DECODE: JURNAL PENDIDIKAN TEKNOLOGI INFORMASI
P-ISSN: 2775-2984 | E-ISSN: 2775-1813

Vol. 4, No. 1, Maret 2024, Hal: 191-203
Doi: http://dx.doi.org/10.51454/decode.v4i1.308
https://journal.umkendari.ac.id/index.php/decode

Prediksi Kelulusan Tepat Waktu Berdasarkan Riwayat Akademik Menggunakan Metode *Naïve Bayes*

Imam Riadi[1], Rusydi Umar[2], Rio Anggara[2]*

[1] Program Studi Sistem Informasi, Universitas Ahmad Dahlan, Indonesia.
[2] Program Studi Informatika, Universitas Ahmad Dahlan, Indonesia.

Artikel Info

Kata Kunci:

Confusion Matrix;
Data Mining;
Klasifikasi;
Naïve Bayes;
Prediksi Kelulusan.

Keywords:

Confusion Matrix;
Data Mining;
Classification;
Naive Bayes;
Graduation Prediction.

Riwayat Article:

Submitted: 7 Agustus 2023
Accepted: 10 Oktober 2023
Published: 23 Januari 2024

Abstrak: Kelulusan tepat waktu mahasiswa memiliki dampak besar dalam dunia pendidikan. Namun, tidak semua mahasiswa, mampu mencapai prestasi tersebut. Oleh karena itu, diperlukan penelitian yang mendalam dalam menganalisis data kelulusan sebagai upaya mendukung mahasiswa agar berhasil menyelesaikan studi mereka tepat waktu. Penelitian data kelulusan bisa dilakukan menggunakan tehnik klasifikasi data mining. Klasifikasi merupakan salah satu pengolahan dalam data mining dilakukan dengan cara mengelompokkan dengan metode tertentu. Penelitian ini membangun aplikasi dengan implementasi metode *naive bayes* dengan mempertimbangkan parameter menghasilkan klasifikasi mahasiswa lulus tidak tepat waktu dan lulus tepat waktu. Tahapan pada penelitian seperti *load data, cleaning data, selection data, transformation data,* data training, data testing, dan hasil prediksi. Tahapan pengujian akurasi penelitian menggunakan metode *confusion matrix* mendapat akurasi 72% dengan total penerapan data sejumlah 291 dengan detail 273 data training dan 18 data testing. Hasil akurasi menunjukkan bahwa sistem prediksi kelulusan dapat digunakan FTI UAD sebagai salah satu acuan dan pertimbangan fakultas mengambil langkah-langkah kelulusan mahasiswa.

Abstract: *The timely graduation of students has a significant impact on the world of education. However, not all students manage to achieve this milestone. Therefore, in-depth research is needed to analyze graduation data as an effort to support students in successfully completing their studies on time. Research on graduation data can be conducted using data mining classification techniques. Classification is one of the data processing methods in data mining, carried out by grouping data using specific methods. This research builds an application implementing the Naive Bayes method, considering various parameters to classify students into those who graduate on time and those who do not. The research process involves several stages, including data loading, data cleaning, data selection, data transformation, data training, data testing, and result prediction. The accuracy of the research is evaluated using the confusion matrix method, which yields an accuracy rate of 72%. This accuracy is obtained from a total of 291 data points, comprising 273 data points for training and 18 data points for testing. The accuracy results indicate that the graduation prediction system can be used by FTI UAD as a reference and consideration in making decisions related to student graduation.*

Corresponding Author:
Rio Anggara
Email: rio2107048006@webmail.uad.ac.id

PENDAHULUAN

Kelulusan dunia perkuliahan merupakan hal yang paling ditunggu oleh orang tua dan pelajar. Mahasiswa yang terlambat lulus, diberikan waktu untuk menyelesaikan tugas akhir sebagai syarat kelulusan (Nugraha et al., 2018). Kampus memiliki kebijakan untuk memberi batas waktu maksimal, supaya mahasiswa menyelesaikan tugas akhir. Apabila tidak memenuhi batas maksimal maka akan mendapat sanksi DO. Universitas Ahmad Dahlan (UAD) menjadi salah satu lembaga pendidikan swasta terkenal di Yogyakarta. Kampus UAD salah satunya terletak di Jalan Prof. Dr. Soepomo Janturan Daerah Istimewa Yogyakarta 55166. Kampus UAD memiliki 11 fakultas pada tahun 2023, salah satunya Fakultas Teknologi Industri (FTI). FTI UAD terdapat 5 prodi seperti Teknologi Pangan, Tehnik Informatika, Tehnik Kimia, Tehnik Elektro dan Tehnik Industri. Kebijakan penyelesaian tugas akhir dibuat oleh bidang kemahasiswaan, dimana merupakan salah satu aspek penilaian penting keberhasilan kelulusan mahasiswa (Priyatman et al., 2019). Ketepatan waktu mahasiswa memiliki kriteria yang berbeda di setiap program. Setiap Universitas selalu mengusahakan yang terbaik untuk kemajuan kampusnya, termasuk dalam hal Akreditasi setiap jurusan (Hidayat, 2021). Akreditasi menjadi keharusan utama oleh pihak kampus untuk menunjukkan kampus terbaik dan layak untuk menjadi tempat menimba ilmu calon mahasiswa. Label akreditasi jurusan pada setiap ijazah kelulusan menjadi hal sangat dipertimbangkan dalam dunia kerja (Amri, 2020). Badan Akreditasi Nasional Perguruan Tinggi (BAN-PT) dalam Buku III Pedoman Penyusunan Borang, kualitas Universitas di Indonesia parameter dari akreditasi yang dilakukan pihak BAN-PT (Yalidhan & Amin, 2018). Akreditasi dipengaruhi oleh kelulusan mahasiswa, apabila kelulusan tidak lebih dari total mahasiswa baru ditahun berikutnya maka dipastikan akreditasi akan turun. Seorang mahasiswa dapat melanjutkan akademik jika memenuhi beberapa spesifikasi syarat akademik (Dwi et al., 2019). Mahasiswa yang tidak memenuhi persyaratan akademik akan dikenakan drop out (DO). Penurunan jumlah mahasiswa yang lulus terlambat atau menghentikan kuliah bisa dicapai dengan cara mengenali mahasiswa-mahasiswa tersebut dan membuat aturan-aturan yang mendorong mereka untuk mengikuti kurikulum (Larasati et al., 2019). Selain itu, ketika waktu lulus dapat diprediksi, manajemen mahasiswa akan lebih efektif dalam mengurangi jumlah mahasiswa yang mendapatkan sanksi akademik DO. Salah satu teknik untuk menghasilkan prediksi semacam itu adalah teknik data mining.

Data mining adalah teknik dalam kemajuan digital berfungsi meringankan perusahaan atau universitas mencari data informasi penting di gudang data (Saputra & Sibarani, 2020). Data mining dalam arti lain sebuah proses mengekstrak informasi dan pengetahuan baru membantu dalam pengambilan keputusan yang berasal dari sebuah database (Ramadhan & Utami, 2019). Adanya data dalam jumlah besar dan kebutuhan akan informasi, untuk mendukung pengambilan keputusan dalam rangka menciptakan cara berbisnis dan infrastruktur pendukung di bidang teknik komputer merupakan awal munculnya teknologi data mining. Menjadikan informasi sebagai solusi pendukung keputusan dalam dunia perusahaan dan pendidikan (Nugraha et al., 2018). Penelitian dengan algoritma *Naive Bayes Classifier (NBC)* untuk memprediksi kelulusan mahasiswa (Rachmat et al., 2020). NBC adalah metode pengelompokkan atau klasifikasi statistik sederhana dengan menghitung sekelompok kemungkinan melalui tahapan menghitung frekuensi kemunculan suatu data dan campuran dari nilai data telah diberi (Falade et al., 2019). Dengan metode ini dapat diketahui secara detail data yang diperlukan dalam penelitian. Seperti pada umumnya setiap instansi perusahaan maupun perguruan tinggi memiliki data di setiap harinya yang biasanya di kalkulasikan dalam bentuk data bulanan maupun tahunan (Rustam & Annur, 2019).

Data penelitian menggunakan data FTI UAD, seperti data sebelum masuk kuliah dan data mahasiswa setelah lulus kuliah. Data mahasiswa terdiri variabel NIM, nama mahasiswa, asal sekolah, asal provinsi, rata – rata matematika, IPK, TOEFL dan lama studi. Kumpulan data sebelum kuliah dan setelah kuliah di fakultas belum banyak dilakukan analisa lebih lanjut (Puspita & Widodo, 2021). Pada penelitian 2019 dengan objek penelitian mata kuliah Pengantar Teknologi Informasi (PTI) menggunakan 3 item yaitu nilai tugas, nilai UTS dan nilai UAS (Pradnyana & Permana, 2018). Penggunaan algoritma yang diperbandingkan 4 algoritma. Seperti *Naive Bayes (NB)*, *Neural Network (NN)*, *Logistic Regression (LR)* dan *Support Vector Machine (SVM)*. Total semua data diambil 175

mahasiswa, hasil akurasi metode *Naive Bayes* sebesar 89,7%. Hasil akurasi kelulusan dengan algoritma *Neural Network* sebesar 85,7%. Sedangkan algoritma *Logistic Regression* akurasi kelulusan 88,6%. Algoritma menggunakan *Support Vector Machine* akurasi sebesar 95,4%. Hasil riset menunjukkan SVM merupakan klasifikasi yang sangat *robust* (Shedriko & Firdaus, 2022). Pada riset ini, 3 jenis kayu jati digunakan: Semarang, Blora dan Sulawesi. Analisis tekstur menggunakan metode *Gray Level Co-occurrence Matrix* (GLCM) dan jarak spasial 1 piksel. Berdasarkan pengujian dan analisis, algoritma k-NN biasanya diklasifikasikan menjadi 3 jenis kayu jati yaitu Semarang, Blora dan Sulawesi dengan tingkat akurasi diatas 70%. Klasifikasi terbaik untuk kayu jati sulawesi adalah metode *Naive Bayes* dengan tingkat akurasi 82,7%. Karena metode *Naive Bayes* memberikan akurasi paling tinggi meskipun mengandung sedikit data latih, maka akurasi algoritma k-NN dipengaruhi oleh banyaknya data latih yang digunakan dalam pencarian (Waliyansyah & Fitriyah, 2019).

Studi tahun 2020 menggunakan 75 data latih dan 25 data uji. Atribut yang digunakan dalam segmentasi pelanggan adalah jumlah pembelian, jangka waktu, dan lokasi. Hasil pengujian dari 25 pengujian sistem klasifikasi data terdapat 23 jawaban benar dan 2 jawaban salah. Hasil dengan menggunakan metode *Confusion Matrix* menunjukkan nilai akurasi mencapai 92%, nilai presisi mencapai 100%, nilai *recovery* mencapai 91% (Putro et al., 2020). Penelitian pada tahun 2021 menggunakan data penerapan klasifikasi karya ilmiah(tugas akhir) dari perpustakaan Teknik Informatika Universitas Malikussaleh sebanyak 170 data, mencakup 150 data latih dan 20 sebagai data uji. Hasil penelitian mendapat akurasi rata-rata 86,68%, dan waktu pemrosesan rata-rata untuk setiap pengujian adalah 5,7406 detik (Nurdin et al., 2021).

Penelitian lain dengan implementasi aplikasi WEKA, hasil proses klasifikasi kelas pilihan lulus terdiri 15 kategori. Kelas pilihan pertama lebih tinggi yaitu 274 data dan pilihan kedua 73 data dan 14 data gagal. Seleksi kelas dua dan algoritma tidak lulus hingga 93,6288% dari total 338 data dan data rahasia, tetapi tidak cocok. Nilai persentase akurasi rendering efektif dari dataset yang diterima oleh pengklasifikasi Naïve Bayes, hingga 94% (Manullang et al., 2022).

Penelitian ini tentang klasifikasi kelulusan dengan fokus penelitian data mahasiswa berdasarkan riwayat terdahulu yang telah lulus dengan penerapan metode NBC. Sedangkan penelitian terdahulu menggunakan metode NBC bukan menggunakan objek mahasiswa. Metode tersebut akan mengklasifikasikan 2 kategori kemungkinan kelulusan mahasiswa. Hasil penelitian berupa hasil 2 klasifikasi yaitu lulus tidak tepat waktu atau lulus tepat waktu. Dengan klasifikasi ini akan didapatkan prediksi mahasiswa. Penelitian ini dengan pengolahan *data mining* menggunakan *tools framework streamlit* dan *library* dari *python* dengan menggunakan bahasa pemrograman *Python*. Penelitian akan lebih akurat apabila jumlah dataset lebih banyak dan lebih variatif atribut yang digunakan. Dengan adanya penelitian prediksi kelulusan diharapkan menjadi dasar pengambilan keputusan menunjang mahasiswa bisa lulus tepat waktu, meminimalisir mahasiswa lulus tidak tepat waktu atau sampai DO. Penelitian ini juga sebagai pengembangan metode dalam data mining sekaligus kontribusi dalam ilmu pengetahuan.

METODE

Penelitian ini dilakukan untuk klasifikasi data mahasiswa berdasar riwayat akademik mahasiswa yang telah lulus.

Naïve Bayes berdasar pada *Teorema Bayes* dengan rumus sebagai berikut:

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

H adalah hipotesis dalam penelitian. X adalah bukti dalam penelitian. P(X) merupakan probabilitas hipotesis H benar untuk bukti X. P(X|H) merupakan probabilitas bukti X benar untuk hipotesis H. P(H) merupakan robabilitas prior (awal) hipotesis H dan P(X) merupakan probabilitas prior (awal) bukti X. Teorema Bayes adalah dasar dari banyak metode statistik dan algoritma yang

digunakan dalam berbagai bidang, termasuk dalam pemrosesan data dan *machine learning*. Pembahasan beberapa konsep yang relevan dengan dataset penelitian:

1. Probabilitas Awal (P(H)) atau *prior probability*, dalam konteks dataset, bisa merujuk pada probabilitas bahwa suatu kejadian atau hipotesis tertentu terjadi sebelum melihat data yang ada. Sebagai contoh, jika kita ingin memprediksi kelulusan mahasiswa berdasarkan beberapa fitur seperti nilai, durasi studi, dan lainnya, P(H) bisa mengacu pada probabilitas bahwa seorang mahasiswa akan lulus tepat waktu sebelum melihat data spesifik mereka.
2. Probabilitas Akhir (P(X|H)) atau likelihood probability, dalam konteks dataset, mengacu pada probabilitas bahwa data yang diamati (X) muncul ketika hipotesis tertentu (H) diterapkan. Dalam kasus di atas, ini bisa berarti probabilitas bahwa data seperti nilai rata-rata mahasiswa atau durasi studi yang diberikan akan muncul jika mahasiswa tersebut lulus tepat waktu (H).

Penerapan Teorema Bayes pada dataset kelulusan akan melibatkan penggunaan informasi probabilitas awal (*prior probability*) dan probabilitas akhir (*likelihood probability*) untuk memperbarui pemahaman kita tentang suatu hipotesis setelah melihat data. Dalam konteks machine learning, ini sering digunakan dalam algoritma seperti Naive Bayes, yang memperkirakan probabilitas kelas (seperti lulus atau tidak lulus) berdasarkan fitur-fitur yang diamati (seperti nilai, durasi studi, dll.). Ketika diterapkan pada dataset kelulusan, *Teorema Bayes* membantu dalam membuat prediksi atau mengambil keputusan berdasarkan bukti yang ada dalam dataset.

Pengumpulan data penelitian adalah tahapan pengambilan data dan analisis secara sistematis maupun objektif, pengumpulan data yang berfokus pada tujuan menemukan informasi yang peneliti butuhkan untuk mencapai tujuan peneliti (Ramadhayanti, 2018). Data yang terkumpul selanjutnya disebut dengan nama *dataset*. Setelah database terkumpul selanjutnya data proses seleksi secara manual tanpa menggunakan sistem, seleksi ini bertujuan mendapatkan data yang *valid* sebelum dilakukan pemrosesan kedalam sistem. Total *record* dataset yang digunakan sebelum dilakukan *filter* oleh sistem sebanyak 291 dataset. Setelah melakukan penyaringan data *noise* yang *valid*, dihasilkan 93 data. Dataset penelitian yang telah berhasil dikumpulkan terdiri dari atribut NIM, nama mahasiswa, nama SLTA, Prodi (Program Studi), TTL (Tempat Tanggal Lahir), Provinsi, MTK (rata-rata matematika), tahun masuk, IPK, lama studi dan lama studi hari. Dalam penelitian ini hanya menggunakan beberapa atribut tidak semua atribut. Atribut yang akan diolah untuk prediksi ada 6 atribut seperti atribut nama SLTA, Prodi (Program Studi), asal Provinsi, MTK (rata-rata matematika), IPK dan lama studi hari. Atribut pada dataset yang akan digunakan dalam penelitian seperti terlihat pada Tabel 1.

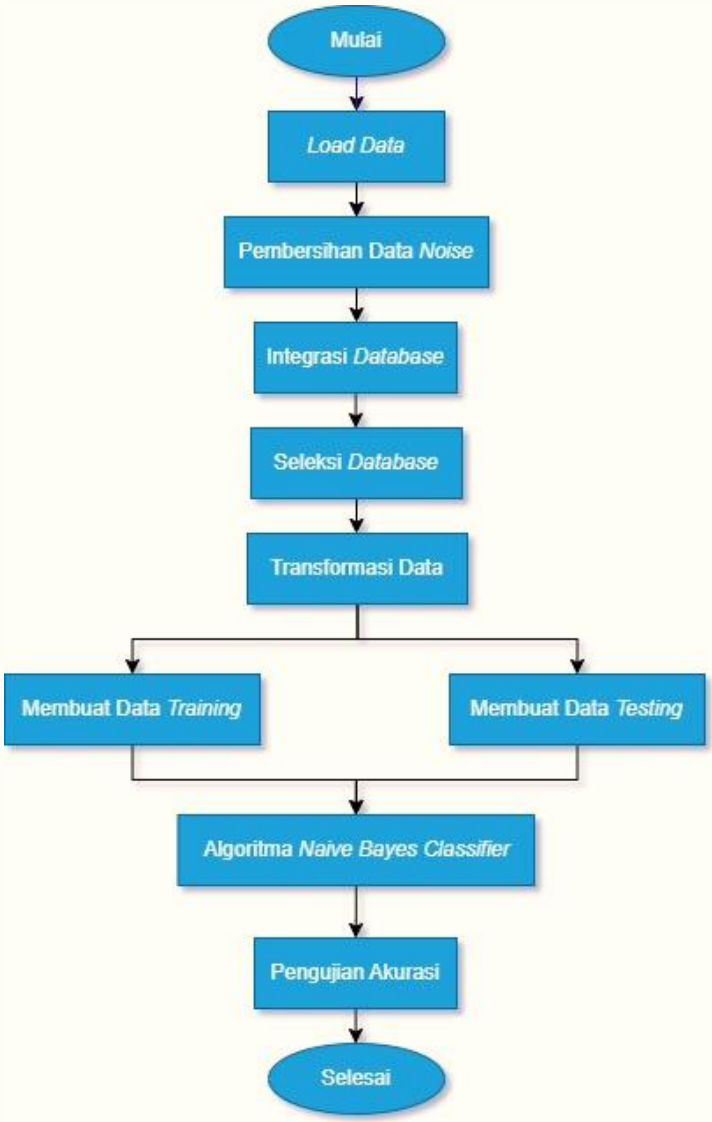
Tabel 1. Hasil Pengumpulan Data Penelitian

No.	NIM	Nama	Nama SLTA	Prodi	Provinsi	MTK	IPK	TOEFL	Lama Studi	Lama Studi Hari
1	1400018012	Mackands Leonardo O	SMA Gadjah Mada	TI	DI Yogyakarta	83	3,46	460	3 Th,11 Bln,30 Hr	1455
2	1400018016	M Rasyid Ridho	MA Negeri 3	TI	DI Yogyakarta	78	3,26	400	3 Th,11 Bln,30 Hr	1455
3	1400018017	Dedy Saputra	SMA Negeri 1, Babakan	TI	Jawa Barat	89	3,57	406	4 Th,1 Bln,23 Hr	1513
....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
92	1500018310	Faiz Isnan A	SMA KH Mustofa	TI	Jawa Barat	82	2,99	416	3 Th,11 Bln,14 Hr	1439

Penjelasan atribut yang akan diolah menjadi kategorial yaitu atribut lama studi. Sedangkan atribut yang akan menjadi numerial yaitu atribut nama SLTA, prodi, provinsi, MTK, IPK, dan TOEFL.

Beberapa penjelasan tentang atribut yang digunakan seperti atribut Prodi merupakan atribut asal program studi saat menempuh sarjana, atribut provinsi merupakan asal provinsi SLTA, atribut IPK, TOEFL dan lama studi merupakan perolehan sarjana. Atribut lama studi hari merupakan target *class* berasal dari hitungan atribut lama studi, didalam atribut lama studi terdapat jumlah dalam satuan tahun. Jumlah satuan tahun dikonversikan kedalam satuan hari, dimana 1 tahun terdapat sebanyak 365 hari. Selebihnya bulan dan hari ditambahkan kedalam total hari yang telah dikonversi. Contohnya dalam atribut lama studi seorang mahasiswa lulus masa studi selama 3 tahun 11 bulan 30 hari. 3 tahun dikonversi dalam satuan hari menjadi 365 hari x 3 = 1.095 hari, 11 bulan = 330 hari maka 1.095 hari ditambah 330 hari ditambah 30 hari sama dengan 1.455 hari.

Perancangan sistem prediksi merupakan acuan dalam penelitian supaya sebagai mana mestinya dan penelitian mendapat hasil yang maksimal. Perancangan sistem prediksi dapat terlihat dibawah ini pada Gambar 1. Perancangan Sistem Prediksi Kelulusan Mahasiswa.



Gambar 1. Perancangan Sistem Prediksi Kelulusan Mahasiswa

Penjelasan secara lengkap Gambar 1. Perancangan Sistem Prediksi Kelulusan Mahasiswa diatas sebagai berikut:

1. *Load data* yaitu suatu proses memasukkan dan mengolah data ke dalam program berupa file dokumen Excel untuk pemrosesan.

2. Pembersihan data *(data cleaning)* adalah tahapan pertama dimana data *noise* (nilai ganjil), data yang tidak digunakan dalam proses penambangan data, data yang nilainya tidak lengkap dibersihkan dari data mahasiswa yang dimuat. Contohnya pada satu nama mahasiswa tidak diketahui nilai TOEFL, maka nama tersebut akan dihapus karena dianggap data *noise.*
3. Integrasi *database* adalah proses menggabungkan data dari basis data yang berbeda menjadi satu kumpulan basis data. Diketahui data riwayat sebelum kuliah dan data setelah kuliah. Data mahasiswa FTI akan digabungkan menjadi database baru gabungan dari data sebelum kuliah dan sesudah kuliah berdasarkan nama mahasiswa.
4. Seleksi database merupakan fase di mana menentukan parameter data mana yang akan dipakai dalam tahapan penambangan, dikarenakan data yang valid akan diambil dari database. Tidak semua atribut diambil dalam proses penambangan, atribut yang akan diproses dalam metode yaitu atribut nama SLTA, Prodi (Program Studi), asal Provinsi, MTK (rata-rata matematika), IPK dan lama studi hari. Selain atribut tersebut akan dihapus.
5. Transformasi data (*data transformation*) merupakan fase di mana data dikonversi. Dalam survei ini, data awal dalam format numerik dikonversi ke kategori dan sebaliknya(Allan, 2020). Seperti atribut asal sekolah dan provinsi awal data format string diubah kedalam numerik dalam kategori. Pada atribut lama studi data awal format numerik dikonversi dalam format string sesuai kategori.
6. Data pelatihan (*data training*) adalah data yang digunakan untuk membuat algoritma prediksi yang funsinya sebagai acuan sistem mengambil keputusan. Data pelatihan yang digunakan adalah dari mahasiswa FTI lulusan tahun 2014-2015 sebanyak 273 data.
7. *Data testing* adalah sekumpulan data dengan atribut-atribut yang akan diuji untuk menampilkan hasil prediksi. Data testing diolah menggunakan penerapan algoritma sampai mengetahui performa seberapa akurat dari penelitian ini. Penelitian menggunakan 18 data yang diambil dari database untuk data testing.
8. Penerapan algoritma merupakan proses penerapan pada *system* penelitian menggunakan algoritma *Naive Bayes Classifier*. Metode klasifikasi ini dipakai dalam penelitian untuk menentukan mahasiswa lulus tepat waktu atau tidak tepat waktu dengan atribut yang telah ditentukan dalam penelitian.
9. Pengujian akurasi, yaitu pengecekan akurasi dengan menggunakan tabel *Confusion Matrix*. *Confusion matrix* adalah metode yang digunakan untuk mengukur kinerja algoritma klasifikasi sebagai larik bernilai empat yang mewakili hasil klasifikasi (Yusuf dkk., 2022). Keempat nilai tersebut merupakan nilai *True Positive* (TP), yaitu mahasiswa diharapkan lulus tepat waktu (positif) dan benar-benar (benar) lulus tepat waktu, nilai *True Negative* (TN) adalah mahasiswa tidak boleh lulus tepat waktu (negatif) dan benar. (Benar) tidak lulus tepat waktu, nilai *False Positive* (FP) adalah mahasiswa yang seharusnya lulus tepat waktu (positif) tetapi sebenarnya tidak lulus tepat waktu (*false*) dan *False Negative* (nilai FN) adalah mahasiswa yang seharusnya tidak lulus lulus tepat waktu (Negatif) tetapi sebenarnya lulus tepat waktu. Rumus lengkap dari rumus *confusion matrix* diatas terlihat seperti Tabel 3.

Tabel 2. Rumus Confusion Matrix

	Actual		
Predicted	*Class*	*Positive*	*Negative*
	Positive	*True Positive (TP)*	*False Negative (FN)*
	Negative	*False Positive (FP)*	*True Negative (TN)*

Pada Tabel 3. Rumus *Confusion Matrix* maka, dapat diketahui nilai dari akurasi, presisi dan *recall*. Akurasi adalah nilai yang mewakili akurasi sistem klasifikasi data di mana jumlah data yang diklasifikasikan dengan benar dibagi dengan jumlah total data. Untuk mencari nilai akurasi dapat menggunakan rumus di bawah ini.

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%$$

Presisi adalah nilai yang mewakili data bernilai positif yang diklasifikasikan dengan benar dibagi dengan jumlah data yang diklasifikasikan secara positif. Rumus mencari nilai presisi dapat dituliskan sebagai berikut.

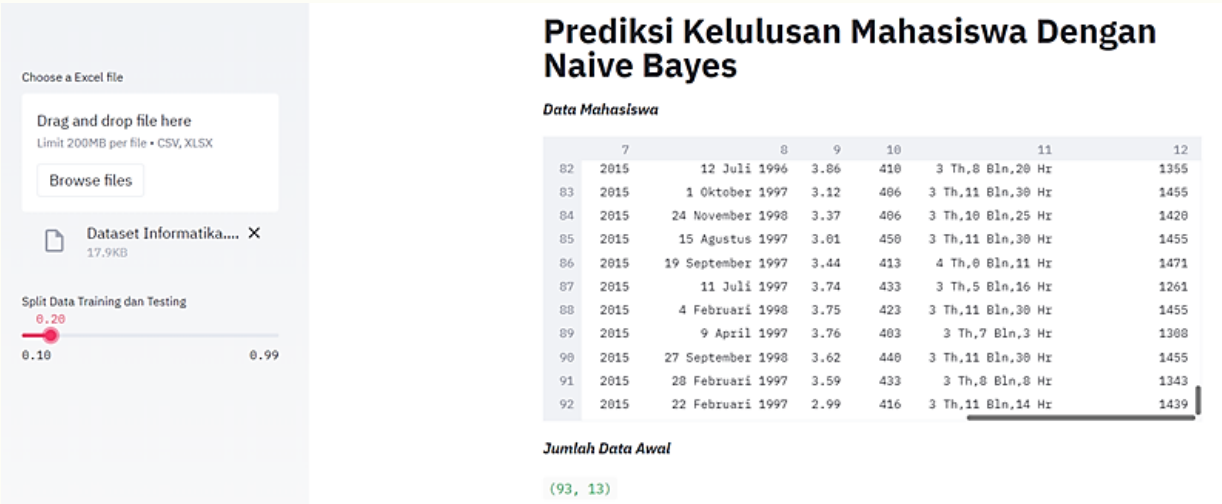
$$\text{Presisi} = \frac{TP}{TP+FP} \times 100\%$$

Recall adalah nilai yang mewakili data nilai positif yang diklasifikasikan dengan benar oleh sistem. Rumus mendapatkan nilai *recall* dapat dituliskan sebagai berikut.

$$\textit{Recall} = \frac{TP}{TP+FN} \times 100\%$$

HASIL DAN PEMBAHASAN

Penelitian ini, dilakukan menggunakan sistem prediksi kelulusan dengan bahasa pemrograman *python framework streamlit* menggunakan metode *naïve bayes classification.* Tahapan pertama dalam prediksi kelulusan sistem melakukan upload data mahasiswa dalam format excel yang berada pada penyimpanan lokal komputer. Data telah melakukan proses pembersihan secara manual sebelum akhirnya data akan diproses oleh sistem prediksi. Data penelitian mahasiswa FTI tahun 2014 dan 2015 telah dinyatakan lulus oleh fakultas. Data dalam penelitian fokus dengan atribut NIM, nama, asal sekolah, asal provinsi sekolah, prodi, rata-rata matematika, nilai IPK, TOEFL, lama studi, dan lama hari studi. Pada sistem prediksi kelulusan diuji salah satu prodi informatika FTI UAD jumlah *dataset* 93 data mahasiswa. Pada sistem prediksi kelulusan *fase* pertama yaitu *upload dataset.* Secara automatis *dataset* akan ditampilkan oleh system seperti menampilkan jumlah data awal terlihat pada Gambar 2.



Gambar 2. *Load Dataset* Penelitian

Fase kedua merupakan tahapan pembersihan *dataset* pada sistem prediksi, data siswa yang telah diunggah ke sistem pencarian terlebih dahulu akan dibersihkan dari data yang berisik (bernilai ganjil) dan data yang tidak terpakai dalam proses penemuan (Devita et al., 2018). Terlihat pada Gambar 3. pembersihan data jumlah data awal yang dimasukkan kedalam sistem penelitian dengan total data 93 mahasiswa, setelah pembersihan data oleh sistem menjadi 92 data mahasiswa. Terhapus 1 data mahasiswa karena salah satu atribut kosong atau tidak *valid*.

Fase berikutnya data yang telah melalui proses pembersihan akan di seleksi lagi untuk menentukan *variabel* data yang relevan diambil dari database untuk digunakan dalam proses ekstraksi. Proses seleksi atribut dari mulai atribut awal NIM, nama, asal sekolah, asal provinsi sekolah, prodi, rata-rata matematika, nilai IPK, TOEFL, lama studi, dan lama hari studi. Atribut yang digunakan penelitian hanya beberapa yaitu asal sekolah, provinsi, rata-rata matematika, IPK dan TOEFL. Seperti penelitian (Kharis et al., 2023) menggunakan atribut matematika dalam prediksi kelulusan, peneliti berharap dengan menggunakan atribut ini memperkuat hasil prediksi kelulusan sarjana. Pada atribut

status kelulusan didapat dari total hari kelulusan, data yang awal berbentuk angka (*numeric*) diubah ke dalam bentuk kategori (*categorical*). Apabila total 1.278 hari - 1.460 hari atau 7-8 semester maka akan memiliki status kelulusan tepat waktu (Setiadi, 2020). Sedangkan total hari kelulusan diatas 1.461 atau lebih dari 8 semester maka akan mendapat status kelulusan tidak tepat waktu. Selanjutnya sistem prediksi setelah melakukan seleksi data akan melakukan proses transformasi terlihat pada Gambar 3.

	ASAL SEKOLAH	PROVINSI	KUANT. MATE	KUANT. IPK	R. TOEFL	STATUS KELULUSAN
63	2	3	3	3	1	TEPAT
64	1	2	2	4	3	TEPAT
66	1	2	2	3	2	TEPAT
67	1	1	2	3	1	TEPAT
68	2	3	2	4	1	TEPAT
69	1	2	2	3	1	TEPAT
70	1	2	2	3	3	TIDAK TEPAT
71	1	2	1	4	3	TEPAT
72	1	2	3	3	1	TIDAK TEPAT
73	1	2	2	4	2	TEPAT
74	1	2	2	3	3	TEPAT
75	2	2	2	4	2	TEPAT
76	1	2	2	3	1	TIDAK TEPAT
77	1	2	2	3	2	TEPAT
78	1	2	3	4	4	TIDAK TEPAT
79	1	2	2	3	3	TEPAT
80	1	2	2	3	2	TEPAT
81	1	2	3	3	2	TEPAT
82	1	2	4	4	2	TEPAT
84	1	2	4	3	2	TEPAT
85	1	2	2	3	4	TEPAT
86	2	2	3	3	2	TIDAK TEPAT
87	1	2	3	4	3	TEPAT
88	3	2	2	4	3	TEPAT
89	1	2	4	4	2	TEPAT
90	1	2	2	4	3	TEPAT
91	1	2	2	4	3	TEPAT
92	1	2	2	2	2	TEPAT

Gambar 3. Hasil Seleksi dan Transformasi Data Pada Sistem Prediksi

Atribut data mahasiswa proses tranformasi data pada sistem prediksi kelulusan dijabarkan dibawah ini:

1. Asal Sekolah berisi atribut sekolah mahasiswa dengan kategori SMA (Sekolah Menengah Atas), SMK (Sekolah Menengah Kejuruan) dan MA (Madrasah Aliyah). .
2. Asal Provinsi berisi asal provinsi mahasiswa, apakah berasal dari Wilayah 1, Wilayah 2 atau Wilayah 3. Wilayah – wilayah tersebut dibagi berdasarkan perhitungan kualitas pendidikan di wilayah Indonesia(Nugraha dkk., 2018) seperti pada Tabel 4.

Tabel 3. Kategorial Wilayah Asal Provinsi

Kategori	Keterangan
Wilayah 1	Maluku Utara, Kalimantan Tengah.
Wilayah 2	Kalimanatan Barat, Kep. Bangka, Aceh, NTB, Sulawesi Barat, Sulawesu Selatan, Bali, Jawa Tengah, Jawa Timur, Sulawesi Utara, Riau, Sumatera Barat, Lampung, NTT, Sulawesi Tengah, Banten, DI Yogyakarta, Kalimantan Selatan, Bengkulu, Sumatera Selatan, Gorontalo, Sulawesi Tengah, Jakarta, Jambi, Jawa Barat, dan Sumatera Utara.
Wilayah 3	Papua Barat, Maluku, Kalimantan Timur dan Papua.

3. Rata-rata MTK adalah atribut rata-rata nilai matematika seorang siswa pada saat mendaftar ke Perguruan Tinggi jalur PMDK-Raport. Jenis MTK rata-rata dapat dilihat pada Tabel 5 .

Tabel 4 . Kategorial Rata-rata Matematika

Kategori	Keterangan
Sangat Baik (SB)	93 – 100
Baik (B)	84 – 92
Cukup (C)	75 – 83
Perlu Dimaksimalkan (K)	0 – 74

4. IPK berisi nilai IPK mahasiswa yang telah dinyatakan lulus. Rentang IPK mahasiswa bervariasi mulai dari 2,78 – 3,92. Predikat kelulusan mengacu peraturan Rektor UAD No. 3 Tahun 2016 tentang peraturan akademik terlihat pada Tabel 6.

Tabel 5. Kategorial Nilai IPK

Kategori	Keterangan
Dengan Pujian (Cumlaude/SB)	3,51 – 4,00
Sangat Memuaskan (B)	3,01 – 3,50
Memuaskan (C)	2,76 – 3,00
Cukup (K)	0,00 – 2,75

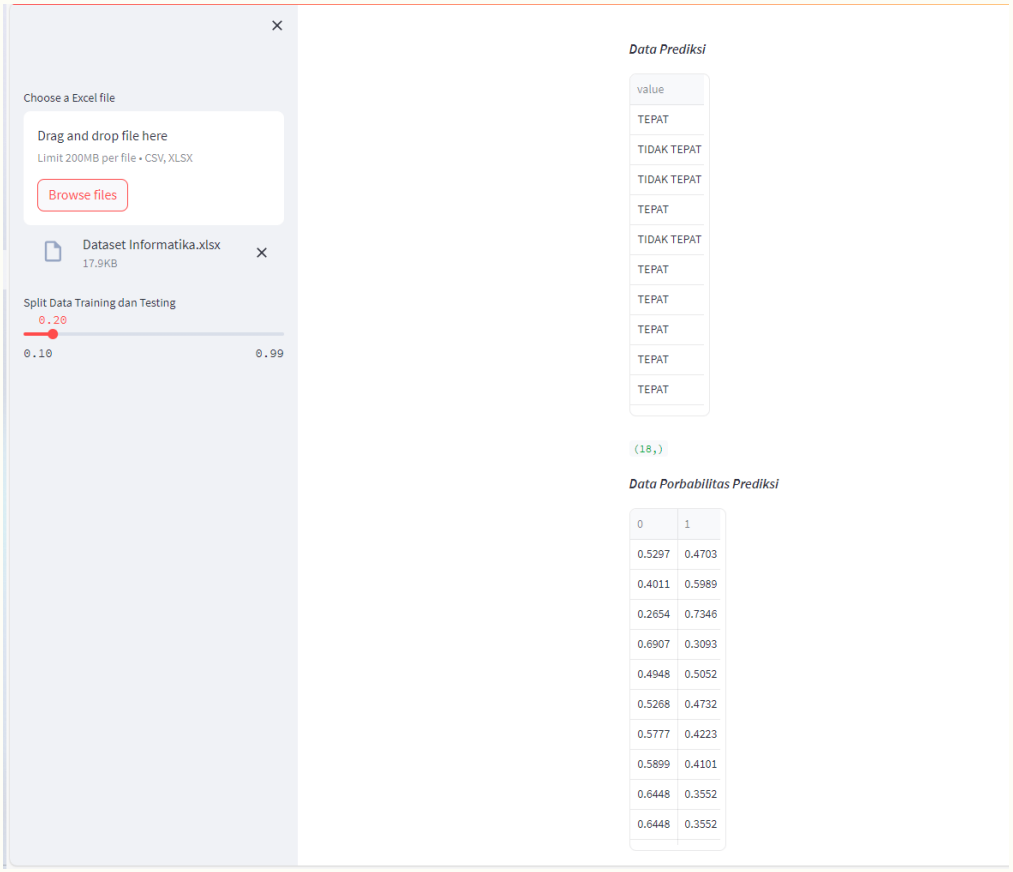
5. TOEFL mahasiswa minimal mendapatkan skor TOEFL sebesar 400. Tidak ada predikat tinggi atau rendah dalam skor TOEFL, maka skor TOEFL terbagi 11 *range* terlihat pada Tabel 7.

Tabel 6. Kategorial Range Nilai TOEFL

Kategori	Nilai TOEFL
Range 11	581-600
Range 10	561-580
Range 9	541-560
Range 8	521-540
Range 7	501-520
Range 6	481-500
Range 5	461-480
Range 4	441-460
Range 3	421-440
Range 2	401-420
Range 1	400

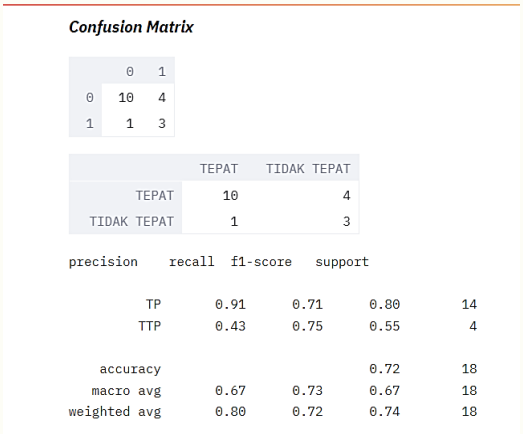
6. Lama masa studi menentukan mahasiswa lulus tepat waktu atau tidak. Mahasiswa dikatakan lulus tepat waktu jika masa studinya tidak lebih dari 4 tahun, sebaliknya jika lebih dari 4 tahun dianggap tidak tepat waktu.

Fase ini, setelah melakukan seleksi dan transformasi pada *dataset*. *Dataset* dibagi menjadi data pelatihan dan pengujian untuk memprediksi apakah siswa akan lulus tepat waktu atau gagal lulus tepat waktu. Proses pemisahan data *training* dan *testing* dilakukan oleh sistem prediksi. Setelah didapat data *testing* data akan dihitung menggunakan rumus (1) yaitu rumus *naïve bayes classifier* dapat dilihat pada Gambar 4. Pada tabel probabilitas prediksi apabila nilai probabilitas 0 lebih besar maka sistem memprediksi tepat waktu sebaliknya apabila nilai 1 lebih besar maka sistem memprediksi tidak lulus tepat waktu.



Gambar 4. Perhitungan *Naïve Bayes Classifier*

Selanjutnya pengujian akurasi pada sistem prediksi menggunakan pengujian *Confusion Matrix* pada sistem prediksi kelulusan menghasilkan akurasi 72%, nilai *recall* 71% dan nilai presisi 91%. Pengujian sistem dengan *Confusion Matrix* terlihat pada Gambar 5.



Gambar 5. Pengujian Confusion Matrix Pada Sistem Prediksi

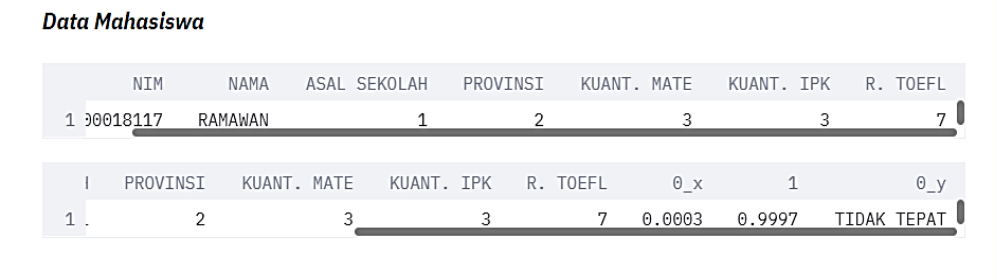
Hasil dari pengujian *Confusion Matrix* perlu dilakukan pengecekan lebih lanjut dengan perhitungan secara manual untuk memastikan hasil sistem prediksi sama. Dataset penelitian menggunakan data 72 mahasiswa sebagai data *training* dan 18 sebagai data *testing* dengan rincian *true positive*: 10, *true negative*: 3, *false positive*: 1, *false negative*: 4.

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% = \frac{10+3}{10+3+1+4} \times 100\% = 0.72 \times 100\% = 72\%$$

$$\textit{Precision} = \frac{TP}{TP+FP} \times 100\% = \frac{10}{10+1} \times 100\% = 0.90 \times 100\% = 90\%$$

$$\textit{Recall} = \frac{TP}{TP+FN} \times 100\% = \frac{10}{10+4} \times 100\% = 0.71 \times 100\% = 71\%$$

Tahap terakhir adalah penerapan algoritma *Naïve Bayes* setelah memisahkan data *training* dan data *testing*. Hasil prediksi dari aplikasi *Naïve Bayes* ditunjukkan pada Gambar 6.



Data Mahasiswa

	NIM	NAMA	ASAL SEKOLAH	PROVINSI	KUANT. MATE	KUANT. IPK	R. TOEFL
1	00018117	RAMAWAN	1	2	3	3	7

	PROVINSI	KUANT. MATE	KUANT. IPK	R. TOEFL	0_x	1	0_y
1	2	3	3	7	0.0003	0.9997	TIDAK TEPAT

Gambar 6. Hasil Sistem Prediksi Metode Naïve Bayes

Data prediksi mahasiswa berasal dari Nusa Tenggara Barat, Asal SMA, nilai matematika 80, IPK 3.47, dan nilai TOEFL 502 menggunakan algoritma *Naïve Bayes* diprediksi Lulus Tidak Tepat Waktu. Hasil kemungkinan lulus tepat waktu sebesar 0,0003 dan lulus tidak tepat waktu sebesar 0,9997, dari hasil kemungkinan nilai lulus tidak tepat waktu lebih besar daripada lulus tepat waktu. Kesimpulan hasil sistem prediksi mahasiswa Ramawan diprediksi lulus tidak tepat waktu.

KESIMPULAN

Penelitian yang dilakukan saat ini menyimpulkan bahwa algoritma *Naïve Bayes* dapat digunakan klasifikasi mahasiswa lulus tepat waktu atau tidak tepat waktu. Proses klasifikasi tersebut menjadi dasar prediksi kelulusan mahasiswa dalam penelitian. Proses klasifikasi pada penelitian dilakukan secara automatis oleh sistem. Proses klasifikasi semakin akurat apabila jumlah data yang menjadi lebih banyak data pelatihan daripada data uji. Penelitian dari algoritma *Naïve Bayes* didalam *system* menggunakan total data 291 terdiri dari 273 data *training* dan 18 data *testing* dengan atribut asal sekolah, provinsi, rata-rata matematika, IPK, TOEFL dan lama studi. Sistem prediksi kelulusan mahasiswa berhasil mendapat akurasi sebesar 72%, presisi sebesar 90% dan *recall* sebesar 71%. Dengan nilai kesalahan klasifikasi sebesar 28%. Kesalahan klasifikasi terjadi karena atribut yang digunakan kurang bervariasi, bervariasi atribut bisa menambah nilai akurasi pada sistem prediksi. Data mahasiswa prodi Tehnik Elektro menggunakan 33 total data, dengan detail *data training* 23 data dan *data testing* 10 data akurasi pada sistem menghasilkan 80%. Kemudian diuji dari data mahasiswa prodi Tehnik Industri menggunakan 75 total data, dengan detail *data training* 58 data dan *data testing* 17 data akurasi pada sistem menghasilkan 82%. Sedangkan data uji mahasiswa prodi Tehnik Kimia menggunakan 90 total data, dengan detail data latih sebanyak 72 data dan data uji sebanyak 18 data, akurasi sistem yang dihasilkan adalah 89%. Dengan demikian dapat disimpulkan bahwa akurasi sistem prediksi kelulusan dengan metode *Naive Bayes* memberikan nilai akurasi rata-rata sebesar 80%. Nilai akurasi rata-rata menunjukkan bahwa memprediksi kelulusan mahasiswa dengan metode *Naïve Bayes* dianjurkan. Namun uji coba menggunakan metode yang lain lebih tepat apabila akurasi mencapai diatas 90%. Penelitian prediksi kelulusan selanjutnya diharapkan dapat membandingkan dan mengembangkan

metode yang lainnya, penelitian menggunakan dataset sebanyak-banyaknya dan menggunakan variabel penelitian yang lebih variatif.

DAFTAR PUSTAKA

Allan, S. (2020). Migration and transformation: A sociomaterial analysis of practitioners' experiences with online exams. *Research in Learning Technology*, *28(2279),* 1-14. https://doi.org/10.25304/rlt.v28.2279

Amri, A. (2020). Dampak Covid-19 Terhadap UMKM Di Indonesia. *Jurnal Brand*, *2*(1), 123-130.

Devita, R. N., Herwanto, H. W., & Wibawa, A. P. (2018). Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa indonesia. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *5*(4), 427–434. https://doi.org/10.25126/jtiik.201854773

Dwi, R., Pambudi, Afif, A., Supianto, & Setiawan, N. Y. (2019). Prediksi Kelulusan Mahasiswa Berdasarkan Kinerja Akademik Menggunakan Pendekatan Data Mining Pada Program Studi Sistem Informasi Fakultas Ilmu Komputer Universitas Brawijaya. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *3*(3), 2194-2200.

Falade, A., Azeta, A., Oni, A., & Odun-ayo, I. (2019). Systematic Literature Review of Crime Prediction and Data Mining. *Review of Computer Engineering Studies*, *6*(3), 56-63. https://doi.org/10.18280/rces.060302

Hidayat, I. (2021). *Prediksi Kelulusan Tepat Waktu Menggunakan Algoritma Svm Dan K-Nearest Neighbour Berbasis Particle SWARM Optimization Di STMIK Eresha*. *XVI*(01), 55-65.

Kharis, S. A. A., Zili, A. H. A., Zubir, E., & Fajar, F. I. (2023). Prediksi Kelulusan Siswa pada Mata Pelajaran Matematika menggunakan Educational Data Mining. *Jurnal Riset Pembelajaran Matematika Sekolah*, *7*(1), 28-36. https://doi.org/10.21009/jrpms.071.03

Larasati, I. D., Supianto, A. A., & Furqon, M. T. (2019). *Prediksi Kelulusan Mahasiswa Berdasarkan Kinerja Akademik Menggunakan Metode Modified K-Nearest Neighbor (MK-NN)*. *3*(5), 4558-4593.

Manullang, J., & Panggabean, J. F. R. (2022). Analisis Kelulusan Mahasiswa Menggunakan Algoritma Naïve Bayes. *Jurnal Sains dan Teknologi ISTP*, *16*(2), 174-179. https://doi.org/10.59637/jsti.v16i2.123

Nugraha, G. S., Hairani, H., & Ardi, R. F. P. (2018). Aplikasi Pemetaan Kualitas Pendidikan di Indonesia Menggunakan Metode K-Means. *Jurnal MATRIK*, *17*(2), 13-23.

Nurdin, N., Suhendri, M., Afrilia, Y., & Rizal, R. (2021). Klasifikasi Karya Ilmiah (Tugas Akhir) Mahasiswa Menggunakan Metode Naive Bayes Classifier (NBC). *Sistemasi*, *10*(2), 268-279. https://doi.org/10.32520/stmsi.v10i2.1193

Pradnyana, G. A., & Permana, A. A. J. (2018). Sistem Pembagian Kelas Kuliah Mahasiswa Dengan Metode K-Means Dan K-Nearest Neighbors Untuk Meningkatkan Kualitas Pembelajaran. *JUTI: Jurnal Ilmiah Teknologi Informasi*, *16*(1), 59. https://doi.org/10.12962/j24068535.v16i1.a696

Priyatman, H., Sajid, F., & Haldivany, D. (2019). Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan Mahasiswa. *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, *5*(1), 62-66. https://doi.org/10.26418/jp.v5i1.29611

Puspita, R., & Widodo, A. (2021). *Perbandingan Metode KNN , Decision Tree , dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS*. *5*(4), 646-654.

Putro, H. F., Vulandari, R. T., & Saptomo, W. L. Y. (2020). Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan. *Jurnal Teknologi Informasi dan Komunikasi (TIKomSiN)*, *8*(2), 19-24. https://doi.org/10.30646/tikomsin.v8i2.500

Rachmat, R., Chrisnanto, Y. H., & Umbara, F. R. (2020). *Sistem Prediksi Mutu Air Di Perusahaan Daerah Air Minum Tirta Raharja Menggunakan K – Nearest Neighbors (K – NN)*. *Prosiding SISFOTEK*, *4*(1), 189-193.

Ramadhayanti, A. (2018). Analisis Strategi Belajar Dengan Metode Bimbel Online Terhadap Kemampuan Pemahaman Kosa Kata Bahasa Inggris dan Pronunciation (Pengucapan/pelafalan) Berbahasa Remaja Saat Ini. *KREDO : Jurnal Ilmiah Bahasa dan Sastra*, *2*(1), 39-52. https://doi.org/10.24176/kredo.v2i1.2580

Rustam, S., & Annur, H. (2019). *Akademik Data Mining (Adm) K-Means Dan K-Means K-Nn Untuk Mengelompokan Kelas Mata Kuliah Kosentrasi*. *11*(3), 260-268. https://doi.org/10.33096/ilkom.v11i3.487.260-268

Setiadi, N. T. (2020). *Prediksi Kelulusan Mahasiswa Berdasarkan Data Berkunjung dan Pinjam Buku di Perpustakaan Menggunakan Metode C4.5*. *8*(2), 24-33. http://dx.doi.org/10.12928/jstie.v8i2.16950

Shedriko, S., & Firdaus, M. (2022). Penentuan Klasifikasi Dengan Crisp-Dm Dalam Memprediksi Kelulusan Mahasiswa Pada Suatu Mata Kuliah. *Semnas Ristek (Seminar Nasional Riset dan Inovasi Teknologi)*, *6*(1), 826-831. https://doi.org/10.30998/semnasristek.v6i1.5814

Saputra, R., & Sibarani, A. J. P. (2020). Implementasi Data Mining Menggunakan Algoritma Apriori Untuk Meningkatkan Pola Penjualan Obat. *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, *7*(2), 262-276. https://doi.org/10.35957/jatisi.v7i2.195

Waliyansyah, R. R., & Fitriyah, C. (2019). Perbandingan Akurasi Klasifikasi Citra Kayu Jati Menggunakan Metode Naive Bayes dan k-Nearest Neighbor (k-NN). *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, *5*(2), 157-163.

Yalidhan, M. D., & Amin, M. F. (2018). Implementasi Algoritma Backpropagation Untuk Memprediksi Kelulusan Mahasiswa. *Klik - Kumpulan Jurnal Ilmu Komputer*, *5*(2), 169. http://dx.doi.org/10.20527/klik.v5i2.152

Yusuf, W., Witri, R., & Juliane, C. (2022). Model Prediksi Penjualan Jenis Produk Tekstil Menggunakan Algoritma K-Nearest Neighbor (K-NN). *(Indonesian Journal on Computer and Information Technology)*, *7*(1), 1–6. https://doi.org/10.31294/ijcit.v7i1.11973