Terbit pada laman web jurnal : **http://ojsamik.amikmitragama.ac.id**

JURNAL JARINGAN SISTEM INFORMASI ROBOTIK (JSR)
Vol. 8 No. 1 TAHUN 2024 E - ISSN : 2579-373X

KOMPARASI ALGORITMA MACHINE LEARNING UNTUK KLASIFIKASI KELULUSAN MAHASISWA

Rapel Aprilius Sigit[1], Zuprizal Kurniawan[2], Rahmaddeni[3]
[123]Teknik Informatika, STMIK Amik Riau, Pekanbaru, Indonesia
Email: [1*] rafelas04@gmail.com, [2.] zuprizalkurniawan03@gmail.com, [3]rahmaddeni@sar.ac.id

Abstrak
Keberhasilan seorang mahasiswa dalam menyelesaikan studinya tepat waktu merupakan faktor kunci dalam pencapaian sebuah lembaga pendidikan tinggi. Algoritma machine learning memberikan pendekatan inovatif dalam analisis data serta prediksi berdasarkan pola yang teridentifikasi. Penelitian ini bertujuan membandingkan algoritma machine learning yang umum digunakan dalam mengklasifikasikan kelulusan mahasiswa, seperti Naive Bayes dan Decision Trees. Data yang diterapkan dalam penelitian ini diperoleh dari kaggle.com dan terdiri dari 4424 entri, yang terbagi menjadi tiga kategori: lulus, drop out, dan masih aktif. Data dapat digunakan untuk melakukan pelatihan setelah tahapan preprocessing, meliputi penghapusan data yang tidak relevan serta transformasi yang diperlukan. Setelah tahapan preprocessing selesai, dilakukan implementasi algoritma Naïve Bayes dan Decision Tree. Hasil penelitian menpresentasikan akurasi Naïve Bayes yakni 70,33% dan Decision Tree yakni 67,09%, dengan F1-score Naïve Bayes 61,81% dan Decision Tree 60,80%. Selain itu, hasil cross-validation menunjukkan akurasi Naïve Bayes sebesar 70,00% dan Decision Tree sebesar 68,29%. Dari hasil tersebut, terbukti bahwa Naïve Bayes memiliki performa yang lebih bagus jika dikomparasi dengan Decision Tree dalam konteks penelitian ini.

Kata Kunci: Kelulusan Mahasiswa, Machine Learning, Klasifikasi, Naive Bayes, Decision Tree

Abstract
The success of a student in completing their studies on time is a key factor in the achievement of a higher education institution. Machine learning algorithms provide an innovative approach to data analysis and prediction based on identified patterns. This research aims to compare machine learning algorithms that are commonly used in classifying student graduation, such as Naive Bayes and Decision Trees. The data applied in this study was obtained from kaggle.com and consists of 4424 entries, which are divided into three categories: graduated, dropped out, and still active. The data can be used for training after the preprocessing stage, which includes the removal of irrelevant data and necessary transformations. After the preprocessing stage was completed, the Naïve Bayes and Decision Tree algorithms were implemented. The results presented Naïve Bayes accuracy of 70.33% and Decision Tree accuracy of 67.09%, with F1-score Naïve Bayes 61.81% and Decision Tree 60.80%. In addition, the cross-validation results show Naïve Bayes accuracy of 70.00% and Decision Tree of 68.29%. From these results, it is evident that Naïve Bayes has better performance when compared with Decision Tree in the context of this research.

Keywords: Student Graduation, Machine Learning, Classification, Naive Bayes, Decision Tree.

1. PENDAHULUAN

Keberhasilan seorang mahasiswa lulus sesuai jadwal adalah elemen utama dalam pencapaian sebuah institusi pendidikan tinggi. Dengan meningkatnya jumlah mahasiswa yang diterima, penting untuk memastikan bahwa tingkat kelulusan dalam waktu yang ditentukan juga meningkat. Jika terdapat banyak mahasiswa yang tidak dapat menyelesaikan studi sesuai jadwal, hal tersebut dapat mengakibatkan peningkatan dalam pengelolaan data pribadi dan akademis bagi semua mahasiswa yang masih terdaftar [1].

Pendidikan tinggi dianggap sebagai fondasi utama dalam pembentukan identitas suatu negara, dengan fokus yang kritis pada pengelompokan kelulusan mahasiswa untuk memastikan efisiensi dan kualitas sistem pendidikan. Seiring dengan pertumbuhan data yang cepat dan kompleksitas variabel yang memengaruhi kelulusan, penggunaan teknologi machine learning (ML) menjadi semakin penting untuk meningkatkan prediksi dan pemahaman mengenai faktor-faktor yang mempengaruhi kesuksesan mahasiswa. Evaluasi kesuksesan program studi di perguruan tinggi mempertimbangkan kelulusan mahasiswa sebagai elemen kunci [2].

Peran vital pendidikan tinggi dalam membangun SDM yang berkualitas dan dapat

berkompetisi tidak dapat diabaikan. Pada usaha untuk meningkatkan kualitas pendidikan tinggi, penilaian terhadap kelulusan mahasiswa menjadi fokus utama. Penerapan teknologi machine learning memberikan peluang untuk meningkatkan efisiensi dan ketepatan dalam menentukan status kelulusan mahasiswa [3].

Dalam situasi ini, kepatuhan terhadap tingkat kelulusan mahasiswa bukan hanya mencerminkan prestasi individu, melainkan juga memiliki dampak yang signifikan pada reputasi dan integritas perguruan tinggi. Oleh karena itu, tindakan strategis perlu diambil untuk mendukung mahasiswa agar dapat menyelesaikan studi sesuai jadwal. Institusi pendidikan tinggi harus terus beradaptasi dengan perkembangan teknologi, termasuk pemanfaatan machine learning, guna meningkatkan efektivitas dalam mengelola dan memahami faktor-faktor yang mempengaruhi pencapaian gelar mahasiswa [4]. Aplikasi dapat dimanfaatkan untuk memperkirakan jumlah kelulusan mahasiswa, tetapi ada juga perguruan tinggi belum memiliki sistem yang dapat memperhitungkan keterlambatan kelulusan mahasiswa. Akibatnya, mereka tidak mampu mengambil langkah pencegahan yang diperlukan [5].

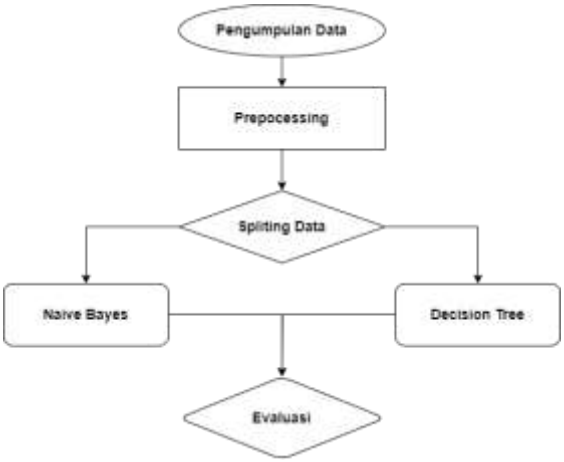
Algoritma machine learning menawarkan pendekatan inovatif dalam pengolahan data dan pembuatan prediksi berdasarkan pola yang terdeteksi. Dalam konteks klasifikasi kelulusan mahasiswa, sejumlah algoritma machine learning telah diajukan dan diimplementasikan. Oleh karena itu, perbandingan atau komparasi antara berbagai algoritma tersebut menjadi esensial guna mengevaluasi kinerja dan efektivitas masing-masing [6].

Klasifikasi, sebagai suatu proses pengelompokan yang terorganisir, mengacu pada teknik pengaturan data atau pengelompokan entitas sesuai dengan aturan yang telah ditentukan. Dalam setiap konteks, klasifikasi melibatkan atribut, termasuk di dalamnya adalah atribut kelas, yang memberikan kelompok pada entitas tersebut [7]. Penerapannya memerlukan pencarian model yang menjelaskan atribut kelas ysng berperan sebagai atribut inputan. Tujuan utama dari klasifikasi adalah mengembangkan model atau algoritma yang dapat meramalkan kelas atau label data berdasarkan atribut-atribut yang dimilikinya [8].

Penelitian ini memiliki tujuan untuk melakukan perbandingan antara beberapa algoritma machine learning yang sering digunakan dalam klasifikasi kelulusan mahasiswa. Algoritma-algoritma tersebut melibatkan Naive Bayes dan Decision Trees. Dengan melakukan perbandingan yang mendalam, penelitian ini bermaksud mengevaluasi keunggulan dan kelemahan tiap algoritma, serta mengidentifikasi algoritma yang paling tepat untuk meningkatkan akurasi prediksi kelulusan mahasiswa.

2. METODOLOGI PENELITIAN

Studi ini dimulai dengan serangkaian tahapan, seperti pengumpulan data yang dibutuhkan sebagai analisis. Data latih serta data uji didistribusikan dengan rasio 70:30. Berikutnya, model naïve bayes dan decision tree dirancang. Langkah terakhir melibatkan pelatihan dan evaluasi kinerja algoritma machine learning yang nantinya akan diuji dalam penelitian. Adapun langkah-langkah ini dipaparkan secara visual pada gambar berikut.



Gambar 1. Alur Penelitian

2.1 Pengumpulan Data

Pada penelitian yang akan diterapkan ini dataset didapat dari website https://www.kaggle.com/datasets/ranzeet013/student-graduation-dataset. Dataset ini terdiri dari 4424 data dengan data latih 70% dan data uji 30%. Kategorisasi data dalam dataset ini terbagi menjadi tiga kelas, yaitu graduation untuk siswa yang berhasil lulus, dropout untuk mahasiswa yang tidak melanjutkan pendidikan mereka dan enrolled untuk mahasiswa yang masih terdaftar atau aktif dalam program pendidikan. Kumpulan data ini menyajikan informasi yang komprehensif mengenai pendaftaran mahasiswa di berbagai kursus, mencakup berbagai aspek seperti karakteristik pribadi, kinerja akademik, dan indikator ekonomi. Data tersebut mencakup berbagai aspek, seperti status perkawinan, cara pendaftaran, dan usia saat pendaftaran, yang menyoroti keragaman mahasiswa dan metode pendaftaran yang mereka pilih. Selain itu, kumpulan data ini mengeksplorasi latar belakang pendidikan mahasiswa, mencatat detail kualifikasi mereka sebelumnya dan kualifikasi orang tua mereka, yang berpotensi memengaruhi perjalanan akademis mereka. Informasi juga mencakup preferensi kehadiran mahasiswa, kewarganegaraan, dan kebutuhan khusus, memberikan pemahaman yang lebih mendalam mengenai populasi mahasiswa.

2.2 Preprocessing

Tahapan ini merupakan proses awal dalam penambangan data yang bertujuan mengubah data menjadi berkualitas, sehingga dapat diproses lebih lanjut. Dalam penelitian, tahapan preprocessing melibatkan proses pembersihan, penyederhanaan huruf, dan tokenisasi pada data [9]. Pada penelitian lain, tahapan preprocessing dilakukan dengan melakukan celaning data, mengintegrasi data, target data, menttransformasi data, dan pemilihan data [10]. Tahapan preprocessing ini penting dilakukan untuk meyakinkan kualitas data dalam proses mining data.

2.3 Spilitting Data

Tahapan kritis dalam pengembangan model machine learning melibatkan proses pelatihan dan pengujian data. Menurut penelitian yang dirujuk, terdapat berbagai metode pembagian data pelatihan dan pengujian yang diterapkan untuk mencapai tingkat akurasi optimal. Sebagai contoh, suatu penelitian memilih untuk membagi dataset dengan rasio 80:20 [11]. Pada umumnya, ada pemisahan antara data pelatihan dan pengujian dalam metode ini. Biasanya, dataset dibagi menjadi dua potongan, di mana satu bagian dimanfaatkan untuk melatih model, sementara yang lainnya untuk menguji performa model. Pembagian data pelatihan dan pengujian bisa bervariasi, seperti 70:30, 80:20, atau 90:10 [12]. Hasil eksperimen menunjukkan bahwa rasio tertentu dalam pembagian data pelatihan dan pengujian dapat berdampak pada akurasi model klasifikasi [13].

2.3.1 Naive Bayes

Naive Bayes digunakan untuk meramalkan peluang, sementara klasifikasi Bayes adalah bentuk klasifikasi statistik yang memproyeksikan kelas suatu entitas berdasarkan probabilitasnya. Naive Bayesian Classifier, yang merupakan bentuk sederhana dari klasifikasi Bayes, diasumsikan bahwa dampak dari nilai atribut pada suatu kelas bersifat independen dari atribut-atribut lainnya [14]. Aturan dasar dari klasifikasi Naive Bayes adalah Teorema Bayes. Teorema Bayes akan dijelaskan dalam persamaan yang mengilustrasikan penggunaannya dalam klasifikasi Naive Bayes.

$$P\ (H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

Keterangan:

X :Data ysng classnya belum diketahui

H :Hipotesis data adalah suatu class spesifik

P(H|X) :Probabilitas hipotesis menurut kondisi (posteriori probability)

P(H) :Probabilitas hipotesis (prior probability)

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

P(X) :Probability H

2.3.2 Decision Tree

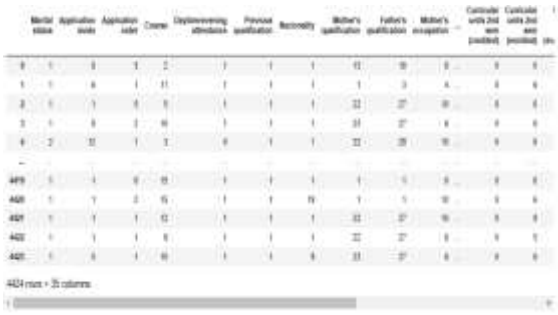
Decision tree merupakan teknik data mining yang sering dipakai pada konstruksi sistem klasifikasi berdasarkan beberapa variabel atau untuk melakukan prediksi. Pendekatan ini mengelompokkan ruang sampel menjadi segmen yang tidak tumpang tindih dan saling melengkapi, di mana masing-masing segmen sesuai dengan simpul daun [15]. Tujuan dari analisis pohon keputusan adalah mengenali model optimal untuk membagi semua catatan menjadi berbagai segmen [16].

2.5 Evaluasi

Untuk mengevaluasi dan menganalisis kinerja algoritma machine learning, pada kasus ini penulis memilih menerapkan metode cross validation ini. Cross validation, juga dikenal sebagai perhitungan rotasi, adalah salah satu teknik yang dimanfaatkan untuk memvalidasi model dengan menggunakan analisis statistik pada sejumlah besar data. Teknik ini sering digunakan untuk mengembangkan model prediktif dan memastikan akurasi pada model-model yang saat ini digunakan. Salah satu metode cross validation yang paling general adalah K-fold cross validation, pada hal ini data dibagi menjadi k bagian yang persis [17].

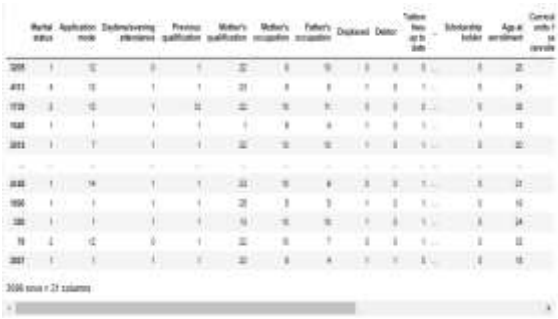
3. HASIL DAN PEMBAHASAN

Hasil yang didapat pada pengujian yang diterapkan dengan Jupyter Notebook dari kasus yang memiliki data terdiri dari 4424 row dan 35 kolom data yang menunjukkan adanya tiga label yang terklasifikasi, yaitu graduation untuk siswa yang berhasil lulus, dropout untuk mahasiswa yang tidak melanjutkan pendidikan mereka dan enrolled untuk mahasiswa yang masih terdaftar atau aktif dalam program pendidikan, dengan menggunakan rasio perbandingan untuk semua metode ialah 70:30. Dan juga kita menggunakan perbandingan 2 algoritma yaitu naïve bayes dan decision tree.



Gambar 2. Kumpulan Data yang Digunakan

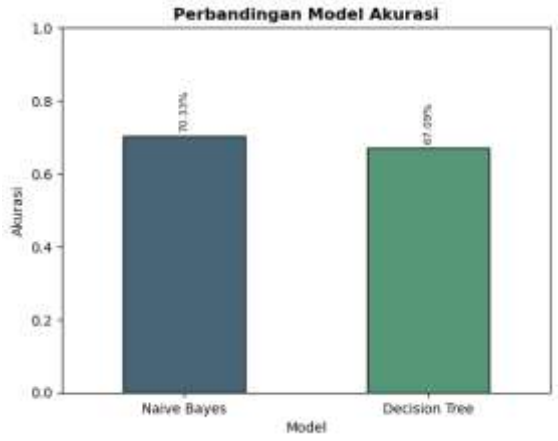
Sebelum data dapat dimasukkan ke dalam model algoritma machine learning untuk pembelajaran, tahap preprocessing. Hal ini esensial untuk memastikan bahwa data sesuai dengan kebutuhan model algoritma, yang mencakup penghapusan data yang tidak memiliki nilai atau bernilai nol, serta melakukan transformasi data yang diperlukan.



Gambar 3. Data Setelah Prepocessing

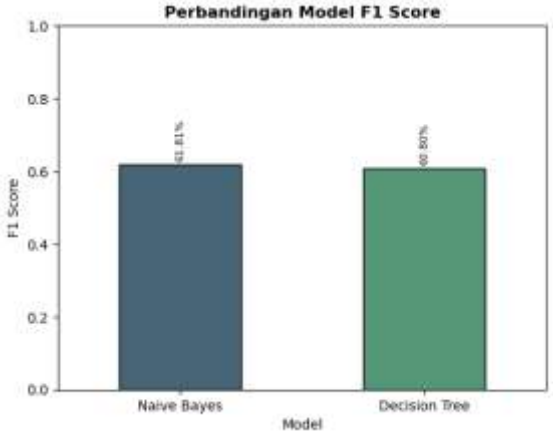
Setelah melalui tahap preprocessing, yang mengurangi jumlah data dari 4424 menjadi 3927 dengan format yang sesuai, langkah selanjutnya adalah membagi data tersebut dengan perbandingan 70-30 antara data latih dan data uji. Setelah data dibagi dilakukan, langkah berikutnya adalah mengimplementasikan algoritma Naïve Bayes dan Decision Tree.

Berikut adalah hasil komparasi akurasi dari algoritma machine learning yang telah diuji, disajikan dalam diagram batang berikut ini.



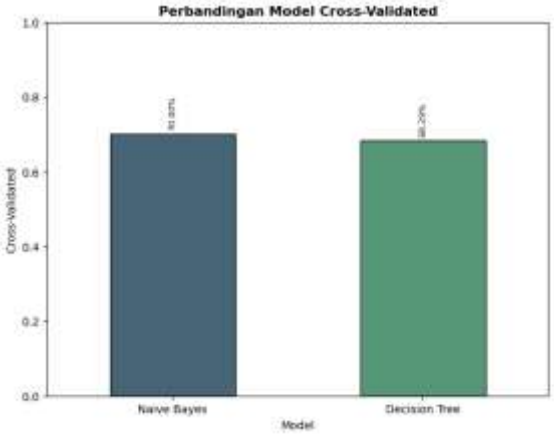
Gambar 4. Perbandingan Akurasi

Dibawah ini terdapat hasil perbandingan skor F1 dari dua model algoritma machine learning yang diuji, disajikan dalam diagram batang berikut ini.



Gambar 5. Perbandingan F1 Score

Berikut ini terlihat hasil evaluasi terhadap dua model algoritma machine learning yang telah diuji, disajikan dalam diagram batang berikut ini.



Gambar 5. Perbandingan Evaluasi Prediksi

Hasil pengujian dan visualisasi yang telah dilakukan dapat disajikan dalam bentuk tabel perbandingan berikut.

Tabel 1. Hasil Perbandingan Data Latih dan Data Uji 70:30

Algoritma	Akurasi	F1 Score	Evaluasi Prediksi
Naive Bayes	70,33%	61,81%	70,00%
Decision Tree	67,09%	60,60%	68,29%

Tabel 2. Hasil Perbandingan Data Latih dan Data Uji 80:20

Algoritma	Akurasi	F1 Score	Evaluasi Prediksi

Naive Bayes	71,53%	61,84%	70,00%
Decision Tree	67,57%	61,64%	67,52%

Tabel 3. Hasil Perbandingan Data Latih dan Data Uji 90:10

Algoritma	Akurasi	F1 Score	Evaluasi Prediksi
Naive Bayes	70,43%	60,73%	70,00%
Decision Tree	69,75%	64,65%	68,04%

4. KESIMPULAN

Berdasarkan studi kasus yang sudah dilakukan dengan menerapkan Dataset kelulusan mahasiswa dari keggle.com yang dikumpulkan berjumlah 4424 data. Dataset ini terbagi menjadi tiga kelas, yaitu graduation untuk siswa yang berhasil lulus, dropout untuk mahasiswa yang tidak melanjutkan pendidikan mereka dan enrolled untuk mahasiswa yang masih terdaftar atau aktif dalam program pendidikan. Pada pengujian memakai metode naive bayes dan decission tree.

Dalam penelitian ini, terjadi proses pre-processing yang melibatkan pembagian data menjadi dua bagian yang terdiri data training dan testing. Selanjutnya, kedua jenis data tersebut diklasifikasikan menggunakan metode decision tree dan naive bayes. Hasil evaluasi menunjukkan akurasi naive bayes sebesar 70,33% dan decision tree sebesar 67,09%. Sementara itu, f1-score untuk naive bayes adalah 61,81% dan untuk decision tree adalah 60,80%. Dalam cross-validation, nilai akurasi naive bayes mencapai 70,00%, sedangkan decision tree mencapai 68,29%. Dari hasil ini, terbukti bahwa algoritma naive bayes mempunyai kinerja yang lebih bagus dibandingkan decision tree dalam penelitian ini.

Untuk penelitian berikutnya, dapat dilakukan langkah pre-processing tambahan guna mengurangi gangguan pada data, menggunakan aplikasi yang lebih efisien, dan mengadakan eksperimen dengan metode alternatif guna meningkatkan akurasi hasil.

5. REFERENCES

[1] E. Haryatmi and S. Pramita Hervianti, "Penerapan Algoritma Support Vector Machine Untuk Model Prediksi Kelulusan Mahasiswa Tepat Waktu," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 2, pp. 386–392, Apr. 2021, doi: 10.29207/resti.v5i2.3007.

[2] H. Priyatman, F. Sajid, and D. Haldivany, "Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan Mahasiswa," *JEPIN (Jurnal Edukasi dan Penelitian Informatika)* , vol. 5, pp. 62–66, Apr. 2019.

[3] L. Setiyani, M. Wahidin, D. Awaludin, and S. Purwani, "Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Data Mining Naïve Bayes : Systematic Review," *Faktor Exacta*, vol. 13, no. 1, p. 35, Jun. 2020, doi: 10.30998/faktorexacta.v13i1.5548.

[4] Nurul Khasanah, Agus Salim, Nurul Afni, Rachman Komarudin, and Yana Iqbal Maulana, "Prediksi Kelulusan Mahasiswa Dengan Metode Naive Bayes," Jul. 2022.

[5] M. R. Qisthiano, T. B. Kurniawan, E. S. Negara, and M. Akbar, "Pengembangan Model Untuk Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu dengan Metode Naïve Bayes," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 3, p. 987, Jul. 2021, doi: 10.30865/mib.v5i3.3030.

[6] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, vol. 7, no. 4, pp. 432–439, Dec. 2021, doi: 10.1016/j.icte.2021.02.004.

[7] A. Nata and Suparmadi, "Analisis Sistem Pendukung Keputusan Dengan Model Klasifikasi Berbasis Machine Learning Dalam Penentuan Penerima Program Indonesia Pintar," Oct. 2022. [Online]. Available: http://jurnal.goretanpena.com/index.php/JSSR

[8] Refido Arjunal Akmal and Aliyah Kurniasih, "Penerapan Algoritma Klasifikasi untuk Menangani Data Tidak Seimbang pada Peningkatan Kualitas Siswa," Jakarta, Oct. 2023. [Online]. Available: https://www.kaggle.com/code/spscientist/student-performance-in-exams/input

[9] L. Hermawan and M. B. Ismiati, "Pembelajaran Text Preprocessing berbasis Simulator Untuk Mata Kuliah Information Retrieval," *TRANSFORMATIKA*, vol. 17, no. 2, pp. 188–199, Jan. 2020.

[10] R. Rahmaddeni, M. K. Anam, Y. Irawan, S. Susanti, and M. Jamaris, "Comparison of Support Vector Machine and XGBSVM in Analyzing Public Opinion on Covid-19 Vaccination," *ILKOM Jurnal Ilmiah*, vol. 14, no. 1, pp. 32–38, Apr. 2022, doi: 10.33096/ilkom.v14i1.1090.32-38.

[11] Arsya Fathiarahma, Apriade Voutama, Taufik Ridwan, and Nono Heryana, "Analisis Text Mining Klasifikasi Kegiatan Keluarga Menggunakan Orange Dengan Metode Naïve Bayes," *Jurnal Teknologi Terpadu*, vol. 9, pp. 35–41, Jul. 2023.

[12] Siti Aisyah, Sri Wahyuningsih, and Fidia Deny Tisna Amijaya, "Peramalan Jumlah Titik Panas Provinsi Kalimantan Timur Menggunakan Metode Radial Basis Function Neural Network," *Jambura Journal of Probability and Statistics*, vol. 2, no. 2, pp. 64–74, Nov. 2021, doi: 10.34312/jjps.v2i2.10292.

[13] Baiq Nurul Azmi, Arief Hermawan, and Donny Avianto, "Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver," *JTIM : Jurnal Teknologi Informasi dan Multimedia*, vol. 4, no. 4, pp. 281–290, Feb. 2023, doi: 10.35746/jtim.v4i4.298.

[14] H. M. Siregar, "Implementasi Metode Naive Bayes Pada Perancangan Aplikasi Sistem Pakar Diagnosa Bronkiektasis," *Bulletin of Information Technology (BIT)*, vol. 1, no. 3, pp. 112–121, Nov. 2020.

[15] S. A. Arnomo, "Analisa Decision Tree untuk Kepuasan Penggunaan Sinyal dari Base Transceiver Station (BTS)," *Jurnal Sistem dan Teknologi Informasi (Justin)*, vol. 9, no. 2, pp. 199–205, Apr. 2021, doi: 10.26418/justin.v9i2.43425.

[16] Bahzad Taha Jijo and Adnan Mohsin Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *JOURNAL OF APPLIED SCIENCE AND TECHNOLOGY TRENDS*, vol. 02, no. 01, pp. 20–28, 2021.

[17] I. K. Nti, O. Nyarko-Boateng, and J. Aning, "Performance of Machine Learning Algorithms with Different K Values in K-fold Cross-Validation," *International Journal of Information Technology and Computer Science*, vol. 13, no. 6, pp. 61–71, Dec. 2021, doi: 10.5815/ijitcs.2021.06.05.