

Studi Komparasi Algoritma *Decision Tree C4.5* dan *K-Nearest Neighbor* pada Klasifikasi Masa Studi dan Tingkat Stres Mahasiswa

Nurul Renaningtias^{1*}, Gushevinalti², Tiara Eka Putri^{3*}, Endina Putri Purwandari⁴, Yusak Stainly Ritonga⁵

^{1,3,4,5}Prodi Sistem Informasi, Universitas Bengkulu, Bengkulu, Indonesia

²Prodi Ilmu Komunikasi, Universitas Bengkulu, Bengkulu, Indonesia

^{*}Email Corresponding Author: nurulrenaningtias@unib.ac.id

Abstract

This research explores the utilization of Educational Data Mining (EDM) to analyze attributes that affect students' academic performance focusing on study duration and stress levels. In this study, the performance of two classification algorithms Decision Tree C4.5 and K-Nearest Neighbor (KNN) was compared in classifying students' study duration and stress levels based on alumni data from those who graduated between 202-2023. Several variables analyzed in this dataset include gender, GPA, the number of credits taken, admission pathway, participation in organizations, and activity as an assistant. The main findings of this study indicate that gender is a significant factor in predicting students' study duration, while GPA substantially impacts students' stress levels. Regarding algorithm performance, KNN outperformed Decision Tree C4.5, achieving an accuracy rate of 71.44% for study duration classification and 64.17% for stress level classification. This research provides valuable insights for higher education institutions in formulating policies to enhance students' academic performance and well-being.

Keywords: *Educational Data Mining; Decision Tree C4.5; K-Nearest Neighbor; Student Study Period; Student Stress Level*

Abstrak

Penelitian ini mengeksplorasi pemanfaatan *Educational Data Mining* (EDM) untuk menganalisis atribut-atribut yang memengaruhi kinerja akademik mahasiswa dengan fokus pada durasi studi dan tingkat stres. Dalam penelitian ini, kinerja dua algoritma klasifikasi yaitu *Decision Tree C4.5* dan *K-Nearest Neighbor* dibandingkan dalam mengklasifikasikan masa studi dan tingkat stres mahasiswa berdasarkan data alumni yang lulus antara tahun 2021-2023. Variabel yang dianalisis dalam dataset ini adalah jenis kelamin, IPK, jumlah SKS yang diambil, jalur masuk, keaktifan organisasi, dan keaktifan asistensi. Temuan utama dari penelitian ini menunjukkan bahwa jenis kelamin merupakan faktor signifikan dalam memprediksi durasi studi mahasiswa, sedangkan IPK memiliki dampak substansial terhadap tingkat stres mahasiswa. Berdasarkan hasil komparasi performa algoritma, KNN lebih unggul dibandingkan *Decision Tree C4.5*, dengan tingkat akurasi sebesar 71,44% untuk klasifikasi durasi studi dan 64,17% untuk klasifikasi tingkat stres. Penelitian ini memberikan wawasan berharga bagi institusi pendidikan tinggi dalam merumuskan kebijakan untuk meningkatkan kinerja akademik dan kesejahteraan mahasiswa.

Kata Kunci: *Educational Data Mining; Decision Tree C4.5; K-Nearest Neighbor; Masa studi mahasiswa; Tingkat stres mahasiswa*

1. Pendahuluan

Tekanan selama menjalani perkuliahan dirasakan oleh banyak mahasiswa, salah satunya di Universitas Bengkulu. Universitas Bengkulu merupakan Perguruan Tinggi di Indonesia yang pada tahun ajaran 2023/2024 memiliki jumlah mahasiswa sebanyak 17.486 orang. Untuk mencapai efisiensi pengelolaan maka pihak universitas dan fakultas harus melakukan pengelolaan secara profesional. Oleh karena itu, diperlukan metode untuk

menghasilkan informasi yang berguna secara ilmiah untuk pengambilan kebijakan atau keputusan [1]. Penggalan data pendidikan merupakan cara yang dapat dilakukan [2] untuk menganalisis sejumlah data yang besar di lingkungan pendidikan dengan tujuan membantu lembaga pendidikan dalam menentukan kebijakan [3][4]. Salah satunya dengan melakukan penggalan data mengenai masa studi dan tingkat stress mahasiswa yang sangat berpengaruh dalam penyelesaian studinya dan merupakan tantangan bagi mahasiswa pada kehidupan keserjanaan di Perguruan Tinggi [5][6].

Menurut kebijakan Badan Akreditasi Nasional Perguruan Tinggi (BAN PT) dalam peraturan No. 4 tahun 2017 tentang Kebijakan Penyusunan Instrumen Akreditasi, lama studi adalah salah satu ukuran dan elemen penilaian dalam akreditasi program studi. Masa studi mahasiswa dan tingkat stres memiliki dampak langsung pada kesejahteraan keberhasilan akademis. *Stres* akademik yang berlebihan dapat menyebabkan penurunan kemampuan akademik, konsentrasi yang buruk dan sulitnya dalam menyelesaikan masalah yang dapat berpengaruh terhadap Indeks Prestasi [7][8]. Oleh karena itu, penting untuk melakukan penelitian yang dapat membantu mengidentifikasi faktor-faktor yang berkontribusi terhadap lama masa studi dan tingkat stres mahasiswa.

Beberapa penelitian mengenai *educational data mining* (EDM) yang telah dilakukan di berbagai negara dalam menganalisis dan menyelesaikan isu-isu yang berkaitan dengan dunia pendidikan seperti prediksi masa studi dan predikat kelulusan [9][5][10], prediksi kinerja siswa dan analisis perilaku siswa [3], serta identifikasi prioritas pembelajaran untuk berbagai kelompok siswa [11]. Metode *data mining* khususnya algoritma *Decision Tree* C4.5 dan *K-Nearest Neighbor* menawarkan pendekatan yang efektif dalam menganalisis dan mengklasifikasikan pola-pola yang kompleks dari data mahasiswa [12].

Oleh karena itu, dalam penelitian ini akan dilakukan studi komparasi algoritma *Decision Tree* C4.5 dan K-NN untuk klasifikasi masa studi dan tingkat stres mahasiswa dengan variabel yang lebih kompleks seperti jenis kelamin, program studi [13], IPK, SKS, tingkat stres mahasiswa [5], jalur masuk [12], masa studi [14], dan keaktifan organisasi [15]. Dataset yang digunakan adalah data alumni Universitas Bengkulu yang lulus pada tahun 2021-2023. Evaluasi untuk mengukur performa model menggunakan *confusion matrix*. Hasil penelitian ini dapat memberikan panduan dalam pengambilan kebijakan di Universitas Bengkulu untuk meningkatkan kualitas pendidikan dan kesejahteraan mahasiswa dan dapat membantu pengelola dalam meningkatkan kinerja akademik.

2. Tinjauan Pustaka

Penelitian dalam penentuan tingkat stress mahasiswa akhir telah dilakukan oleh [6] dengan menerapkan data mining C4.5 menggunakan parameter penilaian seperti Interpersonal, Intrapersonal, akademik dan lingkungan. Penelitian ini menggunakan 100 sampel data mahasiswa dengan menggunakan teknik kuesioner. Penelitian mendapatkan 20 *rules* sebagai pedoman penentuan tingkat stress dan ditemukan bahwa faktor Interpersonal yang menjadi faktor penentu yang menyebabkan stress bagi mahasiswa tingkat akhir. Berdasarkan hasil penelitian ini mendapatkan performa C4.5 memberikan akurasi sebesar 87,88%

Selanjutnya penelitian yang telah dilakukan oleh [5] yang menganalisis performa dari tiga metode yaitu *K-NN*, *Naïve Bayes* dan *Neural Network* dalam membuat model untuk memprediksi kelulusan mahasiswa dengan menggunakan variabel IPK, SKS, status pekerjaan dan tingkat stress mahasiswa dalam menghadapi skripsi. Penelitian ini menggunakan data mahasiswa angkatan 2017-2018. Berdasarkan dari penelitian yang dilakukan diketahui performa model yang paling baik adalah K-NN dengan akurasi sebesar 89% dibandingkan dengan Neural Network 88% dan Naïve bayes sebesar 84%. Pada penelitian ini diketahui bahwa banyak mahasiswa yang lulus tepat waktu berada pada stress level 2.

Selain itu, penelitian menerapkan algoritma *K-Nearest Neighbor* dalam memprediksi ketepatan waktu studi mahasiswa bidik misi jenjang sarjana yang menggunakan variabel jenis kelamin, nilai TOEFL, IP semester 1-4, nilai SMA, nilai UN SMA, program studi dan keaktifan organisasi. Penelitian ini menggunakan data historis mahasiswa bidikmisi jenjang sarjana dari tahun 2013-2017 sebanyak 1593 data. Pengujian dilakukan menggunakan kombinasi percentage split dan nilai k. Berdasarkan dari penelitian ini menemukan bahwa semakin besar nilai k maka nilai akurasi yang didapatkan semakin baik. Performa *K-Nearest Neighbor* mendapatkan hasil akurasi sebesar 93,93% pada nilai k=15 [13][16].

Penelitian mengenai komparasi metode seperti K-NN, naïve bayes, decision tree, ensemble dan linear regression pernah dilakukan untuk menganalisis kinerja siswa dan mengetahui performa algoritma terbaik. Variabel yang digunakan dalam penelitian ini adalah gender, ras, tingkat pendidikan orang tua, pemilihan makan siang, dan kursus persiapan ujian yang mana sampel dari penelitian ini adalah pelajar kelas 3 Sekolah Menengah Atas. Berdasarkan hasil perbandingan *test* dan *score*, algoritma yang memiliki performa tertinggi adalah K-Nearest Neighbor [17][18]. Performa K-NN dalam memprediksi kinerja siswa dan klasifikasi *breast cancer* menunjukkan hasil yang optimal dibanding dengan algoritma lainnya [19][20]. Pada penelitian yang berjudul klasifikasi dengan perbandingan metode decision tree, naïve bayes, dan k-nearest neighbor dalam klasifikasi penjualan dan klasifikasi *heart disease* menggunakan PCA juga menghasilkan decision tree memiliki performa dan akurasi terbaik [21][22].

Dilihat dari berbagai tinjauan literatur sebelumnya, belum ada penelitian yang secara langsung membandingkan kinerja *Decision Tree* C4.5 dengan *K-Nearest Neighbor* dengan menggunakan *Educational Data Mining* untuk klasifikasi masa studi dan tingkat stres mahasiswa. Setelah mengkaji dan menganalisis beberapa penelitian sebelumnya, maka pada penelitian ini dilakukan studi komparasi performa algoritma *Decision Tree* C4.5 dengan *K-Nearest Neighbor* untuk klasifikasi masa studi dan tingkat stress mahasiswa.

3. Metodologi

Tahapan yang digunakan pada penelitian ini berbasis *Knowledge Discovery in Database* dengan tahapan pada penelitian ini adalah sebagai berikut.

- 1) *Selection data*. Variabel target pada penelitian ini adalah masa studi dan tingkat stress sedangkan variabel prediktor yaitu Jenis Kelamin, Program Studi, IPK, SKS, Jalur Masuk, Keaktifan Organisasi dan Keaktifan Asistensi.
- 2) *Preprocessing data*. Data yang diambil adalah data alumni Universitas Bengkulu yang sudah lulus antara tahun 2021- 2023 menggunakan teknik kuesioner. Pembersihan data yang dilakukan apabila terdapat data yang hilang, data ganda atau bersifat *outlier*.
- 3) *Transformation data*. Berikut tabel adalah kategori untuk variabel prediktor dan variabel target.

Tabel 1. Transformasi data

Variabel Target	Nilai	Kategori
Tingkat Stres	$X < 48$	Rendah
	$48 \leq X < 72$	Sedang
	$X > 72$	Tinggi
Masa Studi	≤ 4 tahun	Lulus Tepat Waktu
	> 4 tahun	Lulus Terlambat
Jenis Kelamin	Perempuan	Perempuan
	Laki-laki	Laki-laki
Program Studi	Teknik Informatika	Teknik Informatika
	Teknik Sipil	Teknik Sipil
	Teknik Mesin	Teknik Mesin
	Teknik Elektro	Teknik Elektro
	Arsitektur	Arsitektur
	Sistem Informasi	Sistem Informasi
IPK	2,00 – 2,75	Memuaskan
	2,76 – 3,50	Sangat Memuaskan
	3,51 – 4,00	<i>Cumlaude</i>
SKS	144 – 150	Cukup
	151 – 165	Lebih
Jalur Masuk	SNMPTN	SNMPTN
	SBMPTN	SBMPTN
	Seleksi Mandiri	Seleksi Mandiri
Jumlah Organisasi	0	Rendah
	1 – 2	Sedang
	> 2	Tinggi
Jumlah Asistensi	0	Rendah

Variabel Target	Nilai	Kategori
Dosen	1 – 2	Sedang
	> 2	Tinggi

Setiap variabel akan dimasukkan kedalam kategori sesuai nilainya, untuk variabel tingkat stres akan dilakukan perhitungan dahulu sebelum dimasukkan kedalam kategori sesuai dengan perhitungan rumus kategorisasi hipotetik dan empirik. Berdasarkan hasil perhitungan akan mendapatkan nilai untuk kategori dari variabel tingkat stress.

- 4) *Data Mining*. Pada tahap ini, diimplementasikan algoritma *Decision Tree C4.5* dan *K-Nearest Neighbor* untuk menemukan model dalam klasifikasi masa studi dan tingkat stres mahasiswa.
- 5) *Evaluation*. Evaluasi model dilakukan dengan menggunakan *Confusion Matrix* dan kurva ROC (*Receiver Operating Characteristic*). Nilai performansi yang digunakan yaitu *accuracy*, *precision*, dan *recall*.

4. Hasil dan Pembahasan

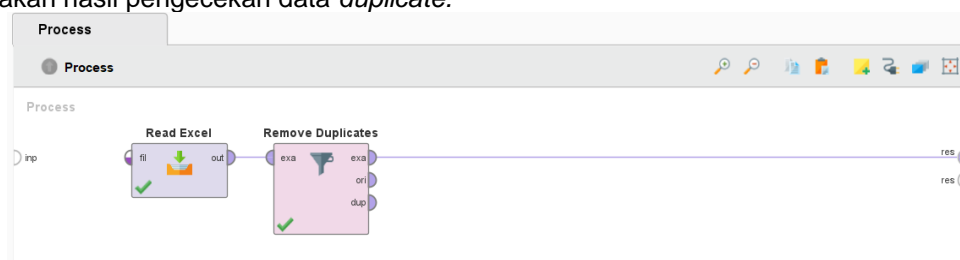
4.1 Hasil Penelitian

Penelitian ini melakukan klasifikasi masa studi dan tingkat stres mahasiswa dengan teknik *data mining* menggunakan algoritma *decision tree C4.5* dan *K-Nearest Neighbor* dan dibantu dengan aplikasi *RapidMiner* [23]. Tahapan penelitian untuk melakukan penambangan data pendidikan terkait dengan masa studi dan tingkat stress mahasiswa menerapkan *Knowledge Discovery in Database* dengan tahapan *selection data*, *preprocessing data*, *transformation data*, *data mining* dan *evaluation*.

Data yang diambil adalah data alumni Universitas Bengkulu yang sudah lulus dari tahun 2021-2023. Variabel target pada penelitian ini adalah masa studi dan tingkat stres sedangkan variabel-variabel prediktornya yaitu jenis kelamin, program studi, IPK, SKS, jalur masuk, jumlah organisasi dan jumlah asistensi dosen. Selanjutnya dilakukan *preprocessing data*, tahap ini dilakukan agar apabila terdapat data yang mengandung *missing value*, data ganda atau bersifat *outlier* dapat ditangani. Pada tahapan ini dilakukan pemrosesan data dengan melakukan *cleaning* data yang terduplikat ataupun bersifat *outlier*, kemudian melakukan pengecekan data yang hilang dilakukan dengan proses *checking missing value* dengan bantuan dari tools *RapidMiner*. Berikut merupakan tahapan *preprocessing data*:

- 1) Pengecekan data duplicate

Duplicate merupakan proses untuk mengidentifikasi baris duplikat dalam *dataset*. Fungsi ini mengembalikan *boolean Series* di mana *True* mengindikasikan baris duplikat. Berikut merupakan hasil pengecekan data *duplicate*.

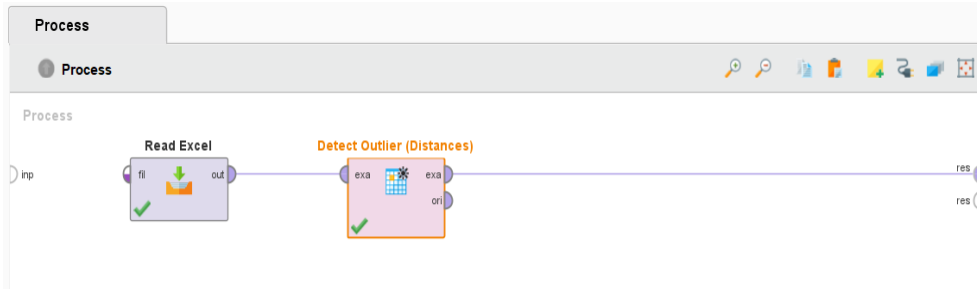


Gambar 1. Pengecekan data *duplicate*

Hasil *preprocessing duplicate* menunjukkan bahwa pada dataset tidak mengandung data yang bersifat duplikat. Hal ini ditunjukkan pada kolom tidak ada data yang bersifat duplikat.

- 2) Pengecekan data outlier

Outlier merupakan proses untuk menemukan dan menyingkirkan item yang tidak relevan dari kumpulan data ketika muncul nilai-nilai ekstrem baik secara univariat maupun multivariat. Nilai-nilai ekstrem ini secara signifikan berbeda atau jauh dari sebagian besar nilai lain dalam kumpulannya.

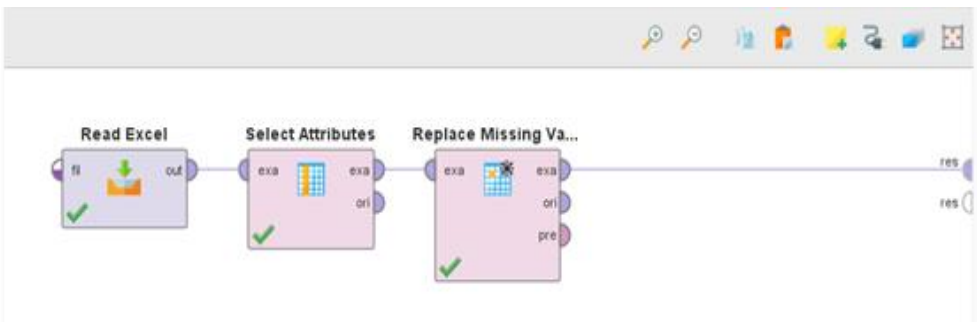


Gambar 2. Pengecekan data outlier

Hasil *preprocessing outlier* menunjukkan bahwa pada dataset tidak mengandung data yang bersifat outlier. Hal ini ditunjukkan pada kolom outlier yang menyatakan nilai *false* pada setiap baris data yang pada dataset.

3) Pengecekan data missing value

Missing value merupakan proses untuk mengidentifikasi baris data yang tidak memiliki nilai dalam *dataset*. Fungsi ini berguna untuk mengisi kembali data yang kosong dengan nilai rata-rata kolom yang mengandung *missing values*.



Gambar 3. Pengecekan data missing value

Hasil dari *preprocessing missing value* menunjukkan bahwa pada *dataset* tidak mengandung data yang memiliki nilai atau kosong. Hal ini ditunjukkan pada kolom *Missing* yang menyatakan nilai 0 pada setiap kolom data yang ada pada dataset.

Tahap yang dilakukan selanjutnya adalah transformasi data. Pada tahap ini dilakukan pengelompokan data sesuai dengan kategori yang telah ditentukan pada Tabel 1. Beberapa hasil dari transformasi data ditunjukkan pada Tabel 2.

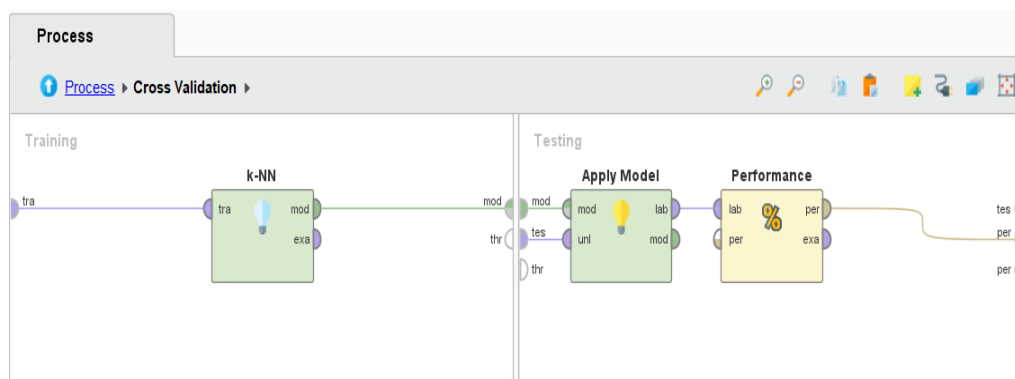
Tabel 2. Hasil transformasi data

Masa Studi	Program Studi	IPK	SKS	Jalur Masuk	Jumlah Organisasi	Jumlah MK Asistensi	Tingkat Stres
Lulus terlambat	Sistem Informasi	Cumlaude	Cukup	SNMPTN	Tinggi	Rendah	Sedang
Lulus tepat waktu	Sistem Informasi	Cumlaude	Cukup	SNMPTN	Sedang	Sedang	Sedang
Lulus terlambat	Sistem Informasi	Cumlaude	Cukup	SBMPTN	Tinggi	Tinggi	Rendah
Lulus terlambat	Teknik Mesin	Sangat Memuaskan	Cukup	SBMPTN	Sedang	Sedang	Rendah
Lulus terlambat	Teknik Mesin	Sangat Memuaskan	Cukup	SBMPTN	Sedang	Sedang	Sedang
Lulus terlambat	Teknik Elektro	Sangat Memuaskan	Cukup	SBMPTN	Sedang	Tinggi	Sedang
Lulus terlambat	Teknik Mesin	Sangat Memuaskan	Cukup	SBMPTN	Tinggi	Sedang	Sedang
Lulus terlambat	Sistem Informasi	Cumlaude	Cukup	Seleksi Mandiri	Sedang	Sedang	Sedang
Lulus terlambat	Arsitektur	Cumlaude	Cukup	SBMPTN	Sedang	Rendah	Rendah

Masa Studi	Program Studi	IPK	SKS	Jalur Masuk	Jumlah Organisasi	Jumlah MK Asistensi	Tingkat Stres
Lulus terlambat	Arsitektur	Cumlaude	Lebih	SBMPTN	Sedang	Sedang	Sedang
Lulus tepat waktu	Arsitektur	Cumlaude	Lebih	SBMPTN	Sedang	Sedang	Sedang
Lulus terlambat	Informatika	Sangat Memuaskan	Cukup	SBMPTN	Tinggi	Sedang	Sedang
Lulus terlambat	Teknik Sipil	Sangat Memuaskan	Cukup	Seleksi Mandiri	Sedang	Sedang	Sedang
Lulus terlambat	Arsitektur	Sangat Memuaskan	Cukup	SBMPTN	Sedang	Rendah	Rendah
Lulus terlambat	Teknik Elektro	Sangat Memuaskan	Cukup	SBMPTN	Sedang	Rendah	Sedang
Lulus terlambat	Informatika	Cumlaude	Cukup	SBMPTN	Tinggi	Sedang	Sedang
Lulus terlambat	Sipil	Sangat Memuaskan	Cukup	Seleksi Mandiri	Sedang	Sedang	Rendah

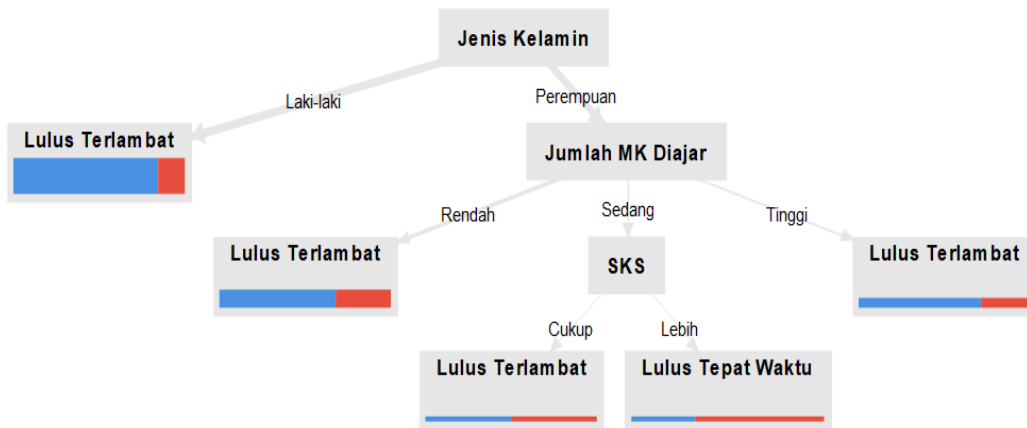
Hasil transformasi data akan diolah dan diproses dengan teknik data mining. Pada tahapan ini dilakukan pemilihan teknik yang tepat untuk melakukan klasifikasi tingkat stres mahasiswa. Analisis data dilakukan dengan algoritma *K-Nearest Neighbor* dan *Decision Tree C4.5*.

Algoritma *K-Nearest Neighbor* (K-NN) adalah teknik untuk mengklasifikasikan objek berdasarkan data pembelajaran yang jaraknya paling dekat dengannya. Algoritma *K-Nearest Neighbor* digunakan untuk menemukan kasus dengan memperkirakan seberapa dekat kasus baru dengan kasus lama. Data pengujian dan nilai jarak data pelatihan diurutkan dimulai dengan nilai terendah [5]. Berikut merupakan model dari klasifikasi dengan algoritma (k-NN).



Gambar 4. Model klasifikasi dengan algoritma k-NN

Algoritma kedua yang digunakan adalah *decision tree C4.5*. Algoritma ini membentuk sebuah pohon keputusan dengan perhitungan nilai *entropy* dan *gain* masing-masing variabel. Berdasarkan perhitungan variabel dengan *gain* tertinggi akan menjadi akar dari pohon keputusan yang terbentuk. Perhitungan nilai *entropy* dan *gain* setiap variabel membentuk pohon keputusan. Hasil perhitungan terlihat bahwa nilai *gain* untuk IPK merupakan hasil yang tertinggi untuk masa studi dan nilai *gain* untuk jenis kelamin merupakan hasil yang tertinggi untuk tingkat stres mahasiswa. Ini menunjukkan bahwa jenis kelamin merupakan faktor signifikan dalam memprediksi durasi studi mahasiswa, sedangkan IPK memiliki dampak substansial terhadap tingkat stres mahasiswa. Hasil *rules* yang terbentuk untuk klasifikasi masa studi dan tingkat stress mahasiswa ditunjukkan pada Gambar 5 dan 6.



Gambar 5. Pohon keputusan klasifikasi masa studi

Pohon keputusan diatas membentuk 5 *rules* untuk klasifikasi masa studi, sebagai berikut:

- 1) Jika Jenis Kelamin mahasiswa adalah laki-laki maka dapat diprediksi mahasiswa akan Lulus Terlambat
- 2) Jika Jenis Kelamin mahasiswa adalah perempuan dengan jumlah MK yang diajar rendah maka dapat diprediksi mahasiswa akan Lulus Terlambat
- 3) Jika Jenis Kelamin mahasiswa adalah perempuan dengan jumlah MK yang diajar sedang dan memiliki SKS yang cukup maka dapat diprediksi mahasiswa akan Lulus Terlambat
- 4) Jika Jenis Kelamin mahasiswa adalah perempuan dengan jumlah MK yang diajar sedang dan memiliki SKS yang lebih maka dapat diprediksi mahasiswa akan Lulus Tepat Waktu
- 5) Jika Jenis Kelamin mahasiswa adalah perempuan dengan jumlah MK yang diajar tinggi maka dapat diprediksi mahasiswa akan Lulus Terlambat



Gambar 6. Pohon keputusan klasifikasi tingkat stres

Pohon keputusan diatas membentuk 6 *rules* untuk klasifikasi tingkat stress mahasiswa, sebagai berikut:

- 1) Jika IPK mahasiswa Sangat Memuaskan maka dapat diprediksi mahasiswa akan mengalami Stres tingkat Sedang
- 2) Jika IPK mahasiswa adalah *Cumlaude* dengan Jalur Masuk SBMPTN dan jumlah SKS cukup maka dapat diprediksi mahasiswa akan mengalami Stres tingkat Rendah

- 3) Jika IPK mahasiswa adalah *Cumlaude* dengan Jalur Masuk SBMPTN dan jumlah SKS lebih maka dapat diprediksi mahasiswa akan mengalami Stres tingkat Sedang
- 4) Jika IPK mahasiswa adalah *Cumlaude* dengan Jalur Masuk SNMPTN dan jumlah SKS cukup maka dapat diprediksi mahasiswa akan mengalami Stres tingkat Sedang
- 5) Jika IPK mahasiswa adalah *Cumlaude* dengan Jalur Masuk SNMPTN dan jumlah SKS lebih maka dapat diprediksi mahasiswa akan mengalami Stres tingkat Rendah
- 6) Jika IPK mahasiswa adalah *Cumlaude* dengan Jalur Masuk Seleksi Mandiri maka dapat diprediksi mahasiswa akan mengalami Stres tingkat Sedang

Setelah dilakukan *mining* data selanjutnya dilakukan evaluasi. Tahap ini digunakan untuk mengevaluasi hasil-hasil klasifikasi dari masa studi dan tingkat stres mahasiswa yang dihasilkan oleh kedua algoritma dan menentukan algoritma terbaik. Nilai *accuracy*, *precision* dan *recall* digunakan untuk melakukan evaluasi dengan metode *confusion matrix*. Tahap evaluasi bertujuan untuk menentukan hasil kegunaan dari model yang telah dibuat pada langkah sebelumnya. Metode validasi yang digunakan peneliti adalah 10 *fold-cross validation*.

Model yang dihasilkan algoritma *Decision tree* C4.5 dan K-NN dengan performansi masa studi masuk ke dalam klasifikasi kelas Lulus Tepat Waktu dan Lulus Terlambat dan untuk penentuan tingkat stres mahasiswa masuk ke dalam klasifikasi kelas Rendah, Sedang, dan Tinggi. Evaluasi tingkat *accuracy* dari algoritma digunakan metode *confusion matrix*. Hasil komparasi dan performa algoritma *K-Nearest Neighbor* dan *Decision Tree* C4.5 ditunjukkan pada Tabel 3 berikut.

Tabel 3. Hasil komparasi metode *K-Nearest Neighbor* dan *Decision Tree* C4.5

Confusion Matrix	Masa Studi		Tingkat Stres	
	k-NN	Decision Tree C4.5	k-NN	Decision Tree C4.5
<i>Accuracy</i>	71.44 %	68.64 %	64.17 %	59.02 %
<i>Precision</i>	74.29 %	75.53 %	68.04 %	66.29 %
<i>Recall</i>	93.98 %	85.54 %	88.00 %	78.67 %

Berdasarkan hasil perbandingan terlihat bahwa algoritma *K-Nearest Neighbor* memiliki nilai yang paling baik untuk mengklasifikasikan masa studi dan tingkat stress mahasiswa dibandingkan dengan algoritma *Decision Tree* C4.5. Nilai *accuracy* *K-Nearest Neighbor* dalam mengklasifikasikan masa studi adalah sebesar 71.44% dan tingkat stres mahasiswa sebesar 64.17 % dan *Decision Tree* C4.5 memiliki nilai *accuracy* dalam mengklasifikasikan masa studi adalah sebesar 68.64% dan tingkat stres mahasiswa sebesar 59.02 %.

4.2 Pembahasan

Penelitian klasifikasi masa studi dan tingkat stres mahasiswa dengan teknik *data mining* menggunakan algoritma *decision tree* C4.5 dan *K-Nearest Neighbor* menghasilkan performa algoritma *K-Nearest Neighbor* lebih unggul dan memiliki performa terbaik. Hal ini dapat dikarenakan algoritma ini efektif dalam menangkap pola non linier dalam data dan dapat menangani *noise* serta variasi pada data yang heterogen dengan nilai akurasi sebesar 71.44%. Hal ini relevan dengan penelitian yang telah dilakukan oleh Mardiani dkk [18] yang melakukan komparasi lima algoritma dalam menganalisis kinerja siswa yang menghasilkan K-NN memberikan hasil performa terbaik dan hasil penelitian ini juga relevan dengan penelitian yang pernah dilakukan [5][19][20] yang melakukan komparasi dan menghasilkan performa K-NN lebih unggul dari algoritma lainnya.

5. Simpulan

Pada penelitian ini telah berhasil membangun model dalam melakukan klasifikasi masa studi dan tingkat stres mahasiswa. Berdasarkan hasil penelitian diperoleh bahwa variabel jenis kelamin sangat berpengaruh terhadap masa studi mahasiswa dan variabel IPK sangat berpengaruh terhadap tingkat stres mahasiswa. Hasil penelitian menunjukkan bahwa algoritma *K-Nearest Neighbor* memiliki performa yang lebih baik dalam mengklasifikasikan masa studi dan tingkat stres mahasiswa dengan tingkat *accuracy* sebesar 71.44 %, *precision* 74.29 %, dan *recall* 93.98 %, pada klasifikasi masa studi dan tingkat *accuracy* 64.17%, *precision* 68.04 %, dan *recall* 88.00% pada klasifikasi tingkat stres mahasiswa.

Daftar Referensi

- [1] S.A.A. Kharis and A.H.A. Zili., "Learning Analytics dan Educational Data Mining pada Data Pendidikan", *JRPMS (Jurnal Riset Pembelajaran Matematika Sekolah)*, vol. 6, no. 1, pp. 12-20, 2022.
- [2] M.M. Abu Tair and A. M. Al-Helees., "Mining Educational Data to Improve Students' Performance: A Case Study", *International Journal of Information and Communication Technology Research*, vol 2, no. 2, pp. 21-30, 2012.
- [3] C. Jalota and R. Agrawal., "Analysis of Educational Data Mining using Classification", *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India, pp. 243-247, 2019, doi: 10.1109/COMITCon.2019.8862214.
- [4] N. Renaningtias., J.E., Suseno., and R. Gernowo., "Hybrid Decision Tree and Naïve Bayes Classifier for Predicting Study Period and Predicate of Student's Graduation", *International Journal of Computer Application*, vol.180, pp. 49-58, 2018
- [5] D. Safitri., S. S. Hilabi, and F. Nurapriani., "Analisis Penggunaan Algoritma Klasifikasi dalam Prediksi Kelulusan Menggunakan Orange Data Mining", *RABIT: Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 8, no. 1, pp. 75-81, 2023.
- [6] A. Trifani, A. P. Windarto, and H. Qurniawan, "Penerapan Data Mining Klasifikasi C4.5 dalam Menentukan Tingkat Stres Mahasiswa Akhir", *Jurnal Riset Rumpun Ilmu Teknik*, vol. 1, no. 2, pp. 91–105, 2022 <https://doi.org/10.55606/jurritek.v1i2.414>
- [7] A.M. Goff, "Stresor, Academic performance and learned resourcefulness in baccalaureate nursing student", *International Journal of Nursing Education Scholarship*, vol. 8, no. 1, pp. 923 – 154, 2011, doi:10.2202/1548-923X.2114
- [8] 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, pp. 35-40, 2020, doi: <https://doi.org/10.30998/faktorexacta.v13i1.5548>
- [9] L.O.M. Zulfiqar., N. Renaningtias., and M.Y. Fathoni., "Educational Data Mining in Graduation Rate and Grade Predictions Utilizing Hybrid Decision Tree and Naïve Bayes Classifier", in *Proceedings of the International Conferences on Information System and Technology (CONRIST 2019)*, pp. 151-157, 2019.
- [10] A. Yandi Saputra and Y. Primadasa, "Penerapan Teknik Klasifikasi Untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritma K-Nearest Neighbour Implementation of Classification Method to Predict Student", *Techno Com*, vol. 17, no. 4, pp. 395-403, doi: 10.33633/tc.v17i4.1864
- [11] F. Alshareef, H. Alhakami, T. Alsubait, and A Baz, "Educational Data Mining Applications and Techniques", *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 729-734, 2020, <http://dx.doi.org/10.14569/IJACSA.2020.0110494>
- [12] Y. Apridiansyah, N.D.M. Veronika, and E.D., Putra, "Prediksi Kelulusan Mahasiswa Fakultas Teknik Informatika Universitas Muhammadiyah Bengkulu Menggunakan Metode Naive Bayes", *JSAI (Journal Scientific and Applied Informatics)*, vol. 4, no. 2, pp. 236–247, 2021 DOI: <https://doi.org/10.36085/jsai.v4i2.1701>
- [13] D. Priyambodo, A. Nugroho, and B. Zaman, "Prediksi Ketepatan Waktu Studi Mahasiswa Bidik Misi Menggunakan K-Nearest Neighbour", *Jurnal Pendidikan Teknologi Informasi (JUKANTI)*, vol. 5, no. 2, pp. 16–22. 2022, doi: <https://doi.org/10.37792/jukanti.v5i2.520>
- [14] E. Siswanto., "Optimasi Metode Naïve Bayes Dalam Memprediksi Tingkat Kelulusan Mahasiswa Stekom Semarang", *Jurikom*, vol. 6 no. 1, pp. 1–13, 2019.
- [15] N.P. Indah, Widodo, and H. Ajie, "Kinerja Algoritma Classification And Regression Tree (Cart) dalam Mengklasifikasikan Lama Masa Studi Mahasiswa yang Mengikuti Organisasi di Universitas Negeri Jakarta", *PINTER: Jurnal Pendidikan Teknik Informatika Dan Komputer*, vol. 3, no. 2, pp. 139–145, 2019.
- [16] I.G.I. Suardika, "Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naive Bayes: Studi Kasus Fakultas Ekonomi Dan Bisnis Universitas Pendidikan Nasional", *Jurnal Ilmu Komputer Indonesia*, vol. 4, no. 2, pp. 37– 44, 2019.
- [17] J. Nasir, "Penerapan Data Mining Clustering Dalam Mengelompokkan Buku Dengan Metode K-Means". *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, vol. 11, no.2, pp. 690–703, 2021 DOI: <https://doi.org/10.24176/simet.v11i2.5482>

- [18] E. Mardiani, et al. "Komparasi Metode Knn, Naive Bayes, Decision Tree, Ensemble, Linear Regression Terhadap Analisis Performa Pelajar SMA". *Innovative: Journal Of Social Science Research*, vol. 3, no. 2, 13880–13892, 2023.
- [19] H. Rajaguru, "Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer", *Asian Pacific journal of cancer prevention: APJCP*, vol. 20, no. 12, 3777–3781, 2019.
- [20] S. Wiyono, and T. Abidin, "Implementation of K-Nearest Neighbour (KNN) Algorithm to Predict Student's Performance", *Jurnal SIMETRIS*, vol. 9, no. 2, pp. 873-878, 2018.
- [21] H. Priatmojo, et al. "Perbandingan Klasifikasi Tingkat Penjualan Buah di Supermarket dengan Pendekatan Algoritma Decision Tree, Naive Bayes dan K-Nearest Neighbor", *Jurnal INSAN (Journal of Information Systems Management Innovation)*, vol. 3, no. 1, pp. 21-28, 2023.
- [22] A.D.R. Wibisono et al. "Comparison of K-Nearest Neighbor and Decision Tree Methods using Principal Component Analysis Technique in Heart Disease Classification", *ijodas*, vol. 4, no. 2, pp. 87-96, Jul. 2023.
- [23] A. Mayssara and A. Hassanin, "Klasifikasi Predikat Kelulusan Mahasiswa dengan Metode CRISP-DM dan Naive Bayes pada Fakultas Komunikasi dan Informatika Universitas Muhammadiyah Surakarta", *Paper Knowledge Toward a Media History of Documents*, vol. 16, 2014.