Jurnal Sains Informatika Terapan (JSIT)
E-ISSN: 2828-1659, Volume: 4, Issue: 2, Month: Juny, Year: 2025
Page: 63-71, Available online at: https://rcf-indonesia.org/home/

KLASIFIKASI NILAI UJIAN SISWA BERDASARKAN KEBIASAAN BELAJAR MENGGUNAKAN K-NEAREST NEIGHBOR DAN SUPPORT VECTOR MACHINE

Rafie Aliefa Khana Kotjek[1)], Pricillia[2)], Filzah Shabrina Wijaya[3)], Mochamad Wahyudi[4)], Sumanto[5)], Ade Surya Budiman[6)]
[1,2,3,4,5,6] Program Studi Informatika, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika
Corresponding Author: [3] 15220620@bsi.ac.id

Article Info

Article history:

Received: mei, 20, 2025
Revised: juni 20, 2025
Accepted: juni 25, 2025
Published: juni 26, 2025

Keywords:

Klasifikasi
K-Nearest Neighbor (kNN)
Support Vector Machine (SVM)
Kebiasaan Belajar
Orange Data Mining

ABSTRACT

Kinerja akademik siswa merupakan indikator penting keberhasilan belajar, namun penilaian konvensional sering kali belum optimal dalam memanfaatkan data kebiasaan dan gaya hidup siswa. Penelitian ini bertujuan untuk mengklasifikasikan nilai ujian siswa (Rendah, Sedang, Tinggi) berdasarkan kebiasaan belajar menggunakan algoritma *K-Nearest Neighbor* (kNN) dan *Support Vector Machine* (SVM), serta membandingkan performa keduanya. Data sebanyak 1000 entri siswa dari Kaggle.com diolah melalui tahap pra-pemrosesan yang meliputi diskritisasi nilai ujian menjadi kategori dan pemilihan fitur yang relevan, seperti jam belajar, persentase kehadiran, waktu tidur, dan peringkat kesehatan mental. Pembagian data dilakukan dengan *random sampling* (90% *training* dan 10% *testing*) yang diulang 10 kali. Hasil evaluasi menunjukkan kNN dengan N=10 mencapai akurasi tertinggi 0.982. Sementara itu, SVM dengan *kernel* Linear memperoleh akurasi 0.974 , diikuti RBF dengan 0.939 , dan Polynomial dengan 0.946 , sedangkan *kernel Sigmoid* hanya 0.712. Performa terbaik kNN (N=10) lebih lanjut dikonfirmasi melalui *confusion matrix*, menunjukkan tingkat kesalahan klasifikasi yang sangat rendah dan prediksi yang konsisten. Penelitian ini menyimpulkan bahwa algoritma k-NN, khususnya dengan N=10, adalah pendekatan yang paling akurat dan efektif untuk klasifikasi nilai ujian berdasarkan kebiasaan siswa, mendukung pihak sekolah dalam prediksi dan perencanaan pendidikan yang lebih baik.

1. INTRODUCTION

Prestasi akademik siswa menjadi indikator seberapa baik siswa dapat mencapai pemahaman terhadap materi pelajaran yang diberikan [1]. Pemahaman dan prediksi terhadap kinerja akademik siswa tiap semester dapat dilakukan melalui proses klasifikasi, dengan mempertimbangkan data poin kedisiplinan sebagai tambahan penilaian, guna memperoleh akurasi dari hasil pengujian data [2]. Dengan demikian, peningkatan prestasi akademik siswa tidak hanya terkait aspek pendidikan, tetapi juga memiliki dampak terhadap keberhasilan mereka setelah menyelesaikan masa studi [3]. Penelitian ini mengadopsi pendekatan data mining untuk mengolah data akademik dan non-akademik secara sistematis, serta menawarkan kontribusi baru dengan mengintegrasikan kebiasaan belajar ke dalam proses klasifikasi nilai. Analisis terhadap perilaku siswa dapat mengungkap pola-pola tersembunyi yang berguna dalam perancangan strategi pembelajaran yang lebih personal dan adaptif [4]. Selain itu, Penerapan teknik data mining memberikan peluang bagi institusi pendidikan untuk melakukan pengambilan keputusan yang lebih tepat dan terarah melalui pemanfaatan data historis serta prediksi berbasis analisis pola akademik [5].

Salah satu tantangan dalam peningkatan mutu pendidikan adalah minimnya pemanfaatan data kebiasaan dan gaya hidup siswa dalam memahami faktor-faktor yang memengaruhi hasil belajar. Selama ini, evaluasi akademik cenderung fokus pada nilai akhir dan aspek kognitif, tanpa mempertimbangkan perilaku harian siswa yang juga berpengaruh signifikan terhadap performa belajar. Penelitian menunjukkan bahwa durasi belajar, kehadiran, pola tidur, dan kesehatan mental memengaruhi tingkat konsentrasi, motivasi, dan kesiapan siswa dalam mengikuti pembelajaran. Namun, pendekatan

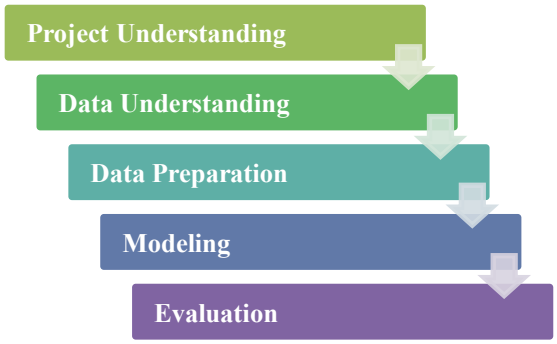
konvensional belum mampu menangkap pola hubungan kompleks dan non-linier dari variabel-variabel tersebut. Akibatnya, institusi pendidikan kekurangan informasi relevan untuk mengidentifikasi risiko atau merancang intervensi berbasis perilaku. Oleh karena itu, dibutuhkan metode berbasis data seperti *data mining* dan *machine learning* yang mampu mengolah informasi kebiasaan siswa secara komprehensif dan menghasilkan model klasifikasi akademik yang akurat.

Beberapa penelitian sebelumnya telah memanfaatkan algoritma klasifikasi *berbasis data mining* dalam menganalisis capaian akademik siswa. Muhaimin et al. [2] menggunakan algoritma K-Nearest Neighbor (k-NN) untuk mengklasifikasikan prestasi siswa berdasarkan nilai rapor dan poin kedisiplinan dari 348 siswa SMA Negeri 2 Batu. Model yang dibangun menunjukkan akurasi tertinggi sebesar 91,39% pada skenario ke-6 dengan *precision* 86%, *recall* 76,4%, dan F1-score 80,92%. Sementara itu, Wulandari et al. [6] menerapkan algoritma *Support Vector Machine* (SVM) untuk memprediksi kelulusan siswa SMA Negeri 1 Kota Lubuklinggau berdasarkan data nilai ujian sekolah. Hasil pengujian menunjukkan akurasi 98,81% untuk kelas XII, 96,49% untuk kelas XI, dan 98,25% untuk kelas X, yang dikategorikan sebagai *excellent classification*. Kedua penelitian tersebut menunjukkan bahwa algoritma klasifikasi dapat diterapkan secara efektif dalam konteks pendidikan.

Namun demikian, belum terdapat studi yang secara langsung membandingkan performa k-NN dan SVM dalam klasifikasi nilai akademik berbasis kebiasaan belajar siswa. Padahal, data seperti durasi belajar, kehadiran, dan pola tidur memiliki potensi signifikan untuk meningkatkan akurasi prediksi akademik. Oleh karena itu, penelitian ini bertujuan melakukan analisis komparatif terhadap algoritma k-NN dan SVM menggunakan platform Orange Data Mining, guna mengevaluasi performa klasifikasi masing-masing metode. Hasil penelitian ini diharapkan dapat menjadi acuan bagi institusi pendidikan dalam mengembangkan sistem evaluasi akademik berbasis data perilaku siswa.

2. MATERIALS AND METHODS

Metode yang digunakan dalam penelitian ini adalah sebagai berikut:



Sumber: Data Science Guide (2023)

Gambar 1. Tahapan Metode CRISP-DM (diadaptasi)

Dalam penelitian ini, tahapan proses yang digunakan mengacu pada model CRISP-DM (*Cross Industry Standard Process for Data Mining*) dengan beberapa penyesuaian sesuai kebutuhan. Model CRISP-DM merupakan salah satu kerangka kerja yang umum digunakan dalam *data mining*, yang mencakup tahapan *project understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*. Namun, pada penelitian ini, tahapan deployment tidak disertakan karena fokus penelitian hanya sampai pada proses evaluasi model klasifikasi. Oleh karena itu, alur kerja disesuaikan menjadi lima tahap utama, yaitu *project understanding*, *data understanding*, *data preparation*, *modeling*, dan *evaluation*. Penyesuaian ini dilakukan untuk menyesuaikan ruang lingkup dan tujuan penelitian, yang tidak berfokus pada penerapan sistem secara langsung tetapi pada pengembangan dan pengujian model klasifikasi berbasis *data mining*. Struktur tahapan ini tetap mempertahankan prinsip dasar CRISP-DM, namun disederhanakan agar relevan dengan konteks studi yang dilakukan.

2.1. *Project Understanding*

Project understanding atau lebih dikenal sebagai *Business understanding* adalah fase awal dalam proyek *data mining* yang berfokus pada pemahaman mendalam terhadap tujuan dan kebutuhan spesifik dari sisi bisnis. Dalam tahap ini, permasalahan atau peluang yang dihadapi bisnis diidentifikasi dan dianalisis secara cermat. Pengetahuan yang diperoleh dari pemahaman bisnis ini kemudian diubah menjadi definisi masalah utama yang jelas dan terukur, yang nantinya dapat diselesaikan secara efektif melalui aplikasi teknik-teknik *data mining* [7].

Penelitian ini dilakukan untuk membangun model klasifikasi yang mampu memprediksi kategori nilai ujian siswa berdasarkan kebiasaan harian mereka. Fokus utama diarahkan pada pemanfaatan data perilaku siswa yang sering diabaikan dalam pendekatan penilaian tradisional. Dengan menggunakan algoritma *k-Nearest Neighbor* (k-NN) dan *Support Vector Machine* (SVM), penelitian ini bertujuan mengungkap hubungan antara aktivitas sehari-hari siswa dan capaian akademik, serta

menghasilkan model prediksi yang akurat untuk mendukung pengambilan keputusan di bidang pendidikan.

2.2. *Data Understanding*

Data understanding adalah fase persiapan esensial dalam proyek data. Tahap ini berfokus pada pengecekan menyeluruh terhadap data yang akan digunakan, dimulai dari pengumpulan data awal hingga identifikasi kualitas data secara mendalam. Dalam proses ini, setiap fitur dalam data akan dideskripsikan secara detail, termasuk karakteristik dan distribusinya, untuk mendapatkan pemahaman komprehensif tentang struktur dan potensi permasalahan data sebelum analisis lebih lanjut [8].

Dataset adalah sebuah koleksi komprehensif dari data yang telah dikumpulkan dan diatur secara sistematis dalam format yang terstruktur. Organisasi ini memfasilitasi kemudahan akses, analisis, dan interpretasi, menjadikannya fondasi esensial untuk berbagai aplikasi seperti penelitian ilmiah, pengembangan model pembelajaran mesin, atau pengambilan keputusan berbasis data [9].

2.3. *Data Preperation*

Data Preperation berfokus pada perhitungan yang akurat. Pada tahapan ini, pembersihan data (data cleaning) merupakan proses penting yang bertujuan untuk menyingkirkan informasi atau data yang tidak relevan agar pengolahan data berjalan lebih efisien dan hasilnya optimal [10].

Data Preperation berfokus pada perhitungan yang akurat. Pada tahapan ini, pembersihan data (data cleaning) merupakan proses penting yang bertujuan untuk menyingkirkan informasi atau data yang tidak relevan agar pengolahan data berjalan lebih efisien dan hasilnya optimal [10].Tahap *data preparation* bertujuan untuk menyiapkan data agar sesuai dengan kebutuhan pemodelan. Proses ini mencakup pemilihan atribut yang relevan (*feature selection*), pengubahan nilai numerik *exam_score* menjadi kategori (*discretization*), penetapan variabel target, serta pembagian data menggunakan metode *random sampling* dengan rasio 90% data latih dan 10% data uji. Seluruh tahapan ini dilakukan agar model dapat bekerja lebih efektif dan akurat.

2.3.1. *Feature Selection*

Langkah pertama yaitu proses *feature selection*. *feature selection* adalah proses penting dalam analisis data dan pembuatan model, di mana kita memilih bagian fitur yang paling relevan dan informatif dari data yang ada. Tujuannya adalah untuk mengurangi dimensi data, meningkatkan kinerja model, menghindari kelebihan informasi, dan mencapai efisiensi [9].

Proses *feature selection* dalam penelitian ini dilakukan secara manual menggunakan Microsoft Excel dengan cara menghapus kolom-kolom yang tidak relevan terhadap tujuan klasifikasi. Atribut yang dipertahankan sebagai fitur adalah jam belajar, persentase kehadiran, waktu tidur, peringkat kesehatan mental, dan nilai ujian karena dianggap memiliki pengaruh langsung terhadap performa akademik siswa. Sementara atribut lain seperti umur, jenis kelamin, dan kebiasaan yang tidak berkaitan langsung dengan pembelajaran dihapus untuk menyederhanakan data dan meningkatkan efektivitas model.

2.3.2. *Discretization*

Langkah kedua yang dilakukan yaitu proses *discretization*, *discretization* adalah sebuah metode pra-pemrosesan data yang berfungsi untuk mengurangi kompleksitas dan jumlah atribut tertentu. Secara spesifik, teknik ini mengubah data yang awalnya bersifat numerik atau kontinu menjadi format data kategorikal, di mana rentang nilai atribut tersebut dibagi ke dalam beberapa interval atau *bin* diskrit yang merepresentasikan kategori-kategori baru [11].

Pada tahap ini, dilakukan pengelompokan nilai *exam_score* ke dalam tiga kategori berdasarkan rentang nilai tertentu, yaitu Rendah untuk nilai di bawah 50, Sedang untuk nilai antara 50 hingga 70, dan Tinggi untuk nilai di atas 70. Proses kategorisasi dilakukan langsung di Microsoft Excel menggunakan rumus logika IF. Nilai-nilai pada kolom *exam_score* diproses satu per satu dengan formula =IF(P2<50;"Rendah";IF(P2<=70;"Sedang";"Tinggi")) dan hasilnya dimasukkan ke dalam kolom baru. Kolom ini kemudian digunakan sebagai variabel target untuk tahap klasifikasi.

2.3.3. *Target Selection*

Langkah ketiga yaitu proses *target selection*, *target selection* adalah proses atau tahap di mana variabel output atau variabel dependen yang ingin diprediksi, diklasifikasikan, atau dianalisis diidentifikasi dan didefinisikan dari kumpulan data yang tersedia.

Target selection dilakukan dengan menetapkan atribut *category* sebagai variabel yang akan diprediksi dalam proses klasifikasi. Atribut ini diperoleh dari hasil *discretization* terhadap nilai *exam_score*. Penetapan dilakukan melalui pengaturan *role* di *widget File* pada Orange Data Mining, di mana kolom *category* diubah perannya menjadi *Target*. Dengan pengaturan ini, sistem mengenali bahwa *category* merupakan kelas *output* yang menjadi fokus dari algoritma klasifikasi, sementara atribut lainnya berfungsi sebagai *input* atau fitur pendukung.

2.3.4. Pembagian Data

Langkah keempat dilakukan pembagian data, pembagian data (*split data*) adalah memisahkan kumpulan data menjadi beberapa bagian, umumnya untuk melatih dan menguji model *machine learning*. Ini dilakukan agar kita bisa mengukur seberapa baik

kinerja model saat dihadapkan pada data baru yang belum pernah digunakan saat pelatihan [12].

Pembagian data dalam penelitian ini menggunakan metode *random sampling* dengan rasio 90% *data training* dan 10% *data testing*. Proses ini dilakukan secara otomatis melalui *widget Test and Score* di Orange Data Mining, yang memungkinkan pemisahan data dilakukan secara acak sesuai proporsi yang telah ditentukan. Untuk menjaga keandalan hasil evaluasi, proses ini diulang sebanyak 10 kali (*repetition*), sehingga model diuji terhadap berbagai kombinasi data acak. Pendekatan ini membantu mengurangi kemungkinan bias akibat pembagian data yang tidak seimbang atau tidak representatif. Penelitian ini hanya menggunakan satu rasio pembagian, yaitu 90:10, dan tidak dilakukan pengujian terhadap rasio alternatif seperti 70:30 atau 80:20. Selain itu, tidak ada pemisahan manual atau pengujian eksternal terhadap data uji, karena seluruh proses evaluasi dilakukan secara internal oleh sistem. Metrik performa seperti akurasi, presisi, dan *recall* diperoleh langsung dari hasil evaluasi yang dihasilkan Orange, sehingga mempercepat proses analisis dan menjaga konsistensi antar percobaan.

2.4. *Modeling*

Modeling dalam penelitian ini menggunakan dua algoritma klasifikasi, yaitu *K-Nearest Neighbors* (k-NN) dan *Support Vector Machine* (SVM). Kedua algoritma dipilih untuk membandingkan performa klasifikasi berdasarkan karakteristik yang berbeda. k-NN digunakan karena kesederhanaannya dan efektivitasnya dalam menghitung jarak antar data.

Algoritma K-Nearest Neighbors (KNN) adalah metode klasifikasi dalam pembelajaran terawasi (*supervised learning*) yang mengklasifikasikan objek atau instans baru. Klasifikasi ini didasarkan pada identifikasi K tetangga terdekat dari objek tersebut dalam data pelatihan. Kelas yang paling sering muncul di antara tetangga-tetangga terdekat inilah yang kemudian menjadi kelas hasil klasifikasi untuk objek baru. Pendekatan KNN bekerja dengan menghitung tingkat kedekatan atau kemiripan antara kasus baru dan kasus-kasus lama (data yang sudah ada), yang perhitungan jaraknya didasarkan pada pembobotan sekumpulan fitur yang tersedia [13].

Support Vector Machine (SVM), sebuah metode klasifikasi yang pertama kali diperkenalkan oleh Vapnik pada tahun 1998, beroperasi dengan mengidentifikasi batas optimal antara dua kelas. Secara fundamental, teknik ini bekerja dengan fokus pada titik-titik data yang paling dekat dengan garis pemisah. Tujuannya adalah untuk membangun *hyperplane* (garis pemisah) terbaik dalam ruang *input*, yang berfungsi sebagai batas dengan margin maksimal untuk memisahkan kelas-kelas tersebut secara efektif [14].

Pengujian dilakukan dengan tiga konfigurasi jumlah tetangga, yaitu $N = 3$, $N = 5$, dan $N = 10$, untuk melihat pengaruh jumlah tetangga terhadap akurasi klasifikasi. Sementara itu, SVM digunakan karena kemampuannya dalam membentuk batas pemisah yang optimal, terutama pada data yang memiliki pola non-linear. SVM diuji dengan empat jenis *kernel* berbeda: RBF (*Radial Basis Function*), *Linear*, *Polynomial*, dan *Sigmoid*. Setiap *kernel* dipilih untuk melihat mana yang paling sesuai dengan karakteristik data siswa yang digunakan dalam penelitian.

2.5. *Evaluation*

Evaluasi model adalah suatu kerangka kerja yang digunakan untuk mengevaluasi suatu sistem atau program. Model evaluasi dapat digunakan untuk mengevaluasi berbagai aspek dari suatu sistem atau program, seperti efektivitas, efisiensi, keandalan, dan keamanan [15].

Evaluasi dilakukan untuk mengukur kinerja model klasifikasi yang dihasilkan oleh algoritma k-NN dan SVM. Pengujian dilakukan menggunakan *confusion matrix* sebagai alat evaluasi utama. Melalui *confusion matrix*, ditinjau seberapa banyak data yang berhasil diklasifikasikan dengan benar maupun salah oleh model. Metrik seperti akurasi, presisi, dan *recall* dihitung secara otomatis untuk mengetahui tingkat efektivitas masing-masing algoritma. Evaluasi ini dilakukan secara internal menggunakan *widget Test and Score*, sehingga hasil pengukuran diperoleh secara sistematis dan konsisten pada setiap pengulangan percobaan.

3. RESULTS AND DISCUSSION

Pada tahap ini, hasil dari penerapan algoritma klasifikasi k-NN dan SVM terhadap *dataset* kebiasaan siswa dalam hubungannya dengan nilai ujian akan dianalisis secara rinci. Pembahasan dimulai dari pemahaman terhadap tujuan proyek dan karakteristik data (*Project Understanding* dan *Data Understanding*), dilanjutkan dengan tahap *Data Preparation* yang mencakup pemilihan atribut, pengelompokan nilai, dan penetapan variabel target. Setelah itu dilakukan pemodelan (*Modeling*) dengan algoritma yang telah ditentukan, serta evaluasi kinerja model berdasarkan metrik akurasi klasifikasi dan *confusion matrix*.

2.1. *Project Understanding*

Penelitian ini bertujuan untuk mengembangkan suatu model klasifikasi yang dapat memprediksi kategori nilai ujian siswa (Rendah, Sedang, Tinggi) berdasarkan pola kebiasaan harian mereka. Latar belakang permasalahan terletak pada keterbatasan pendekatan konvensional yang umumnya hanya berfokus pada penilaian kognitif tanpa mempertimbangkan aspek perilaku dan gaya hidup siswa yang turut memengaruhi performa akademik. Dengan memanfaatkan algoritma *K-Nearest Neighbor* (k-NN) dan *Support Vector Machine* (SVM),

penelitian ini berupaya mengidentifikasi keterkaitan antara variabel-variabel kebiasaan siswa dan hasil capaian akademik mereka, guna menghasilkan model prediksi yang akurat serta dapat mendukung proses pengambilan keputusan dalam konteks pendidikan.

2.2. *Data Understanding*

Dataset yang digunakan dalam penelitian ini diperoleh dari platform www.kaggle.com, yang berisi 1000 entri data siswa lengkap dengan berbagai atribut kebiasaan dan latar belakang pribadi. Struktur data tersebut, sebagaimana disajikan dalam Tabel 1, terdiri dari 16 kolom yang mencakup variabel seperti *student_id*, *age*, *gender*, *study_hours_per_day*, *social_media_hours*, *netflix_hours*, *part_time_job*, *attendance_percentage*, *sleep_hours*, *diet_quality*, *exercise_frequency*, *parental_education_level*, *internet_quality*, *mental_health_rating*, *extracurricular_participation*, dan *exam_score*.

Proses *data understanding* dilakukan melalui eksplorasi awal yang bertujuan untuk menilai kelengkapan data, memeriksa distribusi nilai dari masing-masing atribut, serta mengidentifikasi adanya pola atau kecenderungan umum dalam data. Langkah ini juga melibatkan interpretasi semantik terhadap setiap atribut guna memahami konteks pengukuran dan implikasinya terhadap performa akademik. Selain itu, dilakukan penyesuaian awal terhadap format data agar siap untuk diproses pada tahap berikutnya. Atribut nilai ujian (*exam_score*) diidentifikasi sebagai variabel target (*dependent variable*) yang akan digunakan dalam proses klasifikasi, sedangkan variabel lainnya berfungsi sebagai kandidat fitur (*independent variables*) yang relevan.

Langkah ini memberikan dasar pemahaman yang kuat terhadap struktur dan kualitas *dataset*, serta mendukung pemilihan fitur yang tepat untuk proses pemodelan selanjutnya.

Tabel 1a *Dataset* Kinerja Kebiasaan Siswa

student_id	age	gender	study_hours_per_day
S1000	23	Female	0
S1001	20	Female	6,9
S1002	21	Male	1,4
S1003	23	Female	1
S1004	19	Female	5

Tabel 1b *Dataset* Kinerja Kebiasaan Siswa

social_media_hours	netflix_hours	part_time_job	attendance_percentage
1,2	1,1	No	85
2,8	2,3	No	97,3
3,1	1,3	No	94,8
3,9	1	No	71
4,4	0,5	No	90,9

Tabel 1c *Dataset* Kinerja Kebiasaan Siswa

sleep_hours	diet_quality	exercise_frequency	parental_education_level
8	Fair	6	Master
4,6	Good	6	High School
8	Poor	1	High School
9,2	Poor	4	Master
4,9	Fair	3	Master

Tabel 1d *Dataset* Kinerja Kebiasaan Siswa

internet_quality	mental_health_rating	extracurricular_participation	exam_score
Average	8	Yes	56,2
Average	8	No	100
Poor	1	No	34,3
Good	1	Yes	26,8
Good	1	No	66,4

Sumber: www.kaggle.com (2025)

2.3. *Data Preperation*

Pada tahap ini, data diolah melalui beberapa langkah utama, yaitu pemilihan atribut penting yang memengaruhi nilai ujian, pengelompokan nilai *exam_score* ke dalam tiga kategori, penetapan atribut kategori sebagai target klasifikasi, dan pembagian data latih dan uji menggunakan Orange Data Mining. Proses ini memastikan data siap digunakan untuk klasifikasi secara optimal.

2.3.1. *Feature Selection*

Feature selection dilakukan dengan menyaring atribut-atribut penting secara manual di Excel. Kolom-kolom yang tidak relevan, seperti umur dan jenis kelamin, dihapus dari *dataset*. Lima atribut yang dipertahankan adalah jam belajar, persentase kehadiran, waktu tidur, peringkat kesehatan mental, dan nilai ujian, karena dinilai paling berkaitan dengan pencapaian akademik siswa. Langkah ini membantu menyederhanakan data dan meningkatkan fokus klasifikasi.

	A	B	C	D	E	F
1	study_hours_per_day	attendance_percentage	sleep_hours	mental_health_rating	exam_score	
2	0	85	8	8	56,2	
3	6,9	97,3	4,6	8	100	
4	1,4	94,8	8	1	34,3	
5	1	71	9,2	1	26,8	
6	5	90,9	4,9	1	66,4	
7	7,2	82,9	7,4	4	100	
8	5,6	85,8	6,5	4	89,8	
9	4,3	77,7	4,6	8	72,6	
10	4,4	100	7,1	1	78,9	
11	4,8	95,4	7,5	10	100	
12	4,6	77,6	5,8	3	63,3	
13	3,9	71,7	7,9	1	74,4	
14	3,7	81,1	4,5	9	76,9	
15	3,4	89,3	4,7	10	75,8	
16	2,4	87,4	6,7	9	78,9	
17	3,1	97,5	6,5	7	74	
18	1	92,9	5,6	8	55,2	
19	3,4	94,7	7,5	1	70,8	
20	2	88,3	7,1	5	43,9	
21	1,8	71,1	7,5	2	45,3	

Sumber: Hasil Penelitian (2025)

Gambar 2. *Feature Selection dataset* Kinerja Kebiasaan Siswa

2.3.2. *Discretization*

Discretization dilakukan dengan mengubah nilai ujian menjadi tiga kelas: Rendah, Sedang, dan Tinggi. Proses ini dikerjakan secara manual di Excel dengan menerapkan formula IF untuk mengelompokkan skor setiap siswa berdasarkan batas nilai tertentu. Kolom hasil pengelompokan ditambahkan ke *dataset* sebagai bentuk kategori dari nilai numerik sebelumnya. Hasil akhir dari proses ini menjadi acuan utama dalam menentukan label target yang akan digunakan oleh model klasifikasi pada tahap modeling selanjutnya.

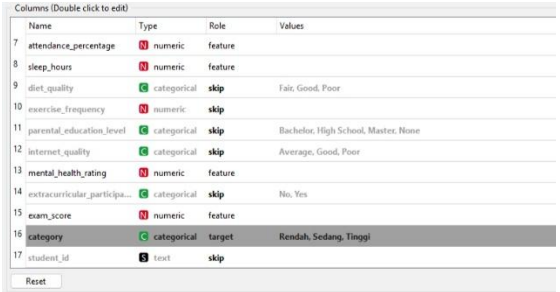
	A	B	C	D	E	F	G
1	study_hours_per_day	attendance_percentage	sleep_hours	mental_health_rating	exam_score	category	
2	0	85	8	8	56,2	Sedang	
3	6,9	97,3	4,6	8	100	Tinggi	
4	1,4	94,8	8	1	34,3	Rendah	
5	1	71	9,2	1	26,8	Rendah	
6	5	90,9	4,9	1	66,4	Sedang	
7	7,2	82,9	7,4	4	100	Tinggi	
8	5,6	85,8	6,5	4	89,8	Tinggi	
9	4,3	77,7	4,6	8	72,6	Tinggi	
10	4,4	100	7,1	1	78,9	Tinggi	
11	4,8	95,4	7,5	10	100	Tinggi	
12	4,6	77,6	5,8	3	63,3	Sedang	
13	3,9	71,7	7,9	1	74,4	Tinggi	
14	3,7	81,1	4,5	9	76,9	Tinggi	
15	3,4	89,3	4,7	10	75,8	Tinggi	
16	2,4	87,4	6,7	9	78,9	Tinggi	
17	3,1	97,5	6,5	7	74	Tinggi	
18	1	92,9	5,6	8	55,2	Sedang	
19	3,4	94,7	7,5	1	70,8	Tinggi	
20	2	88,3	7,1	5	43,9	Rendah	
21	1,8	71,1	7,5	2	45,3	Rendah	

Sumber: Hasil Penelitian (2025)

Gambar 3. Hasil *Discretization dataset* Kinerja Kebiasaan Siswa

2.3.3. *Target Selection*

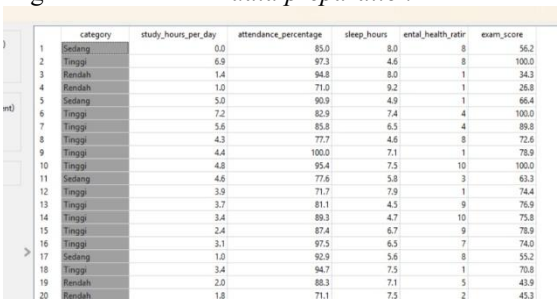
Pada tahap ini, atribut *category* yang memuat kelas Rendah, Sedang, dan Tinggi ditetapkan sebagai target klasifikasi. Penentuan dilakukan di Orange Data Mining dengan mengatur peran kolom tersebut sebagai *Target* dalam *widget File*. Langkah ini memungkinkan model untuk memfokuskan proses pembelajaran terhadap label kategori hasil *discretization*, sehingga semua prediksi diarahkan untuk menentukan kelas nilai ujian berdasarkan *input* dari atribut-atribut kebiasaan siswa.



Sumber: Hasil Penelitian (2025)

Gambar 4. *Target Selection dataset* Kinerja Kebiasaan Siswa

Berikut adalah hasil akhir dari tampilan *dataset* yang sudah dilakukan *data preparation*.



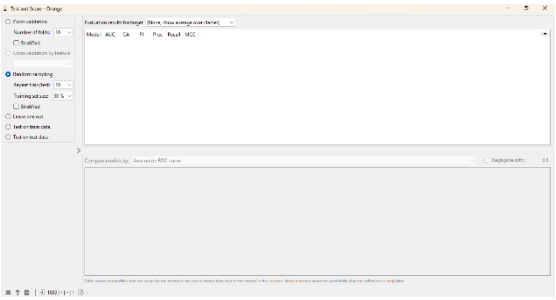
	category	study_hours_per_day	attendance_percentage	sleep_hours	ental_health_ratir	exam_score
1	Sedang	0.0	85.0	8.0	8	56.2
2	Tinggi	6.9	97.3	4.6	8	100.0
3	Rendah	1.4	94.8	8.0	1	34.3
4	Rendah	1.0	71.0	9.2	1	26.8
5	Sedang	5.0	90.9	4.9	1	66.4
6	Tinggi	7.2	82.9	7.4	4	100.0
7	Tinggi	5.6	85.8	6.5	4	89.8
8	Tinggi	4.3	77.7	4.6	8	72.6
9	Tinggi	4.4	100.0	7.1	1	78.9
10	Tinggi	4.8	95.4	7.5	10	100.0
11	Sedang	4.6	77.6	5.8	3	63.3
12	Tinggi	3.9	71.7	7.9	1	74.4
13	Tinggi	3.7	81.1	4.5	9	76.9
14	Tinggi	3.4	89.3	4.7	10	75.8
15	Tinggi	2.4	87.4	6.7	9	78.9
16	Tinggi	3.1	97.5	6.5	7	74.0
17	Sedang	1.0	92.9	5.6	8	55.2
18	Tinggi	3.4	94.7	7.5	1	70.8
19	Rendah	2.0	88.3	7.1	5	43.9
20	Rendah	1.8	71.1	7.5	2	45.3

Sumber: Hasil Penelitian (2025)

Gambar 5. Hasil *Data Preparation dataset* Kinerja Kebiasaan Siswa

2.3.4. Pembagian Data

Pada tahap ini, data dibagi menjadi dua bagian, yaitu 90% untuk pelatihan model dan 10% untuk pengujian akurasi model. Proses pembagian ini dilakukan secara otomatis menggunakan *widget Test and Score* yang tersedia di Orange Data Mining. Untuk memastikan hasil evaluasi tidak bergantung pada satu pembagian saja, proses ini diulang sebanyak 10 kali dengan pengacakan data pada setiap pengulangan. Rasio 90:10 dipilih karena dianggap cukup representatif dan sering digunakan dalam penelitian klasifikasi.



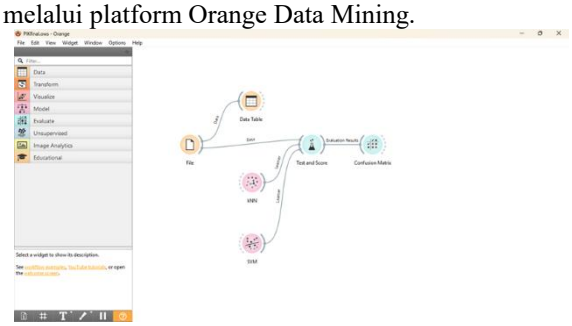
Sumber: Hasil Penelitian (2025)

Gambar 6. Pembagian Data *dataset* Kinerja Kebiasaan Siswa

2.4. *Modeling*

Pada tahap ini, proses pemodelan dilakukan menggunakan algoritma k-NN dan SVM untuk mengklasifikasikan kategori nilai ujian siswa berdasarkan atribut kebiasaan. Algoritma k-NN diimplementasikan dengan tiga konfigurasi jumlah tetangga, yaitu 3, 5, dan 10, guna membandingkan tingkat akurasi yang dihasilkan oleh masing-masing konfigurasi. Hasil dari setiap konfigurasi diuji melalui data yang telah dibagi sebelumnya. Untuk algoritma SVM, digunakan empat jenis *kernel*, yaitu RBF, *Linear*, *Polynomial*, dan *Sigmoid*, yang masing-masing memiliki pendekatan berbeda dalam membentuk batas klasifikasi. Setiap model diuji secara terpisah untuk melihat konfigurasi mana yang

memberikan hasil terbaik terhadap *dataset* yang digunakan, dan seluruh proses pengujian dilakukan melalui platform Orange Data Mining.



Sumber: Hasil Penelitian (2025)

Gambar 7. *Modeling* k-NN dan SVM pada *dataset* Kinerja Kebiasaan Siswa

2.5. *Evaluation*

Evaluation dilakukan untuk mengukur tingkat keberhasilan model dalam mengklasifikasikan data secara akurat. Pada penelitian ini, evaluasi difokuskan pada metrik *Classification Accuracy* (CA), yaitu persentase data yang berhasil diklasifikasikan dengan benar oleh model terhadap total data yang diuji. Metode ini dipilih karena dapat memberikan gambaran umum mengenai performa model secara langsung dan mudah dipahami. Evaluasi dilakukan terhadap dua algoritma yang digunakan, yaitu *k-Nearest Neighbor* (k-NN) dengan tiga konfigurasi nilai *N* (3, 5, dan 10), serta *Support Vector Machine* (SVM) dengan empat jenis *kernel* berbeda (RBF, *Linear*, *Polynomial*, dan *Sigmoid*). Seluruh proses evaluasi dijalankan menggunakan *widget Test and Score* di Orange Data Mining, dan hasil pengujian performa masing-masing konfigurasi disajikan dalam Tabel 2.

Tabel 2. Hasil Perbandingan Evaluasi k-NN dan SVM

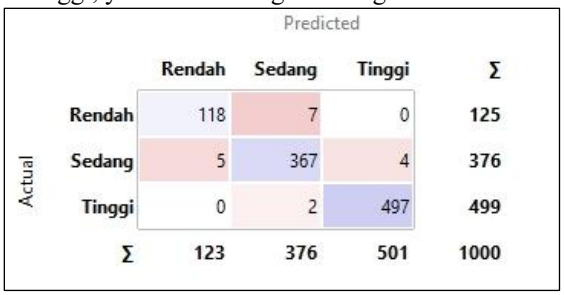
Algoritma	Konfigurasi	*Classification Accuracy* (CA)
k-NN	*N* = 3	0.976
	N = 5	0.976
	N = 10	0.982
SVM	RBF	0.939
	Linear	0.974
	Polynomial	0.946
	Sigmoid	0.712

Sumber: Hasil Penelitian (2025)

Berdasarkan dari Tabel 1 hasil analisis pengujian *Classification Accuracy* (CA) dapat disajikan sebagai berikut:

a. **k-NN dengan *N* = 3** memperoleh nilai akurasi sebesar 0.976. Nilai ini menunjukkan bahwa konfigurasi ini mampu memberikan hasil klasifikasi yang sangat baik dengan tingkat kesalahan yang rendah.
b. **k-NN dengan *N* = 5** juga memperoleh akurasi sebesar 0.976, sama seperti *N* = 3. Hal ini menunjukkan bahwa pada data yang digunakan, perubahan jumlah tetangga dari 3 ke 5 tidak memengaruhi performa secara signifikan.
c. **k-NN dengan *N* = 10** menunjukkan performa terbaik dari seluruh konfigurasi yang diuji, dengan nilai akurasi tertinggi yaitu 0.982. Ini menunjukkan bahwa konfigurasi ini paling optimal dalam proses klasifikasi pada data yang digunakan.
d. **SVM dengan *kernel* RBF** memperoleh nilai akurasi sebesar 0.939. Ini menunjukkan performa yang sangat baik dalam klasifikasi, dan menjadi konfigurasi SVM dengan akurasi tertinggi dalam pengujian ini.
e. **SVM dengan *kernel Linear*** menghasilkan nilai akurasi sebesar 0.974. Ini menunjukkan bahwa linear *kernel* cukup andal untuk digunakan dalam klasifikasi, mendekati performa terbaik dari k-NN.
f. **SVM dengan *kernel Polynomial*** memperoleh akurasi sebesar 0.946. Hasil ini masih tergolong baik, namun lebih rendah dibandingkan *kernel* RBF dan *Linear*, yang berarti *Polynomial kernel* mungkin kurang optimal untuk data ini.
g. **SVM dengan *kernel Sigmoid*** memberikan performa paling rendah di antara semua konfigurasi, dengan nilai akurasi *hanya* 0.712. Ini menunjukkan bahwa konfigurasi ini kurang cocok untuk data yang digunakan karena menghasilkan tingkat kesalahan klasifikasi yang tinggi.

Evaluasi hasil klasifikasi terbaik disajikan melalui *confusion matrix* dari algoritma dengan akurasi tertinggi, yaitu k-NN dengan konfigurasi *N* = 10.



Sumber: Hasil Penelitian (2025)

Gambar 8. Evaluasi *Confusion Matrix* Algoritma k-NN *N* = 10

Berdasarkan *confusion matrix* pada Gambar 8, performa klasifikasi algoritma *k-Nearest Neighbor* (k-NN) dengan konfigurasi *N* = 10 menunjukkan tingkat akurasi yang sangat tinggi. Evaluasi dilakukan terhadap tiga kelas target, yaitu Rendah, Sedang, dan Tinggi, dengan rincian sebagai berikut:

2.5.1. Kelas Rendah

a. Sebanyak 118 data dari total 125 data berhasil diklasifikasikan secara benar ke dalam kelas

Rendah, menunjukkan tingkat akurasi yang tinggi dalam mendeteksi kategori ini.

b. Terdapat 7 data yang mengalami kesalahan klasifikasi ke kelas Sedang, sementara tidak terdapat kesalahan klasifikasi ke kelas Tinggi.

2.5.2. Kelas Sedang

a. Algoritma berhasil mengklasifikasikan secara tepat 367 data ke dalam kelas Sedang.
b. Sebanyak 5 data salah diklasifikasikan ke kelas Rendah, dan 4 data diklasifikasikan secara keliru ke kelas Tinggi.

2.5.3. Kelas Tinggi

a. Sebanyak 497 data dari total 499 berhasil dikenali secara tepat sebagai kelas Tinggi.
b. Hanya 2 data yang diklasifikasikan secara keliru ke kelas Sedang, dan tidak terdapat kesalahan klasifikasi ke kelas Rendah.

Hasil ini menunjukkan bahwa konfigurasi k-NN dengan $N = 10$ memiliki kapabilitas yang sangat baik dalam membedakan antar kelas, dengan tingkat kesalahan klasifikasi yang sangat rendah serta distribusi prediksi yang konsisten terhadap nilai aktual. Tingginya akurasi model ini mencerminkan keandalannya dalam proses klasifikasi pada data yang digunakan dalam jurnal ini.

4. CONCLUSION

Penelitian ini berhasil membangun model klasifikasi nilai ujian siswa berdasarkan data kebiasaan belajar harian dengan menggunakan algoritma *K-Nearest Neighbor* (k-NN) *dan Support Vector Machine* (SVM). Hasil pengujian menunjukkan bahwa algoritma k-NN dengan konfigurasi $N = 10$ memberikan performa terbaik dengan akurasi sebesar 98,2%, mengungguli seluruh konfigurasi *kernel* pada SVM. Hal ini menunjukkan bahwa model klasifikasi berbasis kebiasaan belajar memiliki kemampuan yang tinggi dalam memetakan performa akademik siswa secara presisi.

Temuan tersebut menegaskan bahwa variabel perilaku seperti jam belajar, tingkat kehadiran, pola tidur, dan kondisi kesehatan mental memiliki kontribusi signifikan terhadap capaian akademik, dan dapat dimanfaatkan sebagai prediktor dalam sistem evaluasi pendidikan yang lebih adaptif. Pendekatan ini memberikan alternatif yang lebih komprehensif dibandingkan metode evaluasi konvensional yang hanya berfokus pada aspek kognitif.

Sebagai tindak lanjut, penelitian ini memiliki potensi untuk dikembangkan melalui eksplorasi algoritma lain seperti *Naive Bayes* dan *Decision Tree*, serta penerapan pada *dataset* yang lebih besar dan heterogen guna meningkatkan generalisasi model. Selain itu, integrasi model klasifikasi ini ke dalam sistem informasi akademik dapat menjadi strategi efektif dalam mendukung pengambilan keputusan berbasis data di lingkungan pendidikan.

REFERENCES

[1] F. W. Prima and Z. Fikry, "Pengaruh Keterlibatan Orang Tua terhadap Performa Akademik Siswa Kelas 2 Jurusan IPS di SMAN 4 Kota Sungai Penuh," *Jurnal Pendidikan Tambusai*, vol. 5, no. 2, pp. 3998–4006, 2021.

[2] A. Muhaimin, M. Amin Hariyadi, and M. I. Imamudin, "Klasifikasi Prestasi Akademik Siswa Berdasarkan Nilai Rapor dan Kedisiplinan dengan Metode K-Nearest Neighbor," *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI)*, vol. 7, no. 1, pp. 193–202, 2024, doi: 10.55338/jikomsi.v7i1.2865.

[3] T. Gori, A. Sunyoto, and H. Al Fatta, "Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 1, pp. 215–224, 2024, doi: 10.25126/jtiik.20241118074.

[4] J. Hutahaean and S. Amelia, "Teknologi Pendidikan Penggunaan Teknologi Big Data Untuk Menganalisis Perilaku Teknologi Pendidikan," *Teknologi Pendidikan*, vol. 3, no. 1, pp. 152–160, 2024, doi: 10.56854/tp.v3i1.232.

[5] R. E. Pambudi, H. Purnomo, and R. Irawan, "Pemanfaatan Data Mining Untuk Prediksi Prestasi Akademik Siswa Berdasarkan Pola Kehadiran, Aktivitas Belajar Mengguakan Naive Bayes Logistic Regression," *Jurnal Teknologi Informasi Mura*, vol. 16, no. 2, pp. 132–141, 2024.

[6] C. Wulandari, T. Hasanah Bimastari Aviani, and R. Saputra, "Penerapan Algortima Support Vector Machine (SVM) Untuk Prediksi Tingkat Kelulusan Siswa SMA," *RESOLUSI : Rekayasa Teknik Informatika dan Informasi*, vol. 4, no. 4, pp. 397–407, 2024, [Online]. Available: https://djournals.com/resolusi/article/view/1753

[7] R. Winurputra and D. E. Ratnawati, "PERAMALAN PENJUALAN PRODUK MENGGUNAKAN EXTREME GRADIENT BOOSTING (XGBOOST) DAN KERANGKA KERJA CRISP-DM UNTUK PENGOPTIMALAN MANAJEMEN PERSEDIAAN (STUDI KASUS : UB MART) PRODUCT SALES FORECASTING USING EXTREME GRADIENT BOOSTING (XGBOOST) AND CRISP-DM ," vol. 12, no. 2, pp. 417–428, 2025, doi: 10.25126/jtiik.2025129451.

[8] F. N. Dhewayani, D. Amelia, D. N. Alifah, B. N. Sari, and M. Jajuli, "Implementasi K-Means Clustering untuk Pengelompokkan Daerah Rawan Bencana Kebakaran Menggunakan Model CRISP-DM," *Jurnal Teknologi dan Informasi*, vol. 12, no. 1, pp. 64–77, 2022, doi: 10.34010/jati.v12i1.6674.

[9] H. Tantyoko, D. K. Sari, and A. R. Wijaya, "Prediksi Potensial Gempa Bumi Indonesia Menggunakan Metode Random Forest Dan Feature Selection," *IDEALIS : InDonEsiA journaL Information System*, vol. 6, no. 2, pp. 83–89, 2023, doi: 10.36080/idealis.v6i2.3036.

[10] F. Rahmadayanti, I. Anggraini, and T. Susanti, "Pengklasterisasian Data Penyakit Hipertensi dengan Menggunakan Metode K-Means," *Journal of Information System Research (JOSH)*, vol. 4, no. 2, pp. 737–741, 2023, doi: 10.47065/josh.v4i2.2905.

[11] D. Wintana, Gunawan, H. Sulaeman, and S. Bahri, "Penerapan Multi Layer Perceptron dan Diskrit pada Prediksi Cacat Software," *J-INTECH (Journal of Information and Technology)*, vol. 12, no. 2, pp. 321–329, 2022, [Online]. Available: https://snatika.stiki.ac.id/J-INTECH/article/view/1422/847

[12] F. Putra, H. F. Tahiyat, R. M. Ihsan, R. Rahmaddeni, and L. Efrizoni, "Penerapan Algoritma K-Nearest Neighbor Menggunakan Wrapper Sebagai Preprocessing untuk

Penentuan Keterangan Berat Badan Manusia," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 1, pp. 273–281, 2024, doi: 10.57152/malcom.v4i1.1085.

[13] Devi. Putri, "KLASIFIKASI PENYAKIT GAGAL GINJAL KRONIS DENGAN METODE KNN (STUDI KASUS RS DI KAB GRESIK)," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 9, no. 3, pp. 1739–1748, 2024.

[14] H. Harnelia, "Analisis Sentimen Review Skincare Skintific Dengan Algoritma Support Vector Machine (Svm)," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 2, 2024, doi: 10.23960/jitet.v12i2.4095.

[15] A. Rama, A. Ambiyar, F. Rizal, N. Jalinus, W. Waskito, and R. E. Wulansari, "Konsep model evaluasi context, input, process dan product (CIPP) di sekolah menengah kejuruan," *JRTI (Jurnal Riset Tindakan Indonesia)*, vol. 8, no. 1, p. 82, 2023, doi: 10.29210/30032976000.