktu pendaftaran, nilai dan ciri-ciri penting lainnya. Berikut gambar dataset yang dihasilkan :

Tabel 1. Data Mahasiswa Angkatan 2018

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| NIM | NAMA | …. | NAMA\_MATAKULIAH | SKS | KODE\_NILAI |
| 105811100118 | ISWANDI | **….** | PENDIDIKAN AGAMA ISLAM | 2 | A |
| 105811100118 | ISWANDI | **….** | ILMU SOSIAL BUDAYA DASAR (ISBD) | 2 | B |
| 105811100118 | ISWANDI | **….** | MATEMATIKA TEKNIK 1 | 2 | C |
| 105811100118 | ISWANDI | **….** | HIDROLOGI TEKNIK DASAR | 2 | E |
| 105811100318 | ADELLINA SAHNAZ SUSANTO PUTRI | **….** | PENDIDIKAN AGAMA ISLAM | 2 | B |
| 105811100318 | ADELLINA SAHNAZ SUSANTO PUTRI | **….** | AIK II (PENG. STUDI ISLAM DAN KEMUHAMMADIYAH 2) | 2 | A |
| 105811100318 | ADELLINA SAHNAZ SUSANTO PUTRI | **….** | SEMINAR DAN USULAN SKRIPSI | 2 | E |
| …. | **….** | **….** | **….** | **….** | **….** |
| 105811123518 | RIFKI RAIHAN M | **….** | PENDIDIKAN KEWARGANEGARAAN | 2 | K |
| 105811123518 | RIFKI RAIHAN M | **….** | KOMPREHENSIF AIK | 2 | E |
| 105811123518 | RIFKI RAIHAN M | **….** | PENDIDIKAN AGAMA ISLAM | 2 | B |
| 105811123518 | RIFKI RAIHAN M | **….** | BAHASA INDONESIA | 2 | D |
| 105811123518 | RIFKI RAIHAN M | **….** | PENDIDIKAN PANCASILA | 2 | C |
| 105811123518 | RIFKI RAIHAN M | **….** | AIK II (PENG. STUDI ISLAM DAN KEMUHAMMADIYAH 2) | 2 | B- |
| 105811123518 | RIFKI RAIHAN M | **….** | AIK IV(IBADAH KHASHASHAH DAN KEMUHAMMADIYAHAN 4) | 2 | A- |
| 105811123518 | RIFKI RAIHAN M | **….** | AIK V (AKHLAQ AL KARIEMAH DAN KEMUHAMMADIYAHAN 5) | 2 | A |
| 105811123518 | RIFKI RAIHAN M | **….** | ETIKA PROFESI | 2 | B+ |
| 105811123618 | SITISAITONG BINMA | **….** | PENDIDIKAN AGAMA ISLAM | 2 | A |
| 105811123618 | SITISAITONG BINMA | **….** | PENDIDIKAN PANCASILA | 2 | B |
| 105811123618 | SITISAITONG BINMA | **….** | AIK VII (ILMU FALAQ 2 DAN KEMUHAMMADIYAH AN 7) | 2 | B- |
| 105811123618 | SITISAITONG BINMA | **….** | SEMINAR DAN USULAN SKRIPSI | 2 | E |

Tabel 2. Data Mahasiswa Angkatan 2019

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| NIM | NAMA | …. | NAMA\_MATAKULIAH | SKS | KODE\_NILAI |
| 105811100219 | M.TAUFIK HIDAYAT | **….** | AL ISLAM KEMUHAMMADIYAHAN VI | 1 | B+ |
| 105811100219 | M.TAUFIK HIDAYAT | **….** | AIK III (AQIDAH ISLAM DAN KEMUHAMMADIYAHAN 3) | 1 | A- |
| 105811100219 | M.TAUFIK HIDAYAT | **….** | AIK VII (ILMU FALAQ 2 DAN KEMUHAMMADIYAH 7) | 1 | B |
| 105811100219 | M.TAUFIK HIDAYAT | **….** | PENDIDIKAN AGAMA ISLAM | 2 | A |
| 105811100219 | M.TAUFIK HIDAYAT | **….** | ILMU SOSIAL BUDAYA DASAR (ISBD) | 2 | E |
| 105811100219 | M.TAUFIK HIDAYAT | **….** | PENDIDIKAN PANCASILA | 2 | C |
| 105811100219 | M.TAUFIK HIDAYAT | **….** | KEPEMIMPINAN DAN KEWIRAUSAHAAN | 2 | B- |
| 105811100219 | M.TAUFIK HIDAYAT | **….** | SKRIPSI | 6 | K |
| 105811100219 | M.TAUFIK HIDAYAT | **….** | TEKNIK JALAN RAYA | 2 | B+ |
| 105811100219 | M.TAUFIK HIDAYAT | **….** | PERENCANAAN BANGUNAN PANTAI | 2 | C+ |
| 105811100219 | M.TAUFIK HIDAYAT | **….** | STRUKTUR STATIS TERTENTU | 2 | D |
| 105811104419 | MUH. FIKRA ABDULLAH | **….** | PENDIDIKAN AGAMA ISLAM | 2 | E |
| 105811100119 | NURUL FADLI | **….** | KOMPREHENSIF AIK | 2 | A |
| 105811100119 | NURUL FADLI | **….** | ILMU SOSIAL BUDAYA DASAR (ISBD) | 2 | B |
| 105811100119 | NURUL FADLI | **….** | KEPEMIMPINAN DAN KEWIRAUSAHAAN | 2 | A- |
| 105811100119 | NURUL FADLI | **….** | SEMINAR DAN USULAN SKRIPSI | 2 | K |
| 105811100119 | NURUL FADLI | **….** | ETIKA PROFESI | 2 | E |
| 105811100119 | NURUL FADLI | **….** | METODE NUMERIK | 2 | C |
| 105811100119 | NURUL FADLI | **….** | TEKNIK KONSERVASI WADUK | 2 | B+ |
| 105811100119 | NURUL FADLI | **….** | IRIGASI I | 2 | A- |
| 105811100119 | NURUL FADLI | **….** | TEKNIK SUNGAI I | 2 | B- |
| …. | **….** | **….** | **….** | **….** | **….** |
| 105811122619 | M.SANDI PUTRA | **….** | PENDIDIKAN KEWARGANEGARAAN | 2 | A |
| 105811122619 | M.SANDI PUTRA | **….** | BAHASA INGGRIS TEKNIK II | 2 | C |
| 105811122619 | M.SANDI PUTRA | **….** | AIK VII (ILMU FALAQ 2 DAN KEMUHAMMADIYAH 7) | 1 | B+ |
| 105811122619 | M.SANDI PUTRA | **….** | BAHASA INGGRIS TEKNIK 1 | 2 | E |
| 105811122619 | M.SANDI PUTRA | **….** | ILMU SOSIAL BUDAYA DASAR (ISBD) | 2 | B |
| 105811122619 | M.SANDI PUTRA | **….** | AIK VI (ILMU FALAQ 1 DAN KEMUHAMMADIAYAHAN 6) | 2 | A- |
| 105811122619 | M.SANDI PUTRA | **….** | BAHASA INGGRIS TEKNIK 2 | 2 | C |
| 105811122619 | M.SANDI PUTRA | **….** | KULIAH KERJA PROFESI - PLUS (KKP-PLUS) | 4 | K |
| 105811122619 | M.SANDI PUTRA | **….** | FISIKA TEKNIK | 2 | D |

Tabel 3. Data Mahasiswa Angkatan 2020

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| NIM | NAMA | …. | NAMA\_MATAKULIAH | SKS | KODE\_NILAI |
| 105811100920 | FATIMA AZZAHRA | **….** | AL ISLAM KEMUHAMMADIYAHAN VI | 1 | A- |
| 105811100920 | FATIMA AZZAHRA | **….** | KOMPREHENSIF AIK | 2 | B+ |
| 105811100920 | FATIMA AZZAHRA | **….** | AIK V (AKHLAQ AL KARIEMAH DAN KEMUHAMMADIYAHAN 5 ) | 1 | A |
| 105811100920 | FATIMA AZZAHRA | **….** | BAHASA INGGRIS TEKNIK 2 | 2 | B |
| 105811100920 | FATIMA AZZAHRA | **….** | SEMINAR DAN USULAN SKRIPSI | 2 | E |
| 105811100920 | FATIMA AZZAHRA | **….** | SKRIPSI | 6 | K |
| 105811109620 | FATIHATUL HIDAYAH | **….** | FISIKA TEKNIK | 2 | C |
| 105811109620 | FATIHATUL HIDAYAH | **….** | STATISTIK DAN DASAR-DASAR PROBABILITAS | 2 | B |
| 105811109620 | FATIHATUL HIDAYAH | **….** | PEMOGRAMAN DASAR KOMPUTER | 2 | A |
| …. | **….** | **….** | **….** |  |  |
| 105811110520 | Mahmud | **….** | KOMPREHENSIF AIK | 2 | A |
| 105811110520 | Mahmud | **….** | ILMU SOSIAL BUDAYA DASAR (ISBD) | 2 | B |
| 105811110520 | Mahmud | **….** | KALKULUS II | 2 | K |
| 105811110520 | Mahmud | **….** | SEMINAR DAN USULAN SKRIPSI | 2 | E |
| 105811110520 | Mahmud | **….** | ILMU TANAH DAN TANAMAN | 2 | A- |
| 105811110520 | Mahmud | **….** | TEKNIK PONDASI | 2 | B+ |
| 105811110520 | Mahmud | **….** | TEKNIK SUNGAI I | 2 | C |
| 105811110520 | Mahmud | **….** | TEKNIK SUNGAI I | 2 | A- |
| 105811110520 | Mahmud | **….** | BANGUNAN AIR I | 2 | B+ |
| 105811110520 | Mahmud | **….** | PERENCANAAN PEMBANGKIT LISTRIK TENAGA AIR ( PLTA) | 2 | A |
| 105811110520 | Mahmud | **….** | PENGELOLAAN AIR TANAH (GROUND WATER) | 2 | C+ |
| 105811110520 | Mahmud | **….** | EKONOMI TEKNIK | 2 | B+ |
| 105811110420 | MUH. AL FAREL FADHUL RAHMAN.R | **….** |  |  |  |

Tabel 4. Data Mahasiswa Angkatan 2021

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| NIM | NAMA | …. | NAMA\_MATAKULIAH | SKS | KODE\_NILAI |
| 105811110121 | MUHAMMAD RAFLI | **….** | PENDIDIKAN AGAMA ISLAM | 2 | A- |
| 105811110121 | MUHAMMAD RAFLI | **….** | BAHASA INGGRIS TEKNIK 1 | 2 | A |
| 105811110121 | MUHAMMAD RAFLI | **….** | AIK II (PENG. STUDI ISLAM DAN KEMUHAMMADIYAH 2) | 2 | E |
| 105811110121 | MUHAMMAD RAFLI | **….** | MATEMATIKA DASAR | 3 | C+ |
| 105811110121 | MUHAMMAD RAFLI | **….** | STATISTIK DAN DASAR-DASAR PROBABILITAS | 2 | B |
| 105811110121 | MUHAMMAD RAFLI | **….** | MATEMATIKA TEKNIK 1 | 2 | C |
| 105811110121 | MUHAMMAD RAFLI | **….** | PROBABILITAS LANJUTAN | 2 | B- |
| 105811100221 | Sariana | **….** | AL ISLAM KEMUHAMMADIYAHAN IV | 1 | A |
| 105811100221 | Sariana | **….** | AL ISLAM KEMUHAMMADIYAHAN VI | 1 | K |
| 105811100221 | Sariana | **….** | PENDIDIKAN AGAMA ISLAM | 2 | A- |
| 105811100221 | Sariana | **….** | MATEMATIKA DASAR | 3 | B- |
| 105811100221 | Sariana | **….** | FISIKA TEKNIK | 2 | B |
| 105811100221 | Sariana | **….** | PEMOGRAMAN DASAR KOMPUTER | 2 | B+ |
| 105811100221 | Sariana | **….** | MENGGAMBAR KONSTRUKSI BANGUNAN AIR | 2 | A |
| …. | **….** | **….** | **….** | **….** |  |
| 105811111921 | ANDI ZULKIFLI AGUS | **….** | AL ISLAM KEMUHAMMADIYAHAN IV | 1 | K |
| 105811111921 | ANDI ZULKIFLI AGUS | **….** | PENDIDIKAN AGAMA ISLAM | 2 | A- |
| 105811111921 | ANDI ZULKIFLI AGUS | **….** | BAHASA INDONESIA | 2 | A |
| 105811111921 | ANDI ZULKIFLI AGUS | **….** | AIK II (PENG. STUDI ISLAM DAN KEMUHAMMADIYAH 2) | 2 | B+ |
| 105811111921 | ANDI ZULKIFLI AGUS | **….** | KALKULUS II | 2 | B |
| 105811111921 | ANDI ZULKIFLI AGUS | **….** | APLIKASI KOMPUTER | 2 | C+ |
| 105811111921 | ANDI ZULKIFLI AGUS | **….** | TEKNIK PONDASI II | 2 | C |

Tabel 5. Data Mahasiswa Angkatan 2022

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| NIM | NAMA | …. | NAMA\_MATAKULIAH | SKS | KODE\_NILAI |
| 105811107922 | Nabil Amri Pahlaguna | **….** | PENDIDIKAN AGAMA ISLAM | 2 | E |
| 105811100322 | Novi Julianti | **….** | AL ISLAM KEMUHAMMADIYAHAN III | 1 | A |
| 105811100322 | Novi Julianti | **….** | AL ISLAM KEMUHAMMADIYAHAN IV | 1 | K |
| 105811100322 | Novi Julianti | **….** | ILMU SOSIAL BUDAYA DASAR (ISBD) | 2 | B |
| 105811100322 | Novi Julianti | **….** | KALKULUS I | 2 | B+ |
| 105811100322 | Novi Julianti | **….** | KALKULUS II | 2 | A- |
| 105811100322 | Novi Julianti | **….** | ILMU UKUR TANAH DAN PEMETAAN | 3 | A |
| …. | **….** | **….** | **….** | **….** | **….** |
| 105811109822 | Muhammad Rezky Saputra | **….** | AL ISLAM KEMUHAMMADIYAHAN II | 1 | B |
| 105811109822 | Muhammad Rezky Saputra | **….** | BAHASA INGGRIS TEKNIK II | 2 | A |
| 105811109822 | Muhammad Rezky Saputra | **….** | AIK III (AQIDAH ISLAM DAN KEMUHAMMADIYAHAN 3) | 1 | A- |
| 105811109822 | Muhammad Rezky Saputra | **….** | ILMU SOSIAL BUDAYA DASAR (ISBD) | 2 | B+ |
| 105811109822 | Muhammad Rezky Saputra | **….** | KALKULUS I | 2 | C |
| 105811109822 | Muhammad Rezky Saputra | **….** | MENGGAMBAR TEKNIK | 2 | E |
| 105811109822 | Muhammad Rezky Saputra | **….** | HIDROLIKA I | 2 | C+ |
| 105811109822 | Muhammad Rezky Saputra | **….** | ILMU UKUR TANAH DAN PEMETAAN | 3 | A |

Tabel 6. Data Mahasiswa Angkatan 2023

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| NIM | NAMA | …. | NAMA\_MATAKULIAH | SKS | KODE\_NILAI |
| 105811100223 | MUH. ARAS ASRULLAH | **…** | AL ISLAM KEMUHAMMADIYAHAN II | 1 | K |
| 105811100223 | MUH. ARAS ASRULLAH | **…** | PANCASILA | 2 | A |
| 105811100223 | MUH. ARAS ASRULLAH | **…** | BAHASA INGGRIS TEKNIK I | 2 | A- |
| 105811100223 | MUH. ARAS ASRULLAH | **…** | KALKULUS I | 2 | B+ |
| 105811100223 | MUH. ARAS ASRULLAH | **…** | ILMU UKUR TANAH DAN PEMETAAN | 3 | K |
|  |  | **…** |  |  |  |
| 105811107023 | MUHAMMAD YUSRIL ZUL HAQ S | **…** | AL ISLAM KEMUHAMMADIYAHAN II | 1 | K |
| 105811107023 | MUHAMMAD YUSRIL ZUL HAQ S | **…** | PANCASILA | 2 | A |
| 105811107023 | MUHAMMAD YUSRIL ZUL HAQ S | **…** | BAHASA INGGRIS TEKNIK I | 2 | A- |
| 105811107023 | MUHAMMAD YUSRIL ZUL HAQ S | **…** | PENDIDIKAN AGAMA ISLAM | 2 | B+ |
| 105811107023 | MUHAMMAD YUSRIL ZUL HAQ S | **…** | KALKULUS I | 2 | E |
| 105811107023 | MUHAMMAD YUSRIL ZUL HAQ S | **…** | STATISTIK DAN DASAR-DASAR PROBABILITAS | 2 | B- |
| 105811107023 | MUHAMMAD YUSRIL ZUL HAQ S | **…** | ILMU UKUR TANAH DAN PEMETAAN | 3 | K |
| 105811107123 | MUH.HIKMAL HIDAYAT | **…** | AL ISLAM KEMUHAMMADIYAHAN II | 1 | K |
| 105811107123 | MUH.HIKMAL HIDAYAT | **…** | PANCASILA | 2 | A |
| 105811107123 | MUH.HIKMAL HIDAYAT | **…** | BAHASA INGGRIS TEKNIK I | 2 | B+ |
| 105811107123 | MUH.HIKMAL HIDAYAT | **…** | KALKULUS I | 2 | E |
| 105811107123 | MUH.HIKMAL HIDAYAT | **…** | STATISTIK DAN DASAR-DASAR PROBABILITAS | 2 | A- |
| 105811107123 | MUH.HIKMAL HIDAYAT | **…** | MENGGAMBAR TEKNIK | 2 | E |
| 105811107123 | MUH.HIKMAL HIDAYAT | **…** | FISIKA I | 2 | B |
| 105811107123 | MUH.HIKMAL HIDAYAT | **…** | KIMIA BAHAN BANGUNAN | 2 | K |
| 105811107123 | MUH.HIKMAL HIDAYAT | **…** | ILMU UKUR TANAH DAN PEMETAAN | 3 | K |
| 105811100223 | MUH. ARAS ASRULLAH | **…** | AL ISLAM KEMUHAMMADIYAHAN II | 1 | K |
| 105811100223 | MUH. ARAS ASRULLAH | **…** | PANCASILA | 2 | A |
| 105811100223 | MUH. ARAS ASRULLAH | **…** | BAHASA INGGRIS TEKNIK I | 2 | A- |
| 105811100223 | MUH. ARAS ASRULLAH | **…** | KALKULUS I | 2 | B+ |
| 105811100223 | MUH. ARAS ASRULLAH | **…** | STATISTIK DAN DASAR-DASAR PROBABILITAS | 2 | B |
| 105811100223 | MUH. ARAS ASRULLAH | **…** | FISIKA II | 2 | K |

Tabel 1 hingga 6 di atas menampilkan data mahasiswa yang mencakup berbagai aspek seperti NIM, nama mahasiswa, tahun, periode KRS, kode mata kuliah, jumlah Satuan Kredit Semester (SKS), jumlah mata kuliah wajib dengan nilai baik, jumlah mata kuliah dengan nilai buruk, dan durasi waktu studi. Data ini merupakan data mentah yang diperoleh dari sistem informasi akademik fakultas, yang akan digunakan untuk memprediksi kelulusan mahasiswa.

# Normalisasi Data

Normalisasi data adalah proses pengaturan data dalam basis data untuk mencapai dua tujuan utama: mengurangi redundansi data (menghindari penyimpanan data yang sama lebih dari sekali) dan menjamin integritas data.

Data yang dapat digunakan dipilih dengan cara memilih atribut dan membersihkan data yang tidak terpakai agar dapat diproses atau dianalisis dengan lebih efisien. Proses ini mengubah nilai-nilai variabel agar dapat dibandingkan dalam skala yang sama. Tujuan normalisasi adalah untuk menghindari ketidaksamaan bobot atribut dalam model analisis data.

Tabel 7. Min-Max

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Nim | jumlah\_semester | IPS 1 | SKS 1 | ... | IPS 8 | SKS 8 | total\_SKS | total\_IPK | Unnamed: 22 | Unnamed: 23 |
| count | 9,78E+08 | 978000000 | 978000000 | 978000000 | ... | 978000000 | 978000000 | 978000000 | 978000000 | 0.0 | 1.0 |
| mean | 1,06E+17 | 6.900.818 | 2.600.460 | 18.117.587 | ... | 1.196.421 | 4.871.166 | 94.788.344 | 2.961.943 | NaN | 0.0 |
| std | 7,78E+09 | 3.611.562 | 1.209.952 | 7.144.902 | ... | 1.366.376 | 6.303.300 | 56.406.200 | 0.859140 | NaN | NaN |
| min | 1,06E+17 | 0.000000 | 0.000000 | 0.000000 | ... | 0.000000 | 0.000000 | 0.000000 | 0.000000 | NaN | 0.0 |
| 25% | 1,06E+17 | 4.000.000 | 1.970.000 | 17.000.000 | ... | 0.000000 | 0.000000 | 40.000.000 | 2.812.500 | NaN | 0.0 |
| 50% | 1,06E+17 | 8.000.000 | 3.050.000 | 21.000.000 | ... | 0.420000 | 2.000.000 | 110500000 | 3.210.000 | NaN | 0.0 |
| 75% | 1,06E+17 | 9.000.000 | 3.550.000 | 23.000.000 | ... | 2.345.000 | 10.000.000 | 150000000 | 3.480.000 | NaN | 0.0 |
| max | 1,06E+17 | 13.000.000 | 4.000.000 | 24.000.000 | ... | 4.000.000 | 61.000.000 | 159000000 | 4.000.000 | NaN | 0.0 |

Proses min-maks dilakukan untuk mencakup normalisasi nilai yang mendukung analisis data. Data historis ini memberikan landasan yang kuat untuk melatih model Naïve Bayes, karena mencakup informasi tentang mahasiswa dari berbagai Angkatan.

# Pelabelan Data Mahasiswa

Proses pelabelan data mahasiswa adalah langkah penting dalam penelitian ini, terutama untuk mengidentifikasi dan memahami karakteristik yang membedakan tingkat kelulusan tepat waktu mahasiswa. Pelabelan data ini melibatkan prediksi tingkat kelulusan, di mana status kelulusan dikategorikan menjadi prediksi lulus tepat waktu dan prediksi tidak lulus tepat waktu dalam semester tertentu.

Dalam pelabelan data mahasiswa, variabel atau atribut yang akan digunakan sebagai input adalah NIM, nama, jumlah semester, total SKS, nilai IPS dari semester 1 hingga semester 7, dan yang terakhir adalah nilai IPK.

Kriteria yang digunakan dalam pelabelan ini adalah sebagai berikut: PREDIKSI LULUS TEPAT WAKTU; jika mahasiswa menyelesaikan studinya dalam 8 semester dengan memperoleh total nilai SKS 150 atau sesuai dengan ketentuan yang berlaku di universitas tersebut, serta memenuhi semua persyaratan akademik. PREDIKSI TIDAK LULUS TEPAT WAKTU; jika mahasiswa tidak menyelesaikan studinya dalam jangka waktu yang telah ditentukan dan tidak memenuhi persyaratan akademik.

Berikut adalah gambaran data mahasiswa yang telah dilabeli. Pada ujung tabel, terdapat kolom label untuk mahasiswa yang prediksinya tidak lulus tepat waktu karena nilai IPK yang diperoleh tidak memenuhi standar persyaratan di kampus tersebut.

Tabel 8. Dataset

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No | Nim | Nama | jumlah\_semester | IPS 1 | SKS 1 | … | IPS 8 | SKS 8 | total\_SKS | total\_IPK | label |
| 0 | 105811100117 | NOER MUHAMMAD INDRA MOESLIM RAHMAN | 6 | 2.32 | 18 | **…** | 0.00 | 0 | 46 | 2.54 | TIDAK LULUS TEPAT WAKTU |
| 1 | 105811100217 | SULIMIN | 7 | 0.59 | 9 | **…** | 0.00 | 0 | 23 | 1.96 | TIDAK LULUS TEPAT WAKTU |
| 2 | 105811100317 | UMMU KALSUM | 9 | 3.82 | 22 | **…** | 1.83 | 2 | 156 | 3.61 | LULUS TEPAT WAKTU |
| 3 | 105811100417 | FIRMAN | 13 | 2.68 | 18 | **….** | 2.36 | 20 | 143 | 3.15 | TIDAK LULUS TEPAT WAKTU |
| 4 | 105811100517 | MUH. IRWANSYAH | 13 | 3.32 | 20 | **….** | 3.08 | 24 | 153 | 3.48 | TIDAK LULUS TEPAT WAKTU |
| ….. | ….. | ….. | ….. | ….. | ….. | …. | ….. | ….. | ….. | ….. | ….. |
| 1077 | 105811110022 | Muh. Zulfadli Amrullah | 2.0 | 1.43 | 12.0 | **….** | 0.00 | 0.0 | 14.0 | 0.87 | PREDIKSI TIDAK LULUS TEPAT WAKTU |
| 1078 | 105811110122 | Galih Reyhan Ramadafi | 3.0 | 0.16 | 2.0 | **….** | 0.00 | 0.0 | 9.0 | 0.37 | PREDIKSI TIDAK LULUS TEPAT WAKTU |
| 1079 | 105811110222 | Zul fadli | 3.0 | 3.29 | 22.0 | **….** | 0.00 | 0.0 | 62.0 | 3.13 | PREDIKSI LULUS TEPAT WAKTU |
| 1080 | 105811110322 | MUH. WAHIDIN RUSLAN | 4.0 | 1.87 | 12.0 | **…** | 0.00 | 0.0 | 17.0 | 1.03 | PREDIKSI TIDAK LULUS TEPAT WAKTU |
| 1081 | 105811110422 | ABDUL HADI ARIF | 4.0 | 1.95 | 18.0 | **…** | 0.00 | 0.0 | 35.0 | 1.46 | PREDIKSI TIDAK LULUS TEPAT WAKTU |

# Penerapan Naïve Bayes

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

import sklearn

from sklearn.impute import SimpleImputer

warnings.simplefilter("ignore")

Kode tersebut berfungsi untuk mengimpor berbagai pustaka yang akan digunakan dalam analisis data, visualisasi, serta pemodelan machine learning, dan untuk mengatur beberapa konfigurasi awal.

Mengimpor pustaka pandas digunakan untuk manipulasi dan analisis data, terutama dalam bentuk tabel *(DataFrame),* sedangkan pustaka *numpy* digunakan untuk operasi numerik efisien, terutama pada array multidimensi. Untuk visualisasi data, pustaka *matplotlib.pyplot* dan seaborn digunakan, di mana *matplotlib.pyplot* membuat grafik dan plot, sementara seaborn membuat visualisasi data yang lebih menarik dan informatif, dibangun di atas matplotlib. Pustaka warnings digunakan untuk mengontrol peringatan yang muncul selama eksekusi kode *Python*, sering kali digunakan untuk menyembunyikan peringatan yang tidak diinginkan. Pustaka sklearn *(scikit-learn)* digunakan untuk tugas-tugas *machine learning* seperti preprocessing data, training model, dan evaluasi, dengan SimpleImputer sebagai kelas dalam *scikit-learn* yang digunakan untuk menangani data yang hilang dengan cara menggantinya dengan nilai statistik seperti mean, median, atau modus. Menyembunyikan peringatan mengatur agar semua peringatan diabaikan, yang berguna untuk menghindari peringatan yang tidak relevan atau mengganggu saat menjalankan kode.

#Import manajemen dataset

df=pd.read\_excel("Datanew.xlsx", sheet\_name = 'Sheet1')

df.describe()

Kode ini digunakan untuk mengimpor dan menampilkan ringkasan statistik dari dataset yang disimpan dalam file Excel. Berikut adalah penjelasan rinci dari setiap baris kode:

* 1. Mengimpor dataset dari file Excel:
     1. *pd.read\_excel* adalah fungsi dari pustaka pandas yang digunakan untuk membaca file Excel.
     2. *"Dataundersampling.xlsx"* adalah nama file Excel yang akan dibaca.
     3. sheet\_name='Sheet1' menentukan bahwa data akan diambil dari sheet pertama yang bernama 'Sheet1'.
     4. Data yang diimpor akan disimpan dalam DataFrame df.
  2. Menampilkan ringkasan statistik:

*df.describe()* adalah metode pandas yang memberikan ringkasan statistik dari DataFrame df, seperti mean, standard deviation, min, max, dan quartile values untuk setiap kolom numerik dalam DataFrame.

* 1. Menampilkan DataFrame:

df pada baris terakhir bertujuan untuk menampilkan keseluruhan DataFrame df.

start\_col = 2

end\_col = 20

x= df.iloc[:, start\_col:end\_col+1]

y=df.iloc[:,21]

y

x

Kode ini digunakan untuk memilih subset dari kolom dalam DataFrame df dan membagi data menjadi fitur (x) dan target (y). Berikut adalah penjelasan rinci dari setiap baris kode:

1. Menentukan indeks kolom awal dan akhir:
   1. start\_col = 2 menetapkan bahwa kolom awal untuk subset adalah kolom dengan indeks 2 (kolom ketiga dalam DataFrame, karena indeks dimulai dari 0).
   2. end\_col = 20 menetapkan bahwa kolom akhir untuk subset adalah kolom dengan indeks 20.
2. Memilih subset kolom sebagai fitur (x):
   1. df.iloc[:, start\_col:end\_col+1] menggunakan .iloc untuk memilih kolom dari start\_col hingga end\_col, ditambah satu untuk memastikan kolom dengan indeks end\_col disertakan.
   2. x adalah DataFrame yang berisi kolom-kolom yang dipilih tersebut.
3. Memilih kolom target (y):
   1. df.iloc[:, 21] menggunakan .iloc untuk memilih kolom dengan indeks 21 (kolom ke-22 dalam DataFrame) sebagai target (y).
   2. y adalah Series yang berisi nilai-nilai dari kolom target.
4. Menampilkan target (y):

Menampilkan y, yang berisi nilai-nilai dari kolom target. Jika dijalankan di lingkungan interaktif seperti Jupyter Notebook, ini akan menampilkan nilai- nilai tersebut.

1. Menampilkan fitur (x):

Menampilkan x, yang berisi subset kolom yang dipilih sebagai fitur. Jika dijalankan di lingkungan interaktif seperti Jupyter Notebook, ini akan menampilkan DataFrame tersebut.

from sklearn.model\_selection import train\_test\_split

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x,y,random\_state=0,test\_size=0.2)

Kode ini digunakan untuk membagi dataset menjadi set pelatihan (training set) dan set pengujian (test set). Berikut adalah penjelasan rinci dari setiap baris kode:

1. Mengimpor train\_test\_split dari scikit-learn:

train\_test\_split adalah fungsi dari pustaka scikit-learn yang digunakan untuk membagi dataset menjadi subset pelatihan dan pengujian.

1. Membagi dataset menjadi set pelatihan dan pengujian:
   1. train\_test\_split(x, y, random\_state=0, test\_size=0.1) membagi dataset menjadi empat subset:
      1. x\_train: fitur untuk set pelatihan.
      2. x\_test: fitur untuk set pengujian.
      3. y\_train: target untuk set pelatihan.
      4. y\_test: target untuk set pengujian.
   2. Parameter yang digunakan:
      1. x dan y: DataFrame x dan Series y yang telah dipilih sebelumnya sebagai fitur dan target.
      2. random\_state=0: Menetapkan nilai seed untuk memastikan pembagian dataset yang konsisten setiap kali kode dijalankan. Ini berguna untuk reproduksibilitas hasil.
      3. test\_size=0.1: Menentukan bahwa 10% dari data akan digunakan sebagai set pengujian, sementara 90% sisanya akan digunakan sebagai set pelatihan.

from imblearn.over\_sampling import SMOTE

imputer = SimpleImputer(strategy='mean') #mengubah NAN menjadi nilai rata2

x = imputer.fit\_transform(x)

smote = SMOTE(random\_state=42)

x\_resampled, y\_resampled = smote.fit\_resample(x, y)

Kode tersebut bertujuan untuk menangani nilai yang hilang dalam dataset

dan mengatasi masalah ketidakseimbangan kelas menggunakan teknik oversampling SMOTE *(Synthetic Minority Over-sampling Technique).* Berikut adalah penjelasan rinci dari setiap bagian kode:

1. Mengimpor SMOTE dari imbalanced-learn:

SMOTE adalah teknik oversampling yang digunakan untuk menangani ketidakseimbangan kelas dalam dataset. Ia bekerja dengan membuat sampel sintetis dari kelas minoritas untuk menyeimbangkan jumlah sampel dalam setiap kelas.

1. Menginisialisasi SimpleImputer untuk mengisi nilai yang hilang:
   1. SimpleImputer adalah kelas dari scikit-learn yang digunakan untuk menangani nilai yang hilang dalam dataset.
   2. strategy='mean' menetapkan strategi untuk mengganti nilai yang hilang dengan rata-rata nilai dari kolom tersebut.
2. Mengganti nilai yang hilang dalam fitur (x) dengan nilai rata-rata:
   1. imputer.fit\_transform(x):
      1. fit menghitung rata-rata dari setiap kolom fitur dalam x.
      2. transform menggantikan nilai yang hilang dengan rata-rata yang telah dihitung.
   2. Hasilnya adalah x yang telah diisi dengan nilai rata-rata untuk setiap nilai yang hilang.
3. Menginisialisasi SMOTE:
   1. SMOTE(random\_state=42) menginisialisasi objek SMOTE dengan random\_state=42 untuk memastikan bahwa hasil oversampling dapat direproduksi (konsisten setiap kali kode dijalankan).
4. Menggunakan SMOTE untuk oversampling fitur (x) dan target (y):
   1. smote.fit\_resample(x, y):
   2. fit\_resample menerapkan teknik SMOTE pada fitur x dan target y untuk membuat sampel sintetis dari kelas minoritas.
   3. Hasilnya adalah dua variabel baru:
      1. x\_resampled: Fitur yang telah diresampling.
      2. y\_resampled: Target yang telah diresampling.

from sklearn.naive\_bayes import GaussianNB

from sklearn.impute import SimpleImputer

# Membuat imputer untuk menangani nilai yang hilang

imputer = SimpleImputer(strategy='mean')

# Membuat model Naive Bayes

model = GaussianNB()

# Mengimputasi nilai yang hilang dan melatih model dengan data

x\_train = imputer.fit\_transform(x\_resampled)

model.fit(x\_train, y\_resampled)

Kode ini bertujuan untuk menangani nilai yang hilang dalam dataset,

kemudian melatih model Naive Bayes menggunakan data yang telah diimputasi dan diresampling. Berikut adalah penjelasan rinci dari setiap bagian kode:

1. Mengimpor Gaussian Naive Bayes dari scikit-learn:

GaussianNB adalah implementasi dari algoritma Naive Bayes yang mengasumsikan bahwa fitur mengikuti distribusi Gaussian (normal). Algoritma ini sering digunakan untuk klasifikasi.

1. Mengimpor SimpleImputer dari scikit-learn:

SimpleImputer adalah kelas yang digunakan untuk menangani nilai yang hilang dalam dataset dengan menggantinya menggunakan strategi tertentu, seperti mean, median, atau modus.

1. Membuat imputer untuk menangani nilai yang hilang :

imputer adalah objek SimpleImputer yang diinisialisasi dengan strategi 'mean'. Ini berarti bahwa nilai yang hilang akan diganti dengan rata-rata dari kolom tersebut.

1. Membuat model Naive Bayes:

model adalah objek GaussianNB yang merupakan model klasifikasi Naive Bayes dengan asumsi distribusi Gaussian untuk fitur-fitur.

1. Mengimputasi nilai yang hilang dan melatih model dengan data :
   1. imputer.fit\_transform(x\_resampled):
      1. fit menghitung rata-rata dari setiap kolom fitur dalam x\_resampled.
      2. transform menggantikan nilai yang hilang dengan rata-rata yang telah dihitung.
   2. Hasilnya adalah x\_train, yaitu x\_resampled yang telah diisi dengan nilai rata-rata untuk setiap nilai yang hilang.
2. Melatih model Naive Bayes dengan data yang telah diimputasi dan diresampling:
   1. model.fit(x\_train, y\_resampled):
      1. fit melatih model Naive Bayes menggunakan fitur x\_train dan target y\_resampled.
      2. Model akan mempelajari hubungan antara fitur dan target berdasarkan data yang disediakan.

# Mengimputasi nilai yang hilang pada data pengujian

x\_test\_imputed = imputer.transform(x\_test)

# Melakukan prediksi dengan data pengujian yang telah diimputasi

y\_predict = model.predict(x\_test\_imputed)

Kode ini bertujuan untuk menggunakan model yang telah dilatih untuk membuat prediksi pada data uji dan kemudian menghitung akurasi model tersebut. Berikut adalah penjelasan rinci dari setiap bagian kode :

1. Membuat prediksi pada data uji:
   1. model.predict(x\_test) menggunakan model Naive Bayes yang telah dilatih untuk membuat prediksi berdasarkan fitur-fitur dalam x\_test.
   2. Hasilnya adalah y\_predict, yang berisi prediksi kelas untuk setiap sampel dalam x\_test.
2. Menghitung akurasi model:
   1. sklearn.metrics.accuracy\_score(y\_test, y\_predict) menghitung akurasi model dengan membandingkan prediksi (y\_predict) dengan nilai aktual (y\_test).
   2. accuracy\_score adalah fungsi dari sklearn.metrics yang mengembalikan proporsi prediksi yang benar (benar dibagi dengan total jumlah prediksi).

Langkah-langkah rinci dari fungsi ini:

1. y\_test: Nilai-nilai aktual dari target untuk data uji.
2. y\_predict: Nilai-nilai yang diprediksi oleh model untuk data uji.
3. Akurasi dihitung sebagai jumlah prediksi yang benar dibagi dengan total jumlah sampel.

# Menghitung akurasi

akurasi = sklearn.metrics.accuracy\_score(y\_test, y\_predict)

# Menampilkan hasil akurasi

print(f"Akurasi: {akurasi}")

Kode ini digunakan untuk menampilkan data uji (x\_test) dan nilai akurasi (akurasi) dari model yang telah dilatih dan diuji. Berikut adalah penjelasan rinci dari setiap bagian kode:

1. Menampilkan data uji (x\_test):
   1. print(x\_test) akan mencetak isi dari x\_test ke layar.
   2. x\_test berisi fitur-fitur dari data uji yang digunakan untuk membuat prediksi menggunakan model yang telah dilatih.
   3. Dengan menampilkan x\_test, kita dapat melihat data yang digunakan untuk menguji model.
2. Menampilkan akurasi model (akurasi):
   1. print(akurasi) akan mencetak nilai akurasi ke layar.
   2. akurasi adalah nilai yang dihitung sebelumnya menggunakan
   3. Dengan menampilkan akurasi, kita dapat melihat seberapa baik model melakukan klasifikasi pada data uji.

# Pastikan bahwa x\_test adalah numpy array sebelum mengakses kolomnya

if not isinstance(x\_test, np.ndarray):

    x\_test = np.array(x\_test)

# Pastikan bahwa y\_test adalah numpy array sebelum membuat DataFrame

if not isinstance(y\_test, np.ndarray):

    y\_test = np.array(y\_test)

# Buat DataFrame dengan hasil prediksi

results = pd.DataFrame({

    'Feature 1': x\_test[:, 0],

    'Feature 2': x\_test[:, 1],

    'Predicted': y\_predict,

    'Actual': y\_test

})

# Tampilkan DataFrame

print(results)

# Simpan DataFrame ke dalam file Excel

results.to\_excel('hasil\_prediksi.xlsx', index=False)

Kode ini mengambil prediksi model dan nilai aktual dari data uji, kemudian membuat DataFrame dari hasil tersebut dan menyimpannya ke dalam file Excel. Berikut adalah penjelasan langkah demi langkah dari kode tersebut:

1. Memastikan x\_test adalah numpy array:
   1. Kode ini memeriksa apakah x\_test sudah dalam format numpy array. Jika belum, maka diubah menjadi numpy array menggunakan np.array(x\_test).
   2. Ini diperlukan karena beberapa operasi di bawah ini (seperti mengakses kolom menggunakan slicing) memerlukan x\_test dalam format numpy array.
2. Memastikan y\_test adalah numpy array:
   1. Kode ini memeriksa apakah y\_test sudah dalam format numpy array. Jika belum, maka diubah menjadi numpy array menggunakan np.array(y\_test).
   2. Ini diperlukan karena kita akan membuat DataFrame yang membutuhkan y\_test dalam format numpy array.
3. Membuat DataFrame results:
   1. Kode ini menggunakan pd.DataFrame() dari pandas untuk membuat DataFrame results.
   2. DataFrame ini memiliki empat kolom:
      1. 'Feature 1' dan 'Feature 2': Fitur-fitur dari x\_test. Di sini diasumsikan bahwa x\_test memiliki minimal dua fitur, sehingga kita bisa mengakses kolom pertama dan kedua menggunakan x\_test[:, 0] dan x\_test[:, 1].
      2. 'Predicted': Prediksi yang dihasilkan oleh model (y\_predict).
      3. 'Actual': Nilai aktual dari y\_test.
4. Menampilkan DataFrame results:

Kode ini mencetak DataFrame results ke layar. Jika dijalankan dalam lingkungan seperti Jupyter Notebook, hasilnya akan ditampilkan dalam bentuk tabel.

1. Menyimpan DataFrame ke dalam file Excel:
   1. Kode ini menyimpan DataFrame results ke dalam file Excel dengan nama 'hasil\_prediksi.xlsx'.
   2. index=False mengatur agar indeks baris tidak disertakan dalam file Excel yang disimpan.

# Hasil Pengujian Data Tanpa Menggunakan Smote

Pengujian data tanpa menggunakan SMOTE dilakukan dengan teknik undersampling dan oversampling secara manual, tanpa menggunakan library. Pada teknik undersampling, data dikurangi, sementara pada teknik oversampling, data ditambahkan. Ukuran data yang digunakan untuk undersampling adalah 400 data, dengan masing-masing 200 data untuk label prediksi lulus tepat waktu dan prediksi tidak lulus tepat waktu. Untuk oversampling, digunakan 600 data, dengan masing-masing 300 data untuk label prediksi lulus tepat waktu dan prediksi tidak lulus tepat waktu. Data dibagi menjadi data latih dan data uji dengan perbandingan 90:10, 80:20, dan 70:30.

Data latih digunakan untuk melatih model, sedangkan data uji digunakan untuk menguji kinerja model yang telah dilatih.

* 1. Pengujian Data Dengan Teknik Undersampling

1. Data 200

Tabel 9. Undersampling Data 200

|  |  |
| --- | --- |
| Pembagian Data | Hasil Akurasi |
| 90 : 10 | 90% |
| 80 : 20 | 92% |
| 70 : 30 | 90% |

Pada tabel di atas, dapat dilihat bahwa dengan menggunakan data sebanyak 200 untuk teknik undersampling dengan pembagian data terdiri dari 90 : 10, 80 : 20 dan 70 : 30, menghasilkan akurasi dengan nilai 90%, 92% dan 90%. Dari ketiga pembagian data tersebut, dengan akurasi yang tertinggi diperoleh dari pembagian data 80 : 20 sebesar 92%.

1. Data 300

Tabel 10. Undersampling Data 300

|  |  |
| --- | --- |
| Pembagian Data | Hasil Akurasi |
| 90 : 10 | 96,6% |
| 80 : 20 | 93,3% |
| 70 : 30 | 93,3% |

Tabel tersebut menjelaskan bahwa dengan menggunakan data sebanyak 300 dengan pembagian data terdiri dari 90 : 10, 80 : 20 dan 70 : 30, menghasilkan akurasi tertinggi dengan nilai akurasi sebesar 96,6% pada pembagian data 90 : 10.

1. Data 400

Tabel 11. Undersampling Data 400

|  |  |
| --- | --- |
| Pembagian Data | Hasil Akurasi |
| 90 : 10 | 82,5% |
| 80 : 20 | 86,2% |
| 70 : 30 | 82,5% |

Pada tabel di atas, dapat dilihat bahwa dengan menggunakan data sebanyak 400 dengan pembagian data terdiri dari 90 : 10, 80 : 20 dan 70 :30, menghasilkan akurasi dengan nilai 82,5%, 86,2% dan 82,5%. Dari ketiga pembagian data tersebut, dengan akurasi yang tertinggi diperoleh dari pembagian data 80 : 20 sebesar 86,2%.

Secara keseluruhan, hasil pengujian menunjukkan bahwa pembagian data yang berbeda dapat menghasilkan tingkat akurasi yang bervariasi. Untuk data sebanyak 200, pembagian 80:20 menghasilkan akurasi tertinggi. Untuk data sebanyak 300, pembagian 90:10 menghasilkan akurasi tertinggi. Untuk data sebanyak 400, pembagian 80:20 menghasilkan akurasi tertinggi. Dengan demikian, pemilihan pembagian data yang optimal sangat tergantung pada jumlah data yang digunakan.

* 1. Pengujian Data Dengan Teknik Oversampling
     1. Data 500

Tabel 12. Oversampling Data 500

|  |  |
| --- | --- |
| Pembagian Data | Hasil Akurasi |
| 90 : 10 | 92% |
| 80 : 20 | 86% |
| 70 : 30 | 84,6% |

Pada tabel di atas, dapat dilihat bahwa dengan menggunakan data sebanyak 500 untuk teknik oversampling dengan pembagian data terdiri dari 90 : 10, 80 : 20 dan 70 : 30, menghasilkan akurasi dengan nilai 92%, 86% dan 84,6%. Dari ketiga pembagian data tersebut, dengan akurasi yang tertinggi diperoleh dari pembagian data 90 : 10 sebesar 92%.

2.) Data 600

Tabel 13. Oversampling Data 600

|  |  |
| --- | --- |
| Pembagian Data | Hasil Akurasi |
| 90 : 10 | 91,6% |
| 80 : 20 | 88,3% |
| 70 : 30 | 85% |

Pada tabel di atas, dapat dilihat bahwa dengan menggunakan data sebanyak 600 dengan pembagian data terdiri dari 90 : 10, 80 : 20 dan 70 :30, menghasilkan akurasi dengan nilai 91,6%, 88,3% dan 85%. Dari ketiga pembagian data tersebut, dengan akurasi yang tertinggi diperoleh dari pembagian data 90 : 10 sebesar 91,6%.

3.) Data 700

Tabel 14. Oversampling Data 700

|  |  |
| --- | --- |
| Pembagian Data | Hasil Akurasi |
| 90 : 10 | 91,4% |
| 80 : 20 | 91,4% |
| 70 : 30 | 89,5% |

Pada tabel di atas, dapat dilihat bahwa dengan menggunakan data sebanyak 700 dengan pembagian data terdiri dari 90 : 10, 80 : 20 dan 70 :30, menghasilkan akurasi dengan nilai 91,4%, 91,4% dan 89,5%. Dari ketiga pembagian data tersebut, dengan akurasi yang tertinggi diperoleh dari pembagian data 90 : 10 dan 80 : 20 dengan nilai sebesar 91,4%.

Secara umum, hasil pengujian menunjukkan bahwa teknik oversampling dengan pembagian data 90:10 cenderung menghasilkan akurasi tertinggi pada berbagai ukuran data yang diuji. Pembagian 80:20 juga memberikan hasil yang baik, terutama pada data sebanyak 700, di mana akurasinya setara dengan pembagian 90:10. Ini mengindikasikan bahwa dalam teknik oversampling, pembagian data yang lebih mendekati 90:10 umumnya memberikan hasil akurasi yang lebih optimal.

# Hasil Pengujian Data Dengan Menggunakan Smote

Pada pengujian data menggunakan SMOTE berbeda dengan pengujian tanpa SMOTE. Pengujian dengan SMOTE tidak melibatkan teknik undersampling dan oversampling. Data yang digunakan untuk pengujian ini berjumlah 1.155, dengan 778 data untuk label Prediksi Lulus Tepat Waktu dan 377 data untuk label Prediksi Tidak Lulus Tepat Waktu, sehingga data ini tidak seimbang. Data dibagi menjadi data latih dan data uji dengan perbandingan 90:10, 80:20, dan 70:30. Data latih digunakan untuk melatih model, sedangkan data uji digunakan untuk menguji kinerja model yang telah dilatih.

Tabel 15. Hasil Pengujian Data Dengan Menggunakan Smote

|  |  |
| --- | --- |
| Pembagian Data | Hasil Akurasi |
| 90 : 10 | 89,1% |
| 80 : 20 | 91,8% |
| 70 : 30 | 91,3% |

Pada tabel di atas, dapat dilihat bahwa dengan menggunakan data sebanyak 1.155 dengan 778 data untuk label prediksi lulus tepat waktu dan377 data untuk label prediksi tidak lulus tepat waktu untuk teknik oversampling.

Dimana, dilakukan pembagian data terdiri dari 90 : 10, 80 : 20 dan 70 : 30 yang menghasilkan akurasi dengan nilai 89,1%, 91,8% dan 91,3%. Sehingga, menghasilkan nilai akurasi yang paling tinggi yaitu pada pembagian data 80 : 20 sebesar 91,8%.

# Perhitungan Manual

Pada pengujian ini, perhitungan manual dilakukan untuk mencari nilai akurasi dengan cara menghitung jumlah prediksi yang benar, kemudian dibagi dengan total data uji dan dikalikan dengan 100%. Untuk mendapatkan jumlah prediksi yang benar dan total data uji, data diambil dari hasil prediksi mesin, dan kemudian diproses di Exel. Berikut hasil perhitungan manual.

1. Pengujian Manual Tanpa Menggunakan SMOTE
2. Undersampling Data 200

Tabel 16. Perhitungan Manual Undersampling 200

|  |  |
| --- | --- |
| **Data** | **Akurasi** |
| 90 : 10 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 18 x 100% = 90 %  21 |
| 80 : 20 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 37 x 100% = 92,5%  40 |
| 70 : 30 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 54 x 100% = 90%  64 |

1. Undersampling Data 300

Tabel 17. Perhitungan Manual Undersampling 300

|  |  |
| --- | --- |
| **Data** | **Akurasi** |
| 90 : 10 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 29 x 100% = 96,6 %  30 |
| 80 : 20 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 56 x 100% = 93,3%  60 |
| 70 : 30 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 84 x 100% = 93,3%  90 |

1. Undersampling Data 400

Tabel 18. Perhitungan Manual Undersampling 400

|  |  |
| --- | --- |
| **Data** | **Akurasi** |
| 90 : 10 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 33 x 100% = 82,5 %  40 |
| 80 : 20 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 69 x 100% = 86,2%  80 |
| 70 : 30 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 33 x 100% = 82,5%  40 |

Setelah melakukan pengujian, diporoleh hasil akurasi dari metode Naïve Bayes dengan menggunakan ukuran data sebanyak 200, 300 dan 400, dengan masing – masing pembagian data latih dan data uji sebesar 90:10, 80:20, dan 70:30. Diantara ukuran data yang digunakan diperoleh hasil akurasi yang tinggi sebesar 96,6 %, Dimana hasil ini menggunakan data sebanyak 300 dengan pembagian data 90 : 10.

1. Pengujian Manual Tanpa Menggunakan SMOTE
2. Oversampling Data 500

Tabel 19. Perhitungan Manual Oversampling 500

|  |  |
| --- | --- |
| **Data** | **Akurasi** |
| 90 : 10 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 46 x 100% = 92 %  50 |
| 80 : 20 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 86 x 100% = 86%  100 |
| 70 : 30 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 127 x 100% = 84,6%  150 |

1. Oversampling Data 600

Tabel 20. Perhitungan Manual Oversampling 600

|  |  |
| --- | --- |
| **Data** | **Akurasi** |
| 90 : 10 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 55 x 100% = 91,6 %  61 |
| 80 : 20 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 106 x 100% = 88,3%  120 |
| 70 : 30 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 153 x 100% = 85%  180 |

1. Oversampling Data700

Tabel 21. Perhitungan Manual Oversampling 700

|  |  |
| --- | --- |
| **Data** | **Akurasi** |
| 90 : 10 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 64 x 100% = 91,4 %  70 |
| 80 : 20 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 128 x 100% = 91,4%  140 |
| 70 : 30 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 188 x 100% = 89,5%  210 |

Setelah melakukan pengujian, diporoleh hasil akurasi dari metode Naïve Bayes dengan menggunakan ukuran data sebanyak 500, 600 dan 700, dengan masing – masing pembagian data latih dan data uji sebesar 90:10, 80:20, dan 70:30. Diantara ukuran data yang digunakan diperoleh hasil akurasi yang tinggi sebesar 92%, Dimana hasil ini menggunakan data sebanyak 500 dengan pembagian data 90 : 10.

1. Pengujian Manual Menggunakan SMOTE

Tabel 22. Perhitungan Manual Menggunakan SMOTE

|  |  |
| --- | --- |
| **Data** | **Akurasi** |
| 90 : 10 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 66 x 100% = 89,1 %  74 |
| 80 : 20 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 135 x 100% = 91,8%  147 |
| 70 : 30 | Akurasi = 𝐽𝑢𝑚𝑙𝑎ℎ 𝑃𝑟𝑒𝑑𝑖𝑘𝑠𝑖 𝑦𝑎𝑛𝑔 𝑏𝑒𝑛𝑎𝑟 x 100%  𝑇𝑜𝑡𝑎𝑙 𝐷𝑎𝑡𝑎 𝑈𝑗𝑖  = 201 x 100% = 91,3%  220 |

Setelah melakukan pengujian, dapat dilihat pada tabel di atas. Tabel tersebut menunjukkan bahwa perhitungan manual menggunakan SMOTE menghasilkan nilai akurasi sebesar 91,8% dengan pembagian data 80 : 20.

# Perbandingan Pengujian Menggunakan SMOTE dan Tanpa Menggunakan SMOTE

Dari hasil pengujian kita dapat mengetahui perbandingan pengujian data menggunakan SMOTE dan Tanpa Menggunakan SMOTE. Hasil perbandingan dapat dilihat pada table berikut :

Tabel 23. Hasil Perbandingan

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Hasil Pengujian** | | **Data Latih dan Uji** | | **Akurasi** | **Pembagian Data** | |
|  |  | 90 : 10  80 : 20  70 : 30 | | 90 %  92 %  90 % | 200 | |
| Tanpa Menggunakan SMOTE | Teknik  Undersampling |  | |
|  | 90 : 10  80 : 20  70 : 30 | | 96,6 %  93,3%  93,3% | 300 | |
|  | | | 90 : 10 | 82,5 % | | | 400 |
| 80 : 20 | 86,2% | | |  |
| 70 : 30 | 82,5 % | | |  |
|  | | | 90 : 10 | 92 % | | | 500 |
| Teknik | | | 80 : 20 | 86% | | |  |
| Oversampling | | | 70 : 30 | 84,6 % | | |
|  | | | 90 : 10 | 91,6 % | | | 600 |
| 80 : 20 | 88,3% | | |  |
| 70 : 30 | 85 % | | |  |
|  | | | 90 : 10 | 91,4 % | | | 700 |
| 80 : 20 | 91,4% | | |  |
| 70 : 30 | 89,5 % | | |  |
|  | | | 90 : 10 | 89,1 % | | | 732 |
| Menggunakan SMOTE | | | 80 : 20 | 91,8 % | | |  |
|  | | | 70 : 30 | 91,3% | | |

Teknik undersampling pada data tanpa SMOTE umumnya memberikan hasil yang lebih baik dibandingkan dengan teknik oversampling. Ini terlihat dari akurasi tertinggi yang dicapai oleh teknik undersampling pada data sebanyak 300 dengan pembagian 90:10, yaitu 96,6%. Sebaliknya, teknik oversampling pada data yang sama dengan pembagian data yang berbeda tidak mencapai akurasi setinggi itu. Penggunaan SMOTE menghasilkan akurasi yang baik, khususnya dengan pembagian data 80:20, yang memberikan akurasi 91,8% pada data sebanyak 732.

Namun, hasil ini masih lebih rendah dibandingkan akurasi tertinggi yang dicapai tanpa SMOTE dengan teknik undersampling.

Selain itu, pembagian data 90:10 seringkali menghasilkan akurasi tertinggi dalam pengujian tanpa SMOTE, baik pada teknik undersampling maupun oversampling. Misalnya, teknik undersampling pada data sebanyak 300 dan 400 dengan pembagian 90:10 masing-masing menghasilkan akurasi 96,6% dan 82,5%. Di sisi lain, teknik oversampling pada data sebanyak 500, 600, dan 700 dengan pembagian 90:10 juga menghasilkan akurasi tinggi, yaitu 92%, 91,6%, dan 91,4%. Namun, ketika menggunakan SMOTE, pembagian data 80:20 lebih sering memberikan hasil terbaik, seperti terlihat pada akurasi 91,8% yang dicapai pada data sebanyak 732.

Oleh karena itu, pemilihan teknik dan pembagian data yang optimal sangat bergantung pada metode yang digunakan dan ukuran data yang tersedia. Teknik undersampling tanpa SMOTE cenderung lebih efektif untuk mencapai akurasi tertinggi, terutama dengan pembagian data 90:10. Namun, saat menggunakan SMOTE, pembagian data 80:20 lebih sering memberikan hasil yang lebih baik. Ini menunjukkan bahwa tidak ada pendekatan tunggal yang paling efektif untuk semua situasi, dan evaluasi yang cermat terhadap data dan metode yang digunakan sangat penting untuk mencapai hasil terbaik dalam pengujian model.

# BAB V KESIMPULAN DAN SARAN

1. **Kesimpulan**

Ketidakseimbangan (unbalance) pada dataset sangat mempengaruhi akurasi model Naïve Bayes. Dengan pengujian dataset sebanyak 1.155 data menunjukkan bahwa teknik undersampling menghasilkan akurasi tertinggi sebesar 96,6%, sementara teknik oversampling mencapai akurasi maksimal 92%. Penggunaan SMOTE memberikan hasil yang baik dengan akurasi tertinggi 91,8%. Teknik undersampling cenderung mengurangi akurasi dengan menghilangkan informasi penting dari kelas mayoritas, sedangkan oversampling berisiko overfitting. Oleh karena itu, pemilihan metode penyeimbangan harus dilakukan dengan hati-hati, menyesuaikan sifat data dan tujuan analisis, serta dievaluasi dengan cermat untuk mencapai performa model yang optimal.

# Saran

Dari hasil penelitian yang dilakukan dapat diperoleh saran yaitu, Untuk melakukan evaluasi berkala terhadap performa model setelah penerapan teknik penyeimbangan dan prapemrosesan agar model tetap konsisten dan relevan dengan perubahan data dan dinamika bisnis. Hal ini diharapkan dapat memberikan kontribusi lebih dalam dalam mengatasi ketidakseimbangan data pada model klasifikasi Naïve Bayes.

# DAFTAR PUSTAKA

Choirunnisa, S. (2019). *Metode Hibrida Oversampling dan Undersampling Untuk Menangani Ketidakseimbangan Data Kegagalan Aakademik Universitas XYZ*.

Ericha Apriliyani, & Salim, Y. (2022). Analisis performa metode klasifikasi Naïve Bayes Classifier pada Unbalanced Dataset. *Indonesian Journal of Data and Science*, *3*(2), 47–54. https://doi.org/10.56705/ijodas.v3i2.45

Hari, E., Prastyo, A., Studi, P., & Wiratsongko, R. (2020). *Implementasi Teknik Web Scraping Pada Situs Berita Menggunakan Metode Supervised learning Implementasi Web Scraping Pada Situs Berita Menggunakan Metode Supervised learning IGL Putra Eka Prismana*. https://republika.co.id

Hendrawan, I. N. R., Arya, I. M., Saputra, B., Ayu, G., Cahya, P., & Gede, I. (2021). *Klasifikasi Lama Studi dan Predikat Kelulusan Mahasiswa menggunakan Metode Naïve Bayes*. 50–56. https://doi.org/10.30864/eksplora.v11i1.606

Heranova, O. (2019). Synthetic Minority Oversampling Technique pada Averaged One Dependence Estimators untuk Klasifikasi Credit Scoring. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *3*(3), 443–450. https://doi.org/10.29207/resti.v3i3.1275

Kawani, G. P. (2019). Implementasi Naive Bayes. *Journal of Informatics, Information System, Software Engineering and Applications (INISTA)*, *1*(2), 73–81. https://doi.org/10.20895/inista.v1i2.73

Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti, S., Chamidah, N., Mega Santoni, M., Matondang, N., Ilmu Komputer, F., & Pembangunan Nasional Veteran Jakarta, U. (2017). Terakreditasi SINTA Peringkat 2 Pengaruh Oversampling pada Klasifikasi Hipertensi dengan Algoritma Naïve Bayes, Decision Tree, dan Artificial Neural Network (ANN). *Masa Berlaku Mulai*, *1*(3), 635–641.

Mutmainah, S. (2021). Penanganan Imbalance Data Pada Klasifikasi Kemungkinan Penyakit Stroke. *SNATi*, *1*(1), 10–16. https://doi.org/10.20885/snati.v1i1.2

Ningsi, B. A., & Arofah, I. (2021).

*…………………………………………………………………………………………*

*………………………………………………………………………..*

[*http://ejurnal.binawakya.or.id/index.php/MBI*](http://ejurnal.binawakya.or.id/index.php/MBI) *Vol.15 No.10 Mei 2021 Open Journal Systems*. *15*(10), 5097–5104.

Prihandari, R. C. (2022). *Data Mining: Konsep Dan Apikasi Menggunakan Rapidminer (Series: Supervised Learning Dan Unsupervised Learning)*. 8. <http://repository.uin-suska.ac.id/63073/1/REGITA> CAHYANI PRIHANDARI.pdf

Rahmatullah, S., Utami, E., Informatika, T., Teknik, M., Dian, S., Cendikia, C., Yogyakarta, U. A., Negara, J., Candimas, N., & Lampung, K. (2019). *Jurnal Informasi Dan Komputer Vol : 7 No : 1 2019 PREDIKSI TINGKAT KELULUSAN TEPAT WAKTU DENGAN METODE NAÏVE BAYES DAN K-*

*NEAREST NEIGHBOR Jurnal Informasi Dan Komputer Vol : 7 No : 1 2019*. 7–16.

Septiani, Y., Aribbe, E., & Diansyah, R. (2020). ANALISIS KUALITAS LAYANAN SISTEM INFORMASI AKADEMIK UNIVERSITAS ABDURRAB TERHADAP KEPUASAN PENGGUNA MENGGUNAKAN

METODE SEVQUAL (Studi Kasus : Mahasiswa Universitas Abdurrab Pekanbaru). *Jurnal Teknologi Dan Open Source*, *3*(1), 131–143. https://doi.org/10.36378/jtos.v3i1.560

Sulistiyono, M., Pristyanto, Y., Adi, S., & Gumelar, G. (2021). Implementasi Algoritma Synthetic Minority Over-Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi. *Sistemasi*, *10*(2), 445. https://doi.org/10.32520/stmsi.v10i2.1303

Sutriani, E., & Octaviani, R. (2019). Keabsahan data. *INA-Rxiv*, 1–22.

Technology, I. (2023). *Media 2023 ANALISIS KETIDAKSEIMBANGAN KELAS DALAM PENGEMBANGAN MODEL KLASIFIKASI*. 602–610.

Wicaksana, A., & Rachman, T. (2018). Implementasi Metode Pembelajaran Reading Aloud dalam Meningkatkan Keaktifan Siswa pada Mata Pelajaran Qur’an Hadist. *Angewandte Chemie International Edition, 6(11), 951–952.*, *3*(1), 10–27. [https://medium.com/@arifwicaksanaa/pengertian-use-case-](https://medium.com/%40arifwicaksanaa/pengertian-use-case-) a7e576e1b6bf

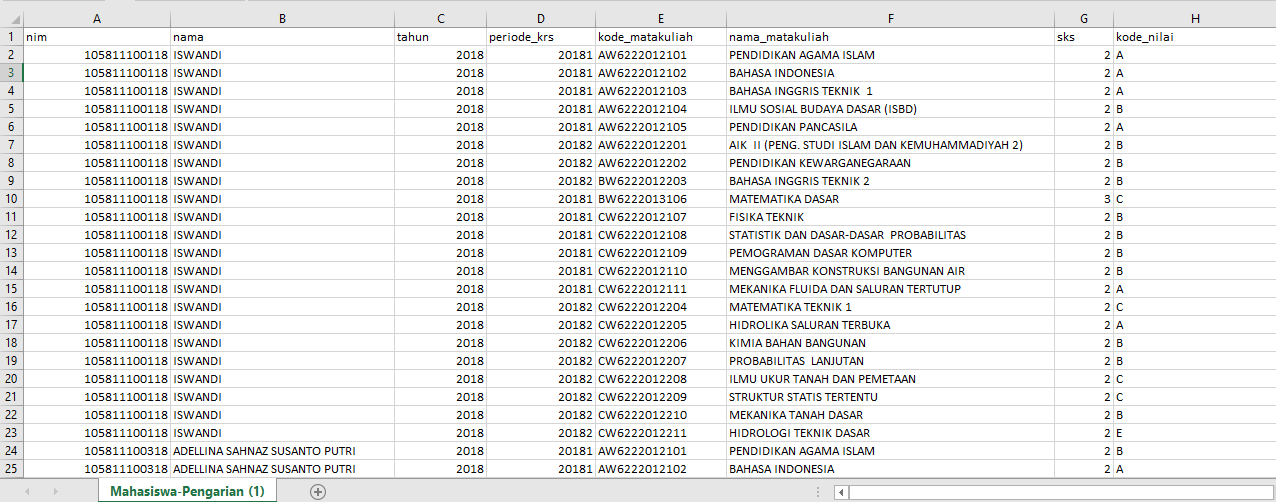
Yustira, N., Witarsyah, D., & ... (2021). Implementasi Algoritma NaÏve Bayes Classification Untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu. *EProceedings …*.

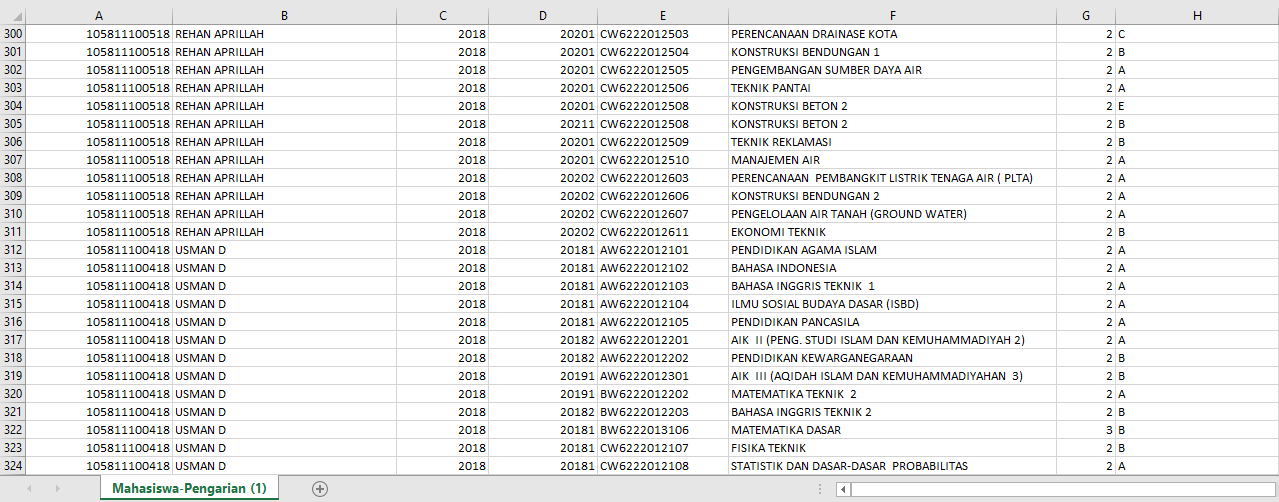
https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/ article/view/16721%0Ahttps://openlibrarypublications.telkomuniversity.ac.id

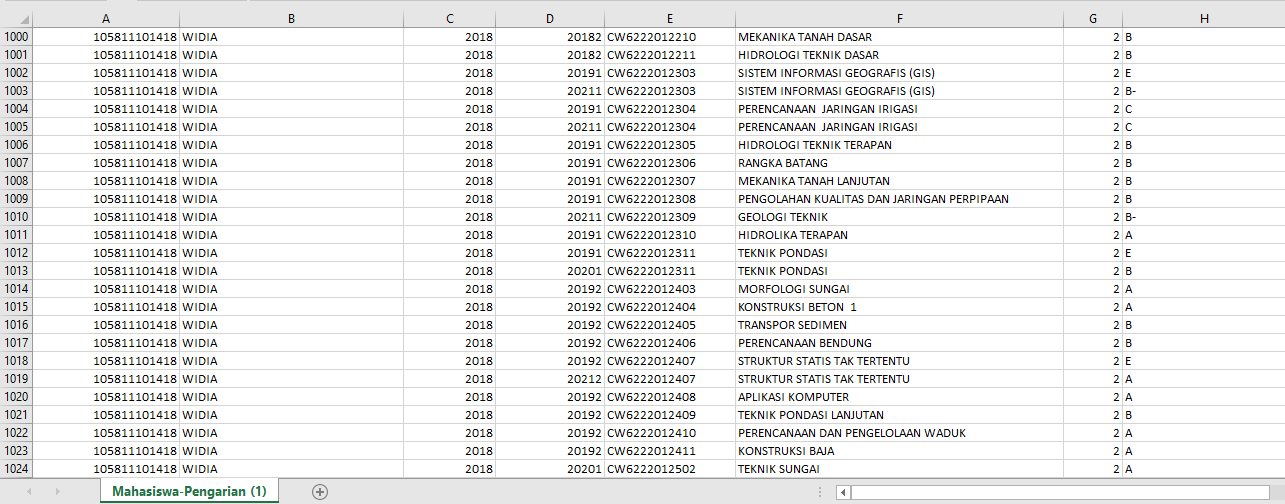
/index.php/engineering/article/view/16721/16429

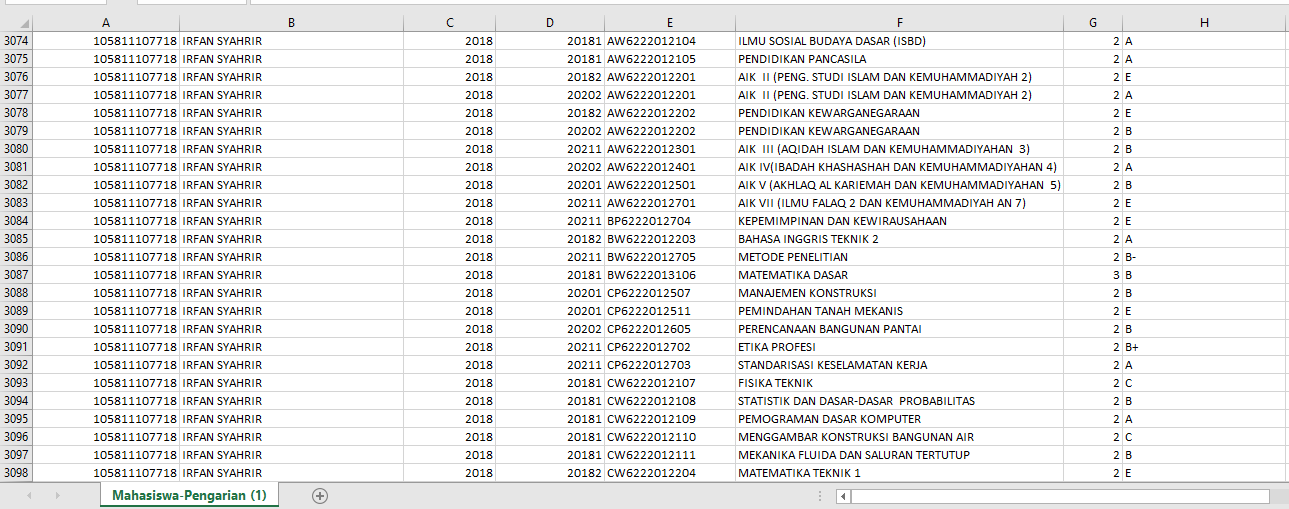
**LAMPIRAN**

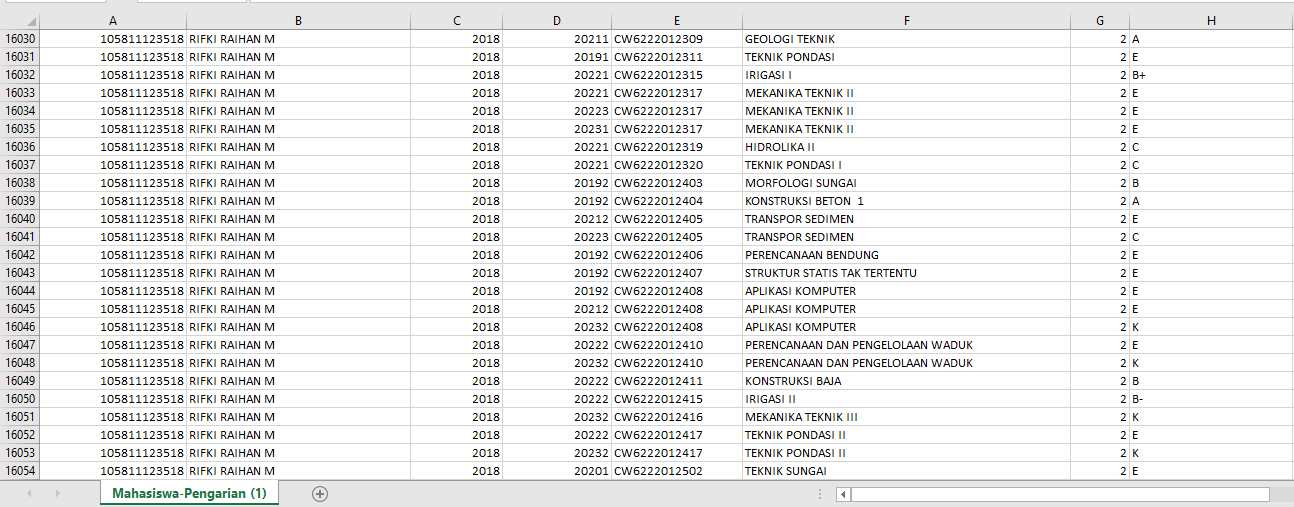
Lampiran 1. Data Mentah Angkatan 2018



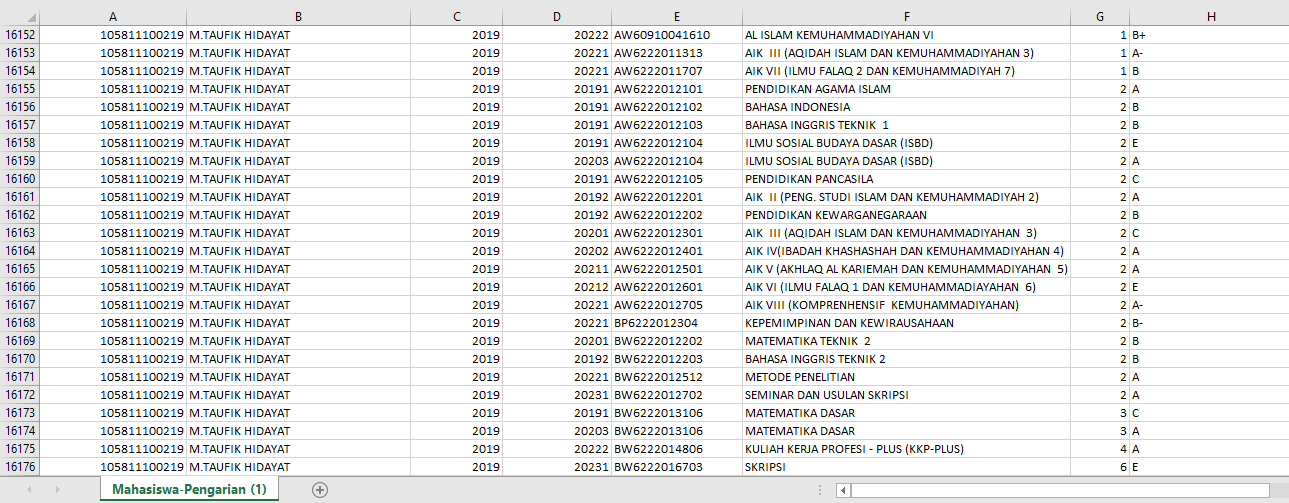


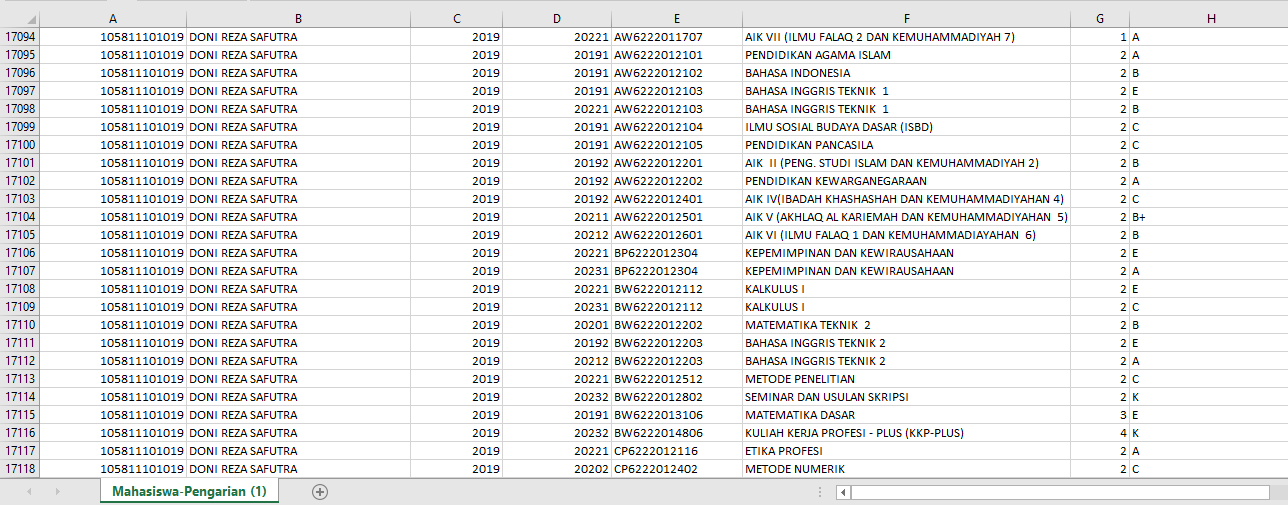


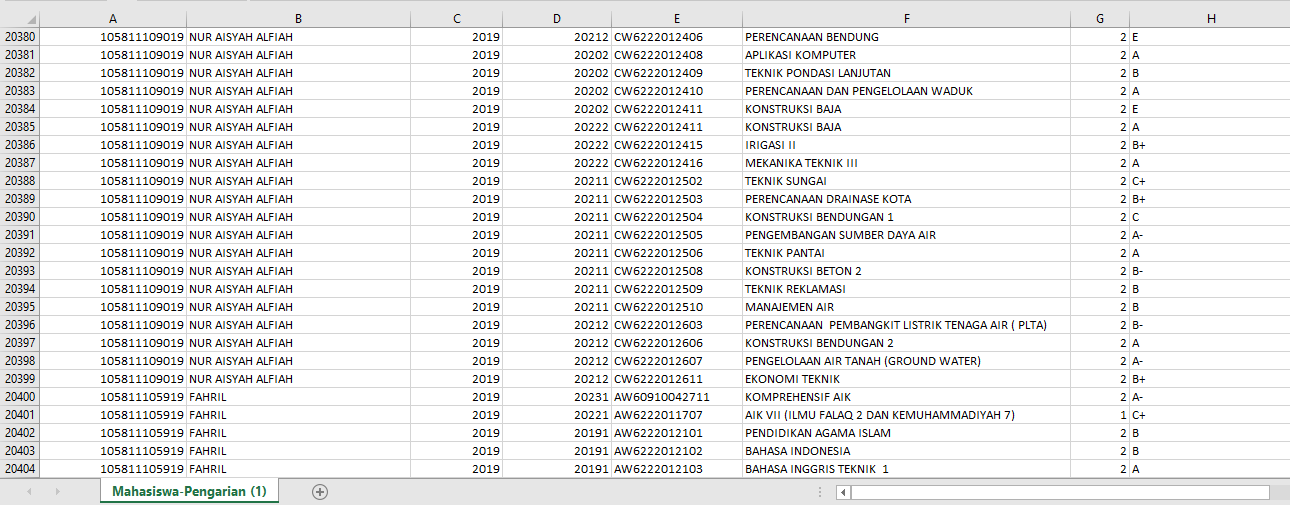


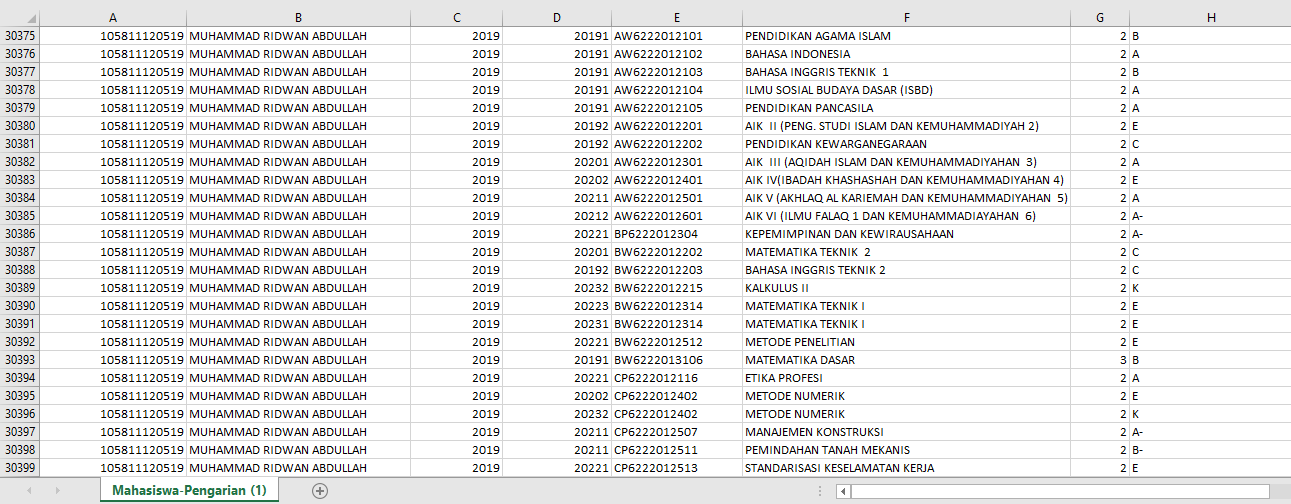


Lampiran 2. Data Mentah Angkatan 2019

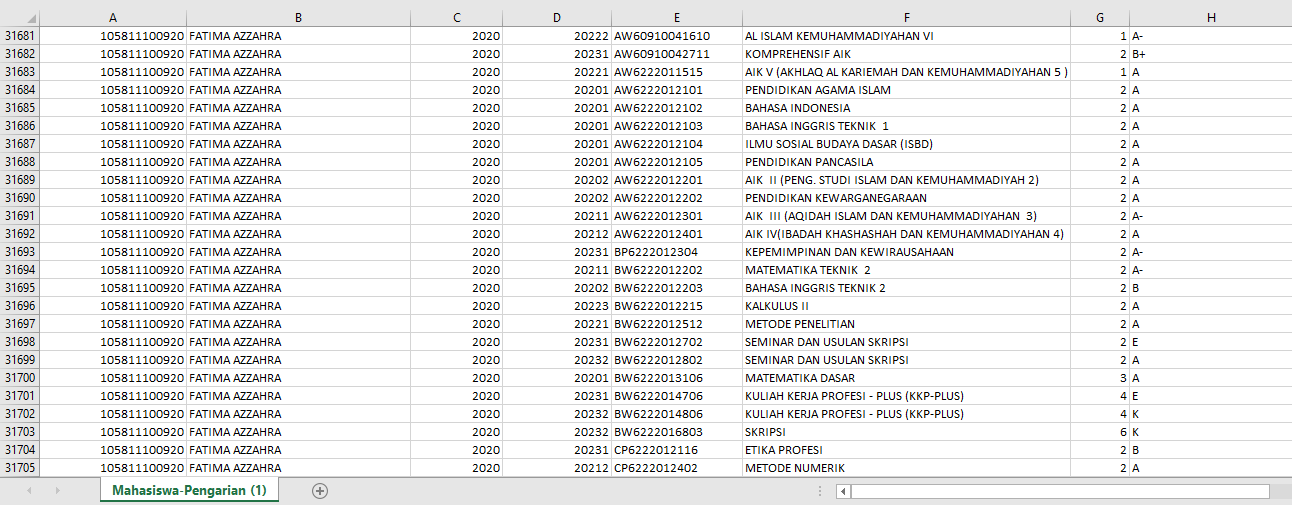


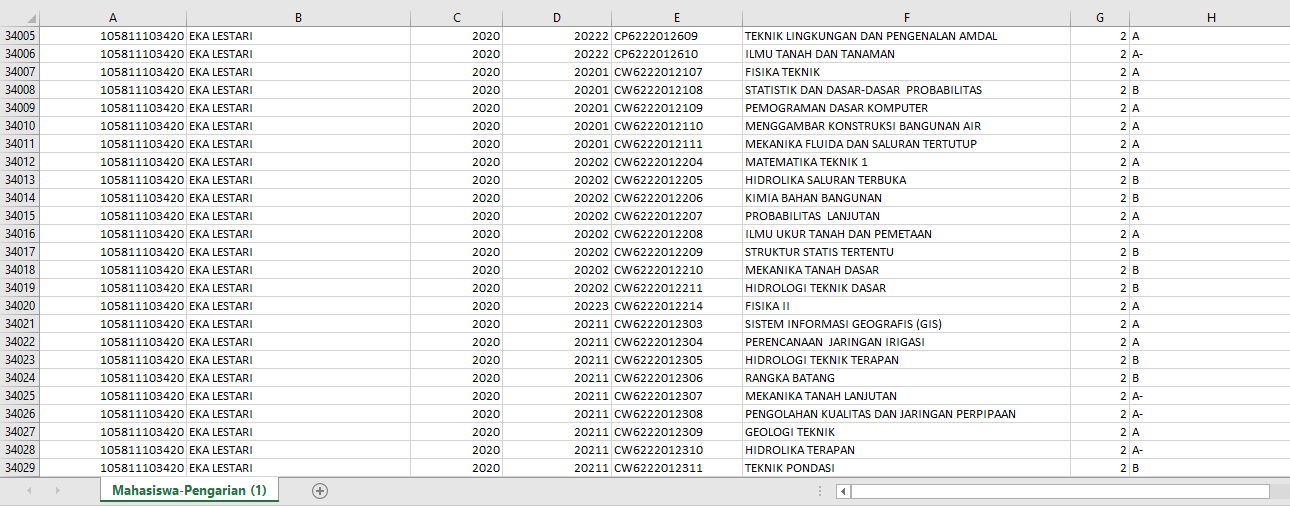


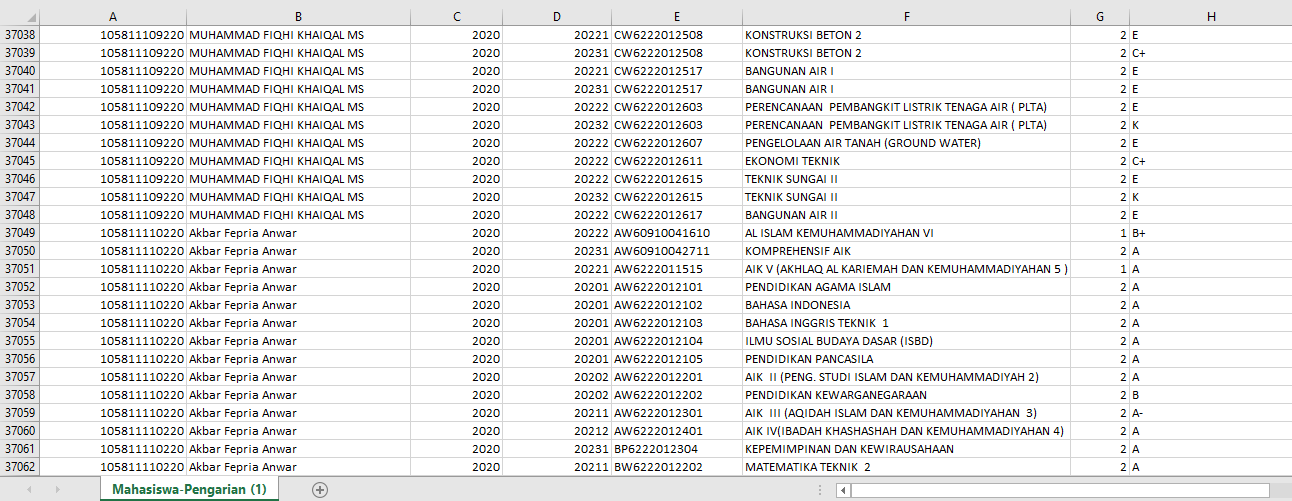




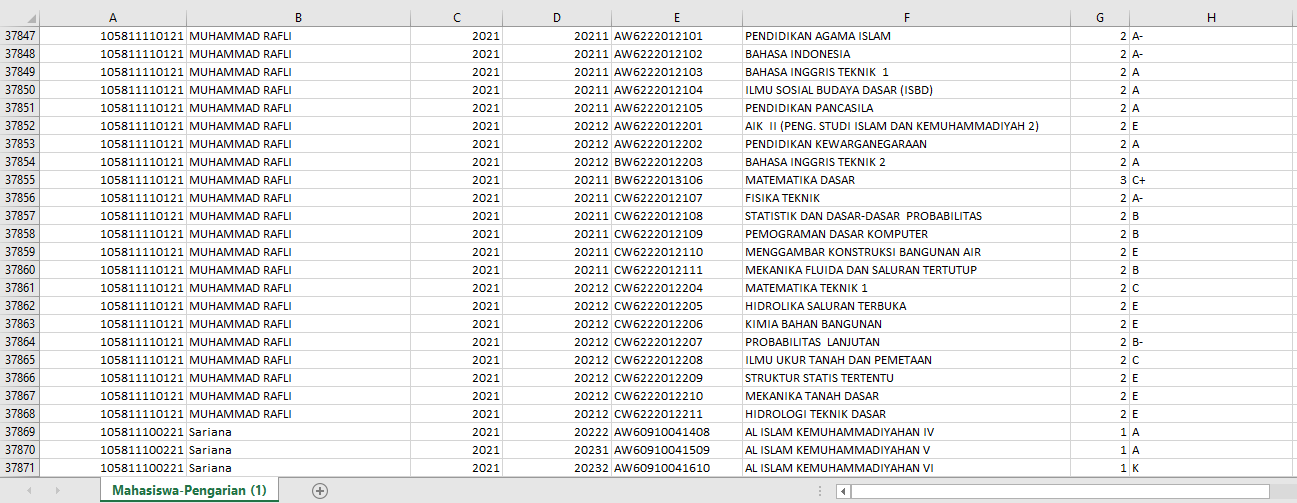
Lampiran 3. Data Mentah Angkatan 2020

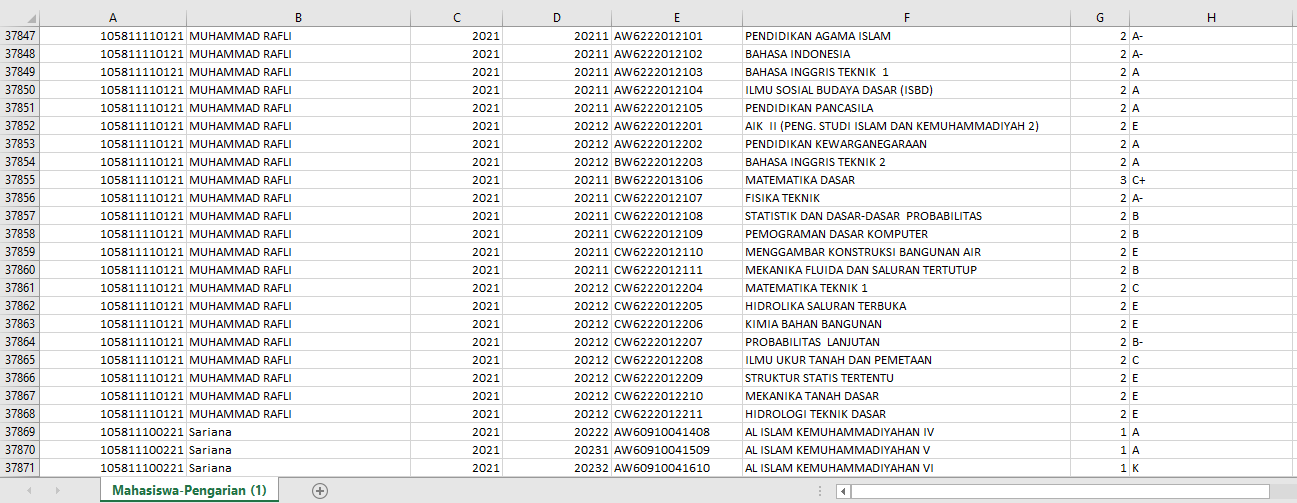


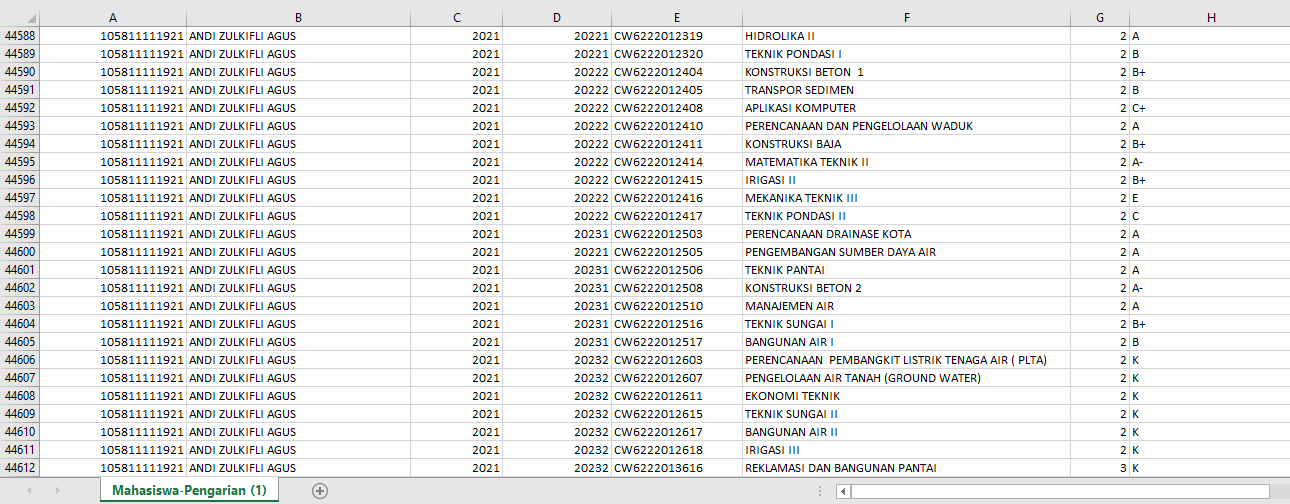




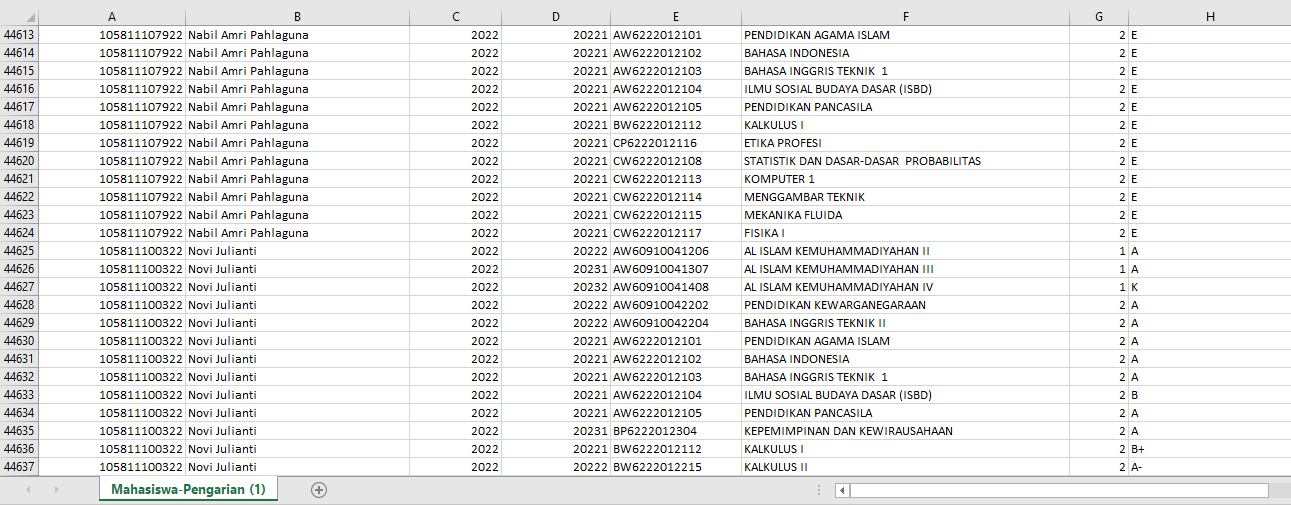
Lampiran 4. Data Mentah Angkatan 2021

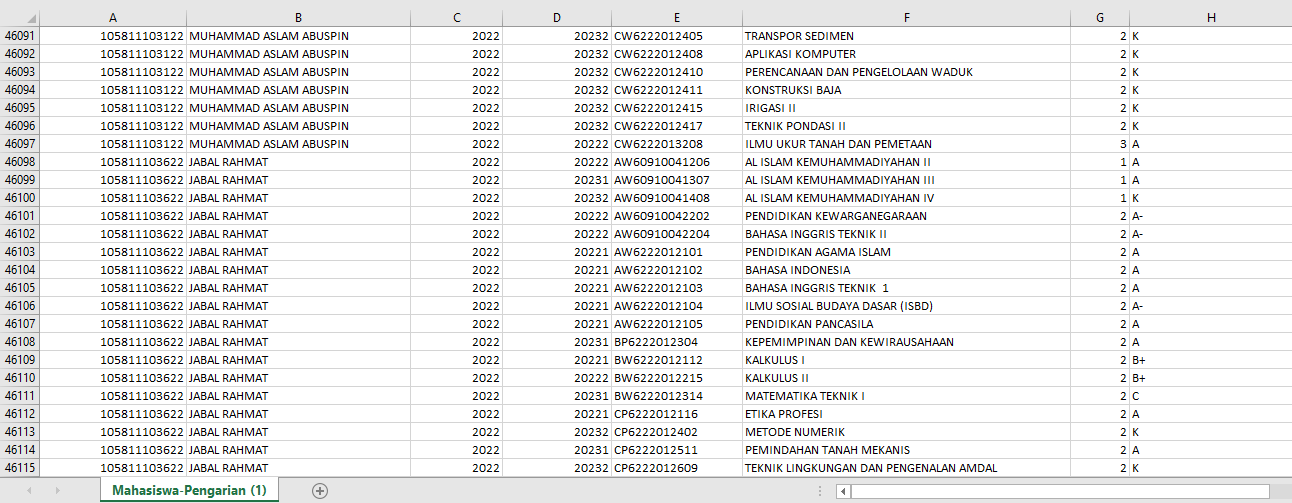


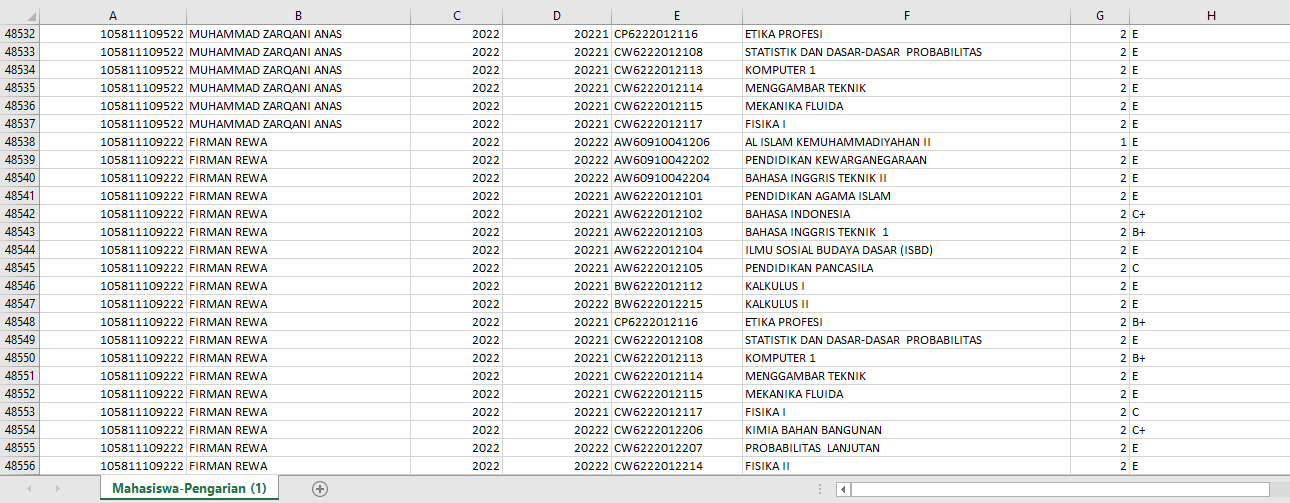




Lampiran 5. Data Mentah Angkatan 2022

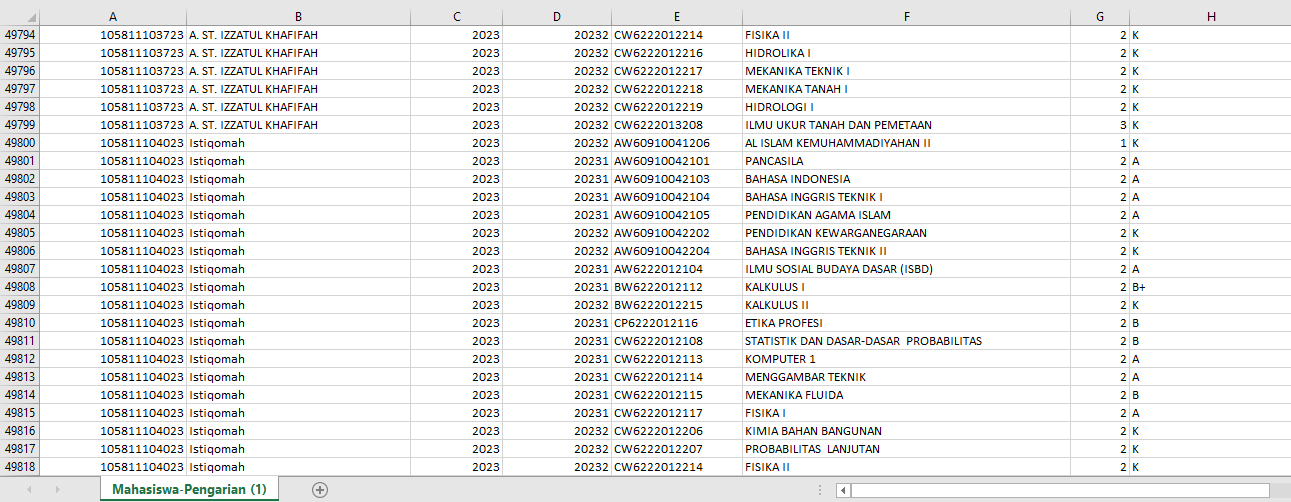


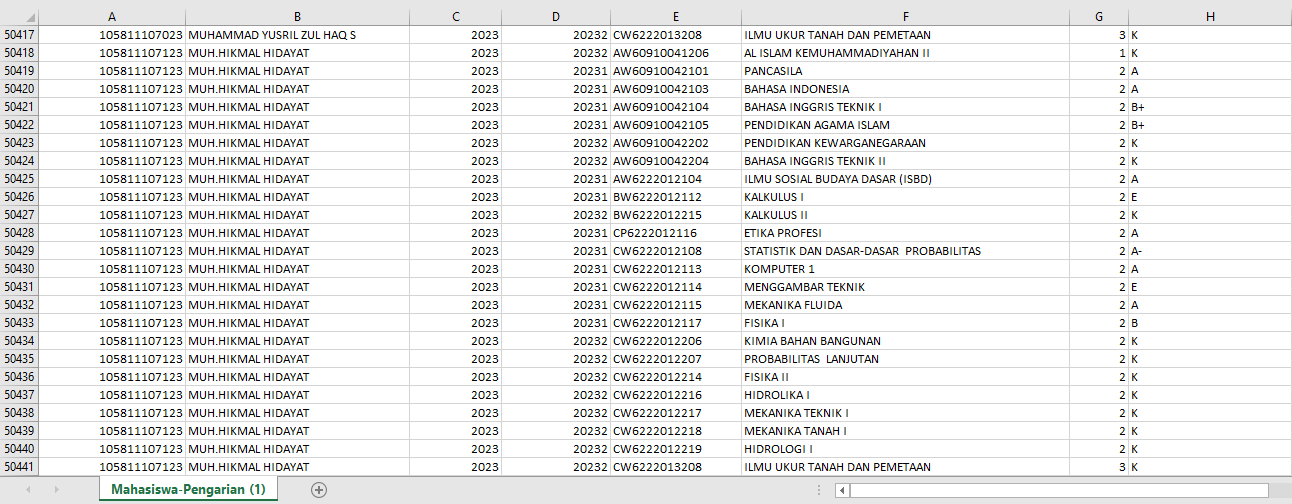




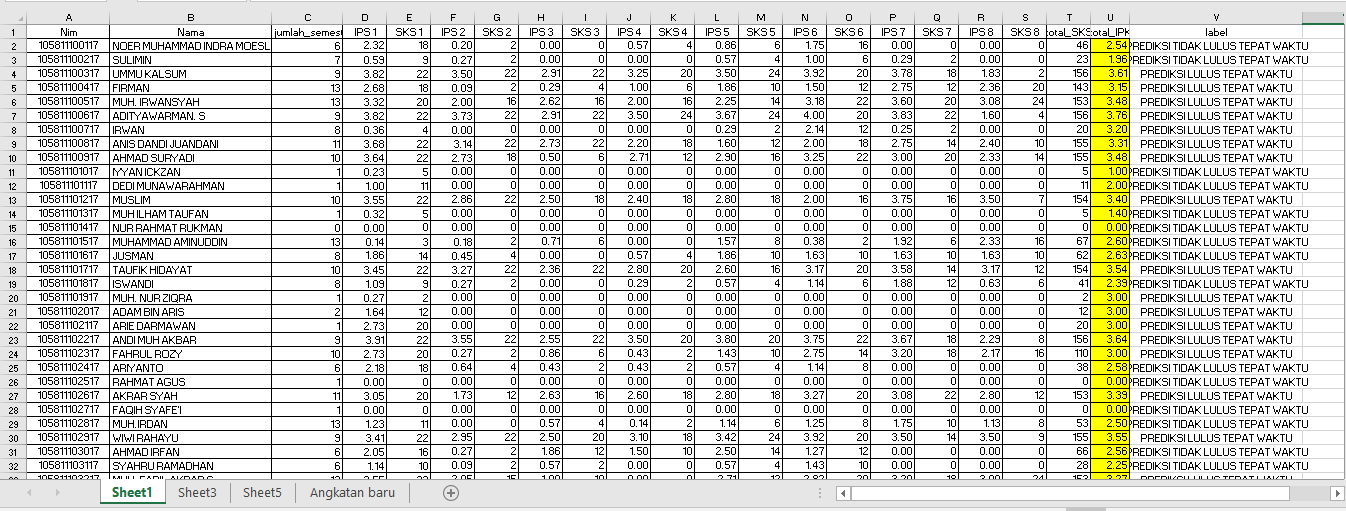
Lampiran 6. Data Mentah Angkatan 2023



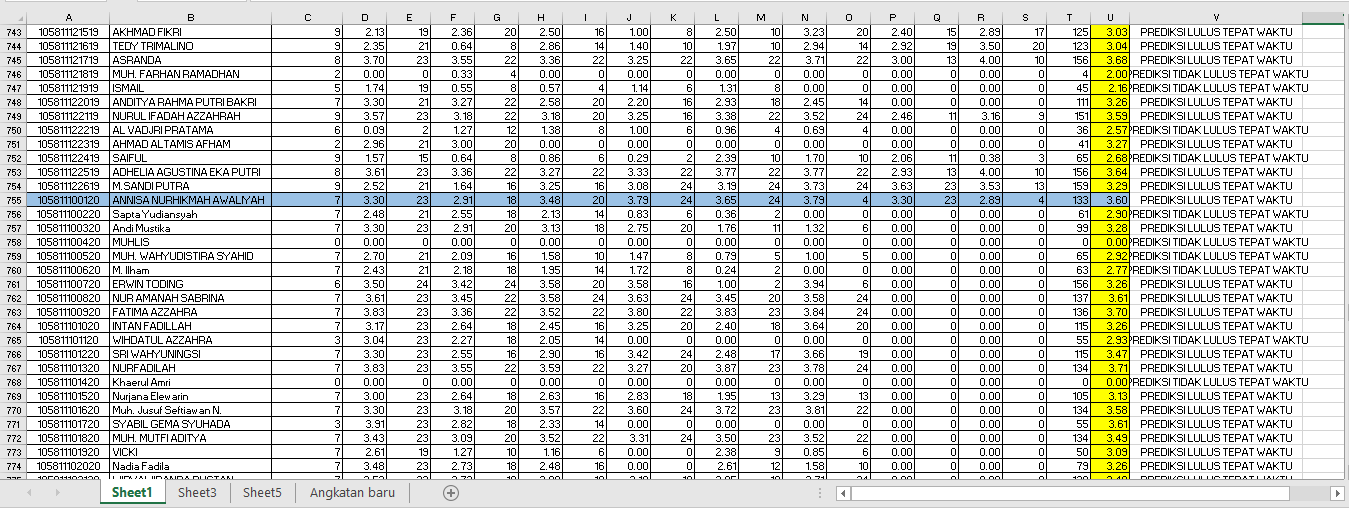


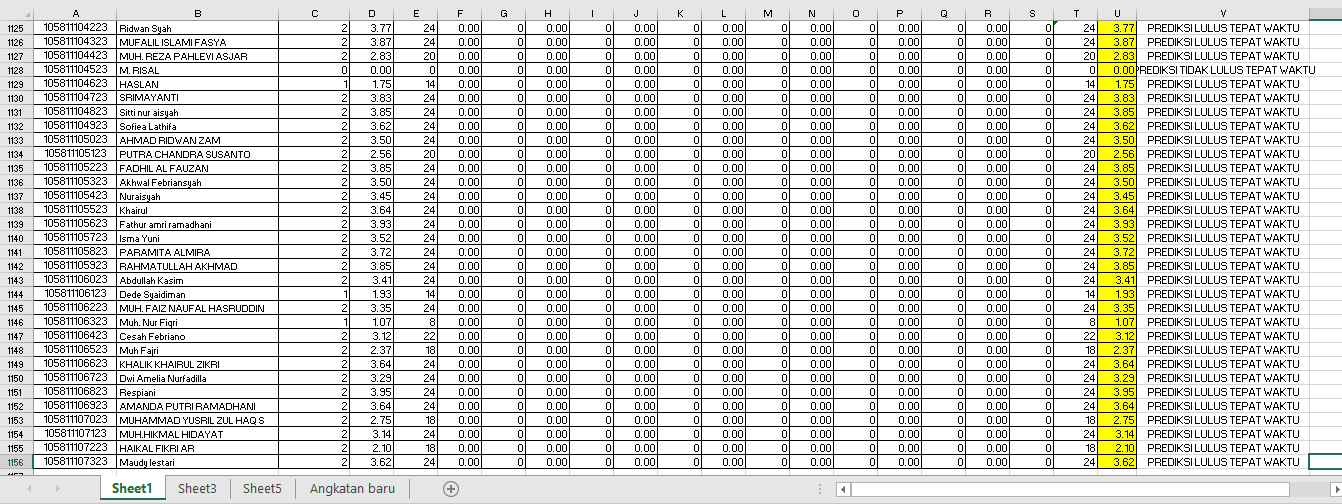


Lampiran 7. Data Mahasiswa Setelah Dilakukan Pelabelan Data

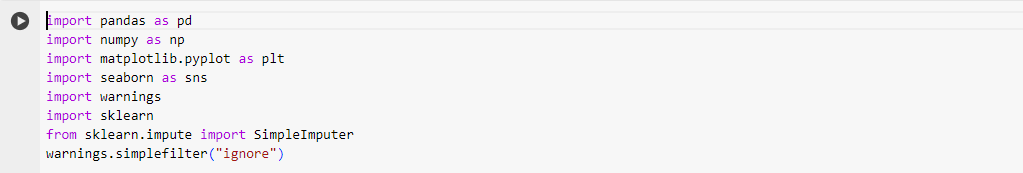








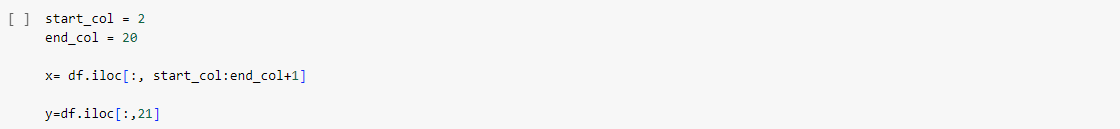
Lampiran 8. Source Code Naïve Bayes Menggunakan SMOTE







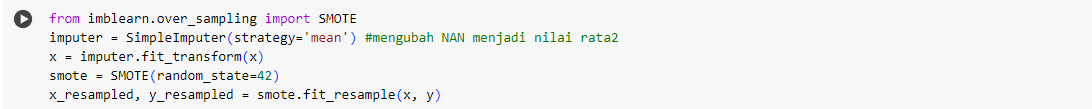


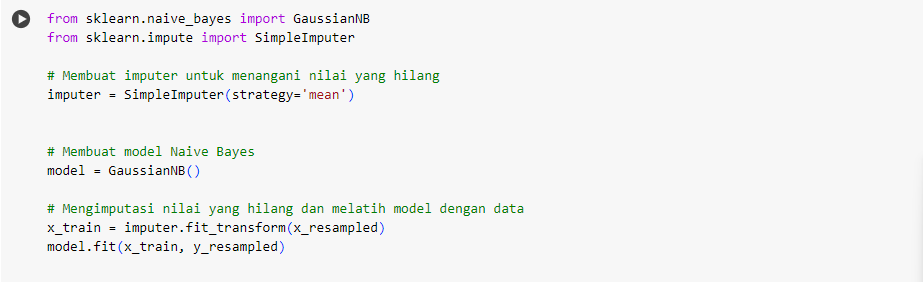


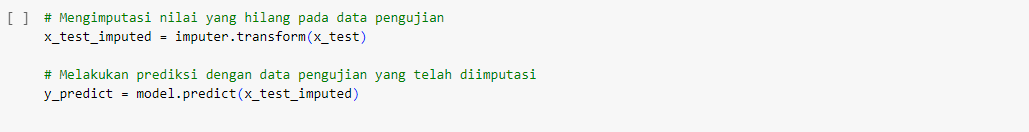


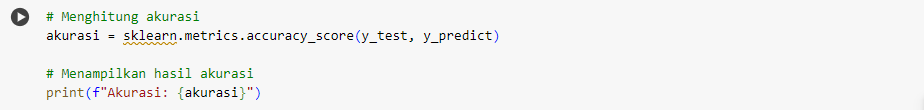


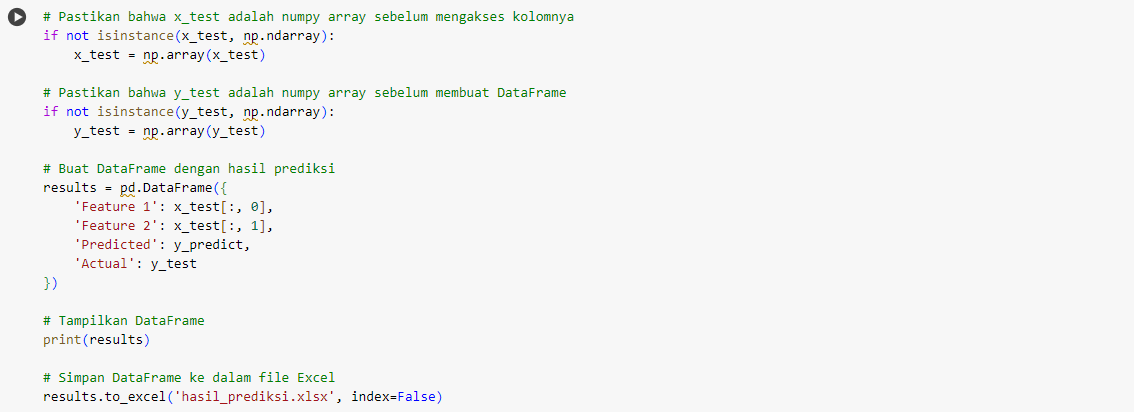








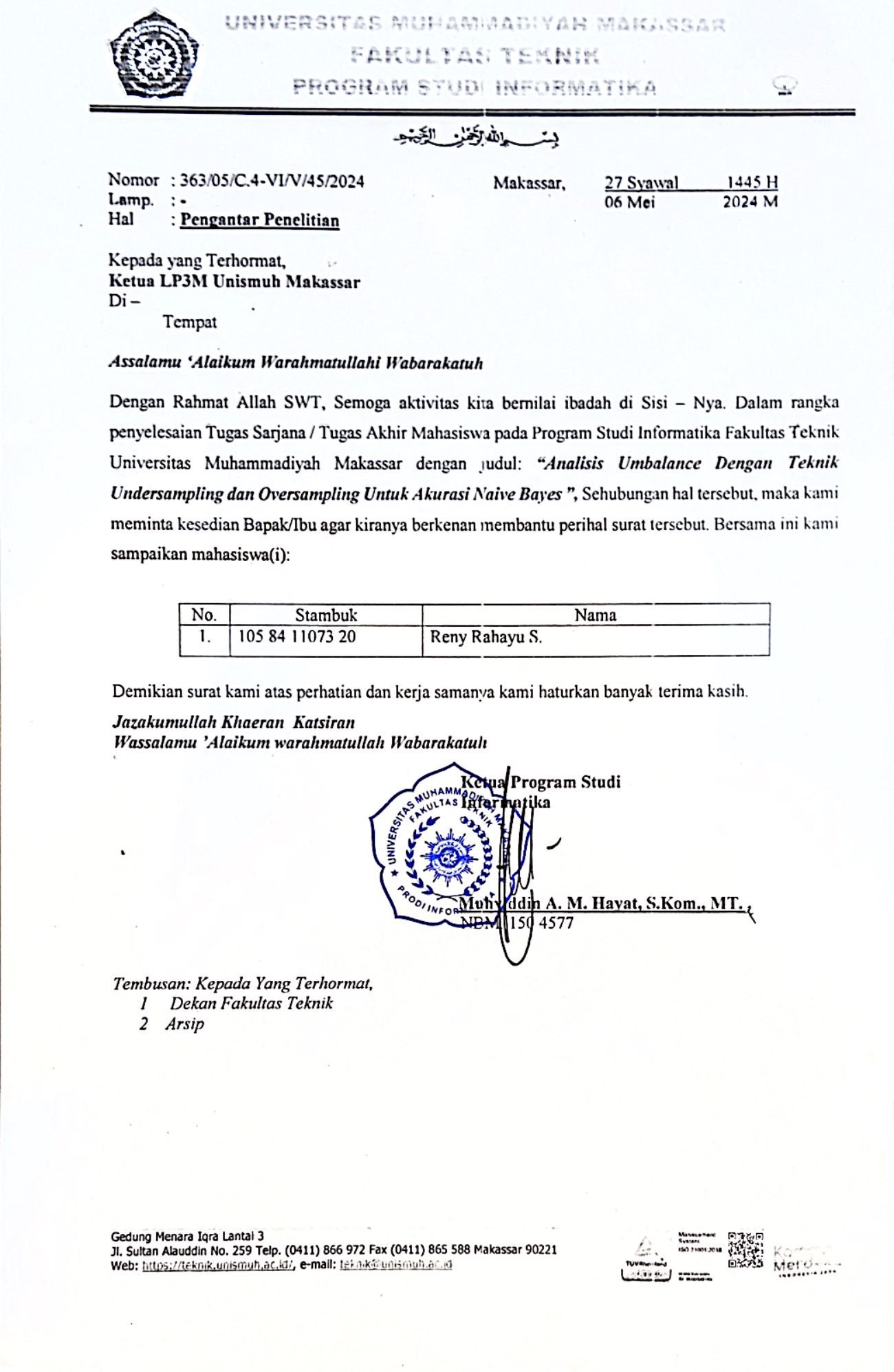




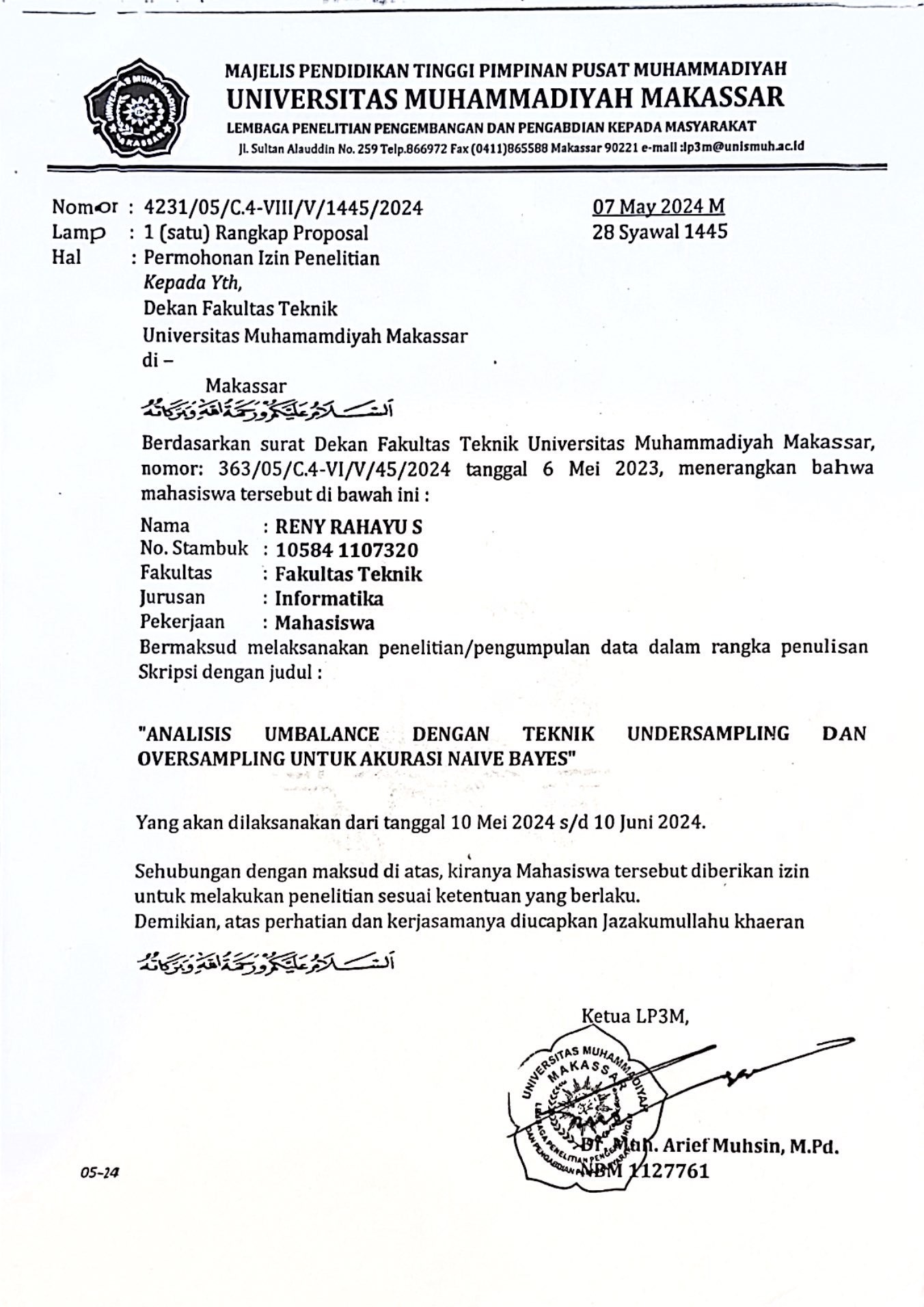
Lampiran 9. Hasil Akurasi Menggunakan SMOTE

|  |  |
| --- | --- |
| Pembagian Data | Hasil Akurasi |
| 90 : 10 | 89,1% |
| 80 : 20 | 91,8% |
| 70 : 30 | 91,3% |

Lampiran 10. Permohonan Penelitian Kepada Kaprodi Informatika



Lampiran 11. Pengantar Penelitian Kepada Ketua Umum LP3M Unismuh Makassar



Lampiran 12. Hasil Scan Plagiasi Per Bab

Lampiran 11. Hasil Scan Plagiasi Per Bab

