(2025), 3 (6): 97-115 **Scientica** 3021-8209

Jurnal Ilmiah Sain dan Teknologi

PREDIKSI MAHASISWA YANG BERPOTENSI MELAKUKAN PENGUNDURAN DIRI MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE

Septian Dwirizkianto [1], Tursina [2], Niken Candraningrum [3]
Jurusan Informatika, Fakultas Teknik, Universitas Tanjungpura
Jl. Prof. Dr. H. Hadari Nawawi, Pontianak 78124
[1] d1041201014@student.untan.ac.id [2] tursina@informatika.untan.ac.id
[3] nikenc@informatika.untan.ac.id

Abstract (English)
Student resignation is a crucial issue that impacts individuals, institutions, and the overall quality of education. This research aims to predict students who have the potential to resign using the Support Vector Machine (SVM) algorithm. The data used comes from students of the Faculty of Engineering, Tanjungpura University, including academic and non-academic data of students from batch 2014 to 2021. The research methodology includes data collection, data preprocessing, model development, testing, and model performance evaluation. The preprocessing process includes attribute selection, data cleaning using Logical Imputation Listwise Deletion, data transformation, feature extraction using Principal Component Analysis (PCA), and normalization using the Min-Max Normalization and Standard Scaler methods. The model is developed using SVM algorithm and hyperparameter tuning, then tested using Mean Squared Error (MSE), Root Mean Square Error (RMSE), and Confusion Matrix. Based on the evaluation results, the SVM model shows good performance with testing data accuracy of 93.54%, MSE of 0.07317, and RMSE of 0.27047. In addition, the prediction results show a high level of conformity with the actual data, which is 96.13%, where out of 83 students who actually resigned, 74 students were successfully detected through the model, and 9 students were not detected. Further analysis shows that students who enter through independent pathways, graduates from high school, and those who experience a decrease in the Semester Achievement Index (IPS) from semester 1 to 4 have a higher tendency to resign. This finding is expected to be a reference for institutions in developing early prevention strategies to reduce student withdrawal rates.

Article History
Submitted: 10 September 2025
Accepted: 13 September 2025
Published: 14 September 2025

Key Words
Student Resignation, Support Vector Machine (SVM), Student Prediction, Machine Learning, Data Mining.

Abstrak (Indonesia)
Pengunduran diri mahasiswa merupakan isu krusial yang berdampak pada individu, institusi, dan kualitas pendidikan secara keseluruhan. Penelitian ini bertujuan untuk memprediksi mahasiswa yang berpotensi mengundurkan diri menggunakan algoritma Support Vector Machine (SVM). Data yang digunakan berasal dari mahasiswa Fakultas Teknik Universitas Tanjungpura, mencakup data akademik dan non-akademik mahasiswa dari angkatan 2014 hingga 2021. Metodologi penelitian mencakup pengumpulan data, preprocessing data, pengembangan model, pengujian, serta evaluasi kinerja model. Proses preprocessing mencakup pemilihan atribut, pembersihan data menggunakan Logical Imputation Listwise Deletion, transformasi data, ekstraksi fitur menggunakan Principal Component Analysis (PCA), serta normalisasi dengan metode Min-Max Normalization dan Standard Scaler. Model dikembangkan menggunakan algoritma SVM dan hyperparameter tuning, kemudian diuji menggunakan Mean Squared Error (MSE), Root Mean Square Error (RMSE), serta Confusion Matrix. Berdasarkan hasil evaluasi, model SVM menunjukkan performa yang baik dengan akurasi data testing sebesar 93,54%, MSE sebesar 0,07317, dan RMSE sebesar 0,27047. Selain itu,

Sejarah Artikel
Submitted: 10 September 2025
Accepted: 13 September 2025
Published: 14 September 2025

Kata Kunci
Pengunduran Diri Mahasiswa, Support Vector Machine (SVM), Prediksi Mahasiswa, Machine Learning, Data Mining.

hasil prediksi menunjukkan tingkat kesesuaian yang tinggi dengan data aktual, yaitu sebesar 96,13%, di mana dari 83 mahasiswa yang benar-benar mengundurkan diri, sebanyak 74 mahasiswa berhasil terdeteksi melalui model, dan 9 mahasiswa tidak terdeteksi. Analisis lebih lanjut menunjukkan bahwa mahasiswa yang masuk melalui jalur mandiri, lulusan dari SMA, serta mereka yang mengalami penurunan Indeks Prestasi Semester (IPS) dari semester 1 hingga 4 memiliki kecenderungan lebih tinggi untuk mengundurkan diri. Temuan ini diharapkan dapat menjadi acuan bagi pihak institusi dalam menyusun strategi pencegahan dini untuk menekan angka pengunduran diri mahasiswa.

I. PENDAHULUAN

Pendidikan tinggi di Indonesia menunjukkan peningkatan jumlah mahasiswa setiap tahunnya. Data dari BPS dan Kemendikbudristek mencatat kenaikan dari 8,6 juta mahasiswa pada tahun 2020 menjadi 9,32 juta pada tahun 2022. Namun, terdapat kesenjangan antara jumlah mahasiswa baru (rata-rata 350 ribu per tahun) dan lulusan (rata-rata 240 ribu per tahun), dengan sekitar 110 ribu mahasiswa per tahun berstatus tidak diketahui, seperti mengundurkan diri, tidak menyelesaikan studi tepat waktu, atau drop out. Fenomena ini menjadi isu kritis di universitas, termasuk Universitas Tanjungpura, karena berdampak pada reputasi dan kinerja akademik institusi.

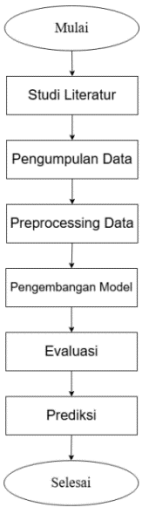
Berdasarkan observasi dan wawancara terhadap beberapa mahasiswa aktif dari berbagai program studi di Fakultas Teknik Universitas Tanjungpura, diketahui banyak dari teman seangkatan mereka yang memutuskan untuk mengundurkan diri. Upaya yang dilakukan untuk mengurangi jumlah mahasiswa yang melakukan pengunduran diri dengan cara melakukan analisis pola atau informasi yang didapatkan dari pusat data perguruan tinggi mengenai faktor - faktor yang memiliki pengaruh terhadap mahasiswa yang memilih mengundurkan diri dengan menggunakan *data mining* [1].

Data mining adalah proses yang menggunakan statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar [2]. Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD). Data Mining atau Knowledge Discovery in Database (KDD) adalah penyaringan data secara implisit dimana sebelumnya tidak diketahui terdapatnya informasi yang potensial [3]. Data mining sendiri dibagi menjadi 6 kelompok berdasarkan tugas atau pekerjaan yang dapat dilakukan, diantaranya adalah Deskripsi, Klasifikasi, Estimasi, Asosiasi, Pengklasteran, dan Prediksi.

Penelitian ini bertujuan untuk memprediksi mahasiswa yang berpotensi mengundurkan diri menggunakan algoritma Support Vector Machine (SVM). Studi sebelumnya menunjukkan bahwa SVM memiliki akurasi hingga 95% dalam memprediksi status [4]. Fokus penelitian ini adalah menganalisis performa SVM dan hasil prediksinya terhadap mahasiswa aktif angkatan 2021, dengan dataset mencakup data mahasiswa aktif dan yang mengundurkan diri dari angkatan 2014–2021. Hasil penelitian diharapkan dapat membantu universitas mengidentifikasi dan mengantisipasi mahasiswa berisiko, sehingga mengurangi tingkat pengunduran diri.

II. METODOLOGI

Penelitian ini menggunakan metodologi yang direpresentasikan dalam diagram alir penlitian yang tertera pada Gambar 1.



Gambar 1. Diagram Alir Metodologi Penelitian

A. *Studi Literatur*

Tahap studi literatur ini bertujuan untuk mendapatkan pemahaman terhadap teori dan dasar pengetahuan yang memiliki keterkaitan dengan penelitian ini. Literatur yang dicari diperoleh dari jurnal, artikel, dan studi kasus berdasarkan penelitian – penelitian terdahulu yang dapat dijadikan acuan dan landasan dalam penelitian. Adapun literatur yang dibaca diantaranya adalah penelitian yang dilakukan oleh Bode, A (2019)[5], Mahanggara & Laksito (2019)[6], Sinanto, dkk (2022)[7].

B. *Pengumpulan Data*

Data yang digunakan dalam penelitian ini merupakan data mahasiswa aktif dan mahasiswa yang telah mengundurkan diri dari angkatan 2014 hingga angkatan 2021 Fakultas Teknik Universitas Tanjungpura. Data yang digunakan didapat dari Biro Akademik dan Kemahasiswaan (BAK) Universitas Tanjungpura. Data yang digunakan dalam proses klasifikasi merupakan data akademik dan data nonakademik mahasiswa. Atribut data yang akan digunakan untuk proses pengembangan model prediksi disesuaikan dengan data – data yang akan diberikan oleh pihak Biro Akademik dan Kemahasiswaan

C. *Preprocessing Data*

Pada tahap *preprocessing* data dilakukan proses pada data mentah agar data tersebut dapat digunakan untuk membuat model prediksi sehingga mendapatkan hasil prediksi yang akurat dan konsisten. Berikut merupakan tahapan-tahapan yang dilakukan pada proses *preprocessing*:

1) *Pemilihan Atribut:* Langkah pertama adalah menentukan atribut atau varibel apa saja yang akan digunakan untuk proses pengembangan model prediksi. Pada penelitian ini memiliki 14 atribut, yaitu jalur masuk, jurusan SLTA, pekerjaan ayah, pekerjaan ibu, pendapatan ayah, pendapatan ibu, status mengundurkan diri, jumlah SKS semester 2 sampai 4, dan nilai IPS semester 1 sampai 4.
2) *Pembersihan Data:* Tahap kedua dalam proses preprocessing yaitu *data cleaning* atau pembersihan data yang dilakukan untuk mengidentifikasi dan mengatasi *missing value* yang dapat menyebabkan bias dalam analisis. Untuk mengisi nilai yang hilang pada data penelitian ini dilakukan dengan menggunakan proses *Logical Imputation*, yaitu mengisi nilai dalam dataset menggunakan nilai pada atribut lain yang berkaitan. Untuk nilai yang hilang dan tidak dapat diisi menggunakan nilai pada atribut lain akan dihapus

menggunakan teknik Listwise Deletion, yaitu menghapus sepenuhnya baris data jika terdapat data yang hilang pada satu kolom atau lebih dalam baris tersebut

3) *Transformasi Data:* Tahapan ketiga dalam proses preprocessing yaitu dilakukan transformasi data atau data transformation. Proses ini dilakukan untuk menyamakan jenis data menjadi data numerik untuk memudahkan dalam proses pengembangan model. Pengkategorian data dapat dilihat pada Tabel **I**.

TABEL I
PENGKATEGORIAN DATA

Atribut Data	Keterangan
Jalur Masuk	1 = SNMPTN 2 = SBMPTN 3 = MANDIRI
Jurusan SLTA	1 = SMA 2 = SMK 3 = MA
Pekerjaan Ayah	0 = Tidak Bekerja 1 = Bekerja
Pekerjaan Ibu	0 = Tidak Bekerja 1 = Bekerja
Pendapatan Ayah	1 = 500.000 2 = 1.000.000 3 = 1.500.000 4 = 2.000.000 5 = 3.000.000
Pendapatan Ibu	1 = 500.000 2 = 1.000.000 3 = 1.500.000 4 = 2.000.000 5 = 3.000.000
Status Mengundurkan Diri	0 = Tidak Mengundurkan Diri 1 = Mengundurkan Diri
SKS Tiap Semester dari Semester 2 - 4	1 = 0 ≤ Jumlah SKS tiap semester ≤ 12 2 = 13 ≤ Jumlah SKS tiap semester ≤ 15 3 = 16 ≤ Jumlah SKS tiap semester ≤ 18 4 = 19 ≤ Jumlah SKS tiap semester ≤ 21 5 = 22 ≤ Jumlah SKS tiap semester ≤ 25
IPS Tiap Semester dari Semester 1 - 4	1 = 0 ≤ Nilai IPS tiap semester ≤ 1.49 2 = 1.5 ≤ Nilai IPS tiap semester ≤ 1.99 3 = 2 ≤ Nilai IPS tiap semester ≤ 2.49 4 = 2.5 ≤ Nilai IPS tiap semester ≤ 2.99 5 = 3 ≤ Nilai IPS tiap semester ≤ 4

4) *Normalisasi Data:* Tahapan selanjutnya pada proses preprocessing yaitu normalisasi data. Normalisasi bertujuan untuk menyesuaikan skala sehingga nilai pada data berada dalam rentang yang sama sehingga dapat meningkatkan kinerja dan konvergensi algoritma. Pada penelitian ini, proses normalisasi dilakukan menggunakan dua metode sesuai dengan kebutuhan dan kecocokan pada proses pengembangan model, yaitu Min-Max Normalization dan Standard Scaler menggunakan persamaan berikut:

- *Min-max Normalization*

$$X_{norm} = X \quad max X^{i-} - X\,min_{Xmin} \qquad (1)$$

- *Standard Scaler*

$$X_{std} = X \quad i_{\sigma}{}^{-}\mu \qquad (2)$$

5) *Penyeimbangan Data*: Penyeimbangan data diperlukan untuk mengatasi ketidakseimbangan distribusi kelas dalam dataset, di mana satu kelas mungkin jauh lebih sedikit daripada kelas lainnya. Hal ini dapat menyebabkan bias model terhadap kelas mayoritas. Dalam penelitian ini, digunakan metode Synthetic Minority Over-sampling Technique (SMOTE) untuk menyeimbangkan data. SMOTE menghasilkan data sintetis baru untuk kelas minoritas melalui interpolasi, sehingga meningkatkan representasinya tanpa sekadar menggandakan data. Dengan demikian, model dapat mempelajari pola kedua kelas secara lebih seimbang, meningkatkan akurasi prediksi, terutama untuk kelas minoritas.

6) *Ekstraksi Fitur:* Ekstraksi fitur merupakan tahap yang bertujuan untuk mengurangi dimensi data dengan tetap mempertahankan informasi yang paling relevan. Proses ini membantu menyederhanakan kompleksitas data, mempercepat waktu komputasi, serta mengurangi risiko overfitting pada model. Dalam penelitian ini, metode ekstraksi fitur yang digunakan adalah Principal Component Analysis (PCA). PCA bekerja dengan mentransformasikan data asli ke dalam sistem koordinat baru berdasarkan komponen utama, yaitu arah dengan variasi data terbesar. Dengan demikian, PCA dapat mereduksi jumlah fitur tanpa mengorbankan informasi penting dalam data.

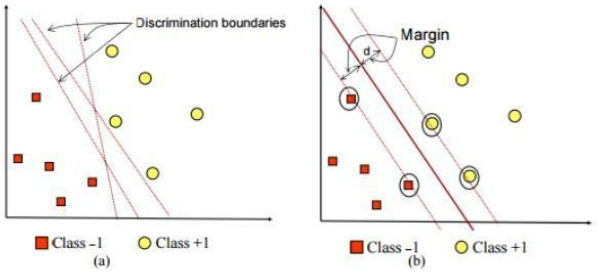
D. Pengembangan Model

Pengembangan model pada penelitian melewati beberapa tahapan:

1. *Pembagian Data: Dataset dibagi menjadi data training dan testing dengan variasi persentase (60:40, 70:30, 80:20, 90:10) untuk membantu evaluasi kinerja model secara objektif.*
2. *Pemisahan Atribut Identitas:* Atribut identitas merupakan atribut dengan nilai unik dalam dataset yang digunakan untuk mengidentifikasi entitas data sebagai referensi. Tujuan dari pemisahan atribut identitas adalah untuk memastikan bahwa hanya atribut yang relevan dengan prediksi yang digunakan dalam pengembangan model sehingga model tidak menyebabkan bias atau hasil yang kurang akurat.
3. *Pemisahan Fitur dan Target:* Fitur merupakan atribut – atribut yang digunakan sebagai variabel sebagai proses prediksi, sedangkan target merupakan variabel yang akan diprediksi. Fitur dan target dipisahkan dengan tujuan agar model dapat lebih fokus dalam mencari pola dari data yang tersedia sehingga dapat meningkatkan akurasi. Pada penelitian ini, fitur yang digunakan berasal dari dataset setelah preprocessing dengan menghilangkan atribut identitas.
4. *Hyperparameter Tuning:* Hyperparameter tuning adalah proses optimasi untuk menentukan kombinasi parameter terbaik guna meningkatkan kinerja model. Berbeda dengan parameter biasa, hyperparameter ditetapkan sebelum pelatihan dan tidak dihasilkan dari data. Proses ini krusial karena memengaruhi akurasi, kemampuan generalisasi, dan efisiensi model.

   Dalam penelitian ini, metode Grid Search digunakan untuk melakukan tuning. Teknik ini mengevaluasi secara sistematis semua kombinasi hyperparameter yang telah ditentukan dalam sebuah parameter grid. Setiap kombinasi diuji menggunakan validasi silang (cross-validation), dan kombinasi dengan performa terbaik dipilih sebagai konfigurasi akhir. Pada penelitian ini, parameter yang akan di tuning meliputi nilai C (cost), jenis kernel, nilai gamma, dan jumlah maksimum iterasi (max iteration).
5. *Pelatihan:* Tahapan pelatihan (training) bertujuan untuk melatih model agar mampu mengenali pola dalam data dengan mempelajari hubungan antara fitur dan target. Pada penelitian ini, metode utama yang dipilih dalam memodelkan hubungan antara fitur dan target adalah algoritma *Support Vector Machine.*

Support Vector Machine (SVM) merupakan metode data mining yang masuk dalam kelas supervised learning. Tujuan utama dari teknik ini adalah mencari hyperplane terbaik dari beberapa hyperplane yang ada. Adapun konsep hyperplane dapat dilihat pada Gambar 2.



Gambar 2. Konsep Hyperplan

Pada Gambar 2. *hyperplane* berada pada posisi ditengah antara dua kelas, yang artinya jarak antara *hyperplane* dengan data berbeda dengan kelas yang memiliki posisi terluar. Pada SVM, data yang berada diluar merupakan data yang paling dekat dengan *hyperplane* yang disebut dengan *support vector*. SVM pada dasarnya bekerja secara *linear*, dan dikembangkan untuk menyelesaikan masalah *non-linear*. *Hyperplane* terbaik didapatkan dengan cara memaksimalkan jarak antar kelas [8].

Dalam teknik SVM, *hyperplane* terbaik adalah *hyperplane* yang terletak di tengah-tengah antara dua set data dari dua kelas yang berbeda. Mencari *hyperplane* terbaik adalah dengan cara memaksimalkan margin. Margin adalah jarak tegak lurus antara *hyperplane* dengan objek terdekat. Objek terdekat disebut juga *support vector* [9]. Model *linear* secara umum yang digunakan dalam SVM untuk mencari *hyperplane* menggunakan persamaan berikut [10]:

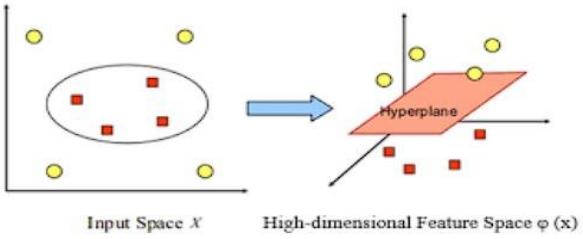
$$y = w^T x_i + b \quad , i = 1,2, \dots, n \tag{3}$$

Metode SVM membagi dataset menjadi dua kelas, kelas pertama akan bernilai +1 (positif) dan kelas kedua akan bernilai -1 (negatif). Dalam menentukan persamaan *hyperplane*, hal yang harus dilakukan terlebih dahulu yaitu menentukaan nilai *w* dan *b*. yang dimana nilai *w* merupakan nilai yang akan berkaitan dengan margin. Dalam menentukan nilai *w* dan b dapat menggunakan persamaan berikut:

$$w = \sum_{i=1}^{N} a_i y_i K(x, x_i) \tag{4}$$

$$b = -\frac{1}{2}(w.x^+ + w.x^-) \tag{5}$$

Pada beberapa kasus penelitian data tidak bisa diolah jika menggunakan klasifikasi metode linier SVM, sehingga untuk mengatasi masalah tersebut dikembangkan sebuah fungsi kernel yang bertujuan untuk menyelesaikan klasifikasi data secara *non-linear*. Pada *non-linear* SVM, data akan dipetakan kedalam sebuah fungsi ke ruang *vector* yang memiliki dimensi lebih tinggi. Pada ruang *vector* yang baru, *hyperplane* yang memisahkan kedua kelas tersebut akan dikonstruksikan [11].



Gambar 3. *Hyperplane*

Pada Gambar 3. proses input data x yang dimasukkan akan dipetakan pada *feature space* dengan menggunakan dimensi yang lebih tinggi dengan menggunakan fungsi kernel. Fungsi $\varphi(x)$ terkadang tidak bisa dilakukan perhitungan namun *dot product* pada kedua *vector* tersebut dapat dihitung dalam *input space* maupun *future space* menggunakan persamaan berikut [12]:

$$\phi(x_i).\phi(x_j) \quad (6)$$

Dalam Gambar 3. yang terletak disebalah kiri dapat dilihat bahwa data x yang berada di dalam *input space* tidak dapat dipisahkan secara *linear*. Sedangkan gambar yang terletak disebelah kanan fungsi φ berhasil memetakan data kedalam sebuah ruang *vector* yang memiliki dimensi lebih tinggi yaitu 3 dengan garis pemisah *hyperplane* yang linier [12]. Yang menggantikan perhitungan *dot product* pada persamaan (4) merupakan fungsi kernel. Adapun persamaan dari fungsi kernel menggunakan persamaan berikut:

$$K(x,x_j) \quad (7)$$

Fungsi kernel disinilah yang merupakan transformasi dari φ yang dimana proses ini dikenal dengan sebutan *Kernel Trick*. Dalam proses memperoleh *support vector* dengan menggunakn *kernel trick* ini hal tersebut akan lebih mudah. Umumnya terdapat empat jenis fungsi karnel yang dapat digunakan yaitu, kernel *linier,* kernel *polynomial,* kernel *gaussian (Radial Basis Function),* dan kernel *sigmoid* [13].

2) *Pengujian:* Untuk memperoleh performa terbaik dari model Support Vector Machine (SVM), dilakukan eksplorasi terhadap empat skenario pengujian berdasarkan kombinasi preprocessing dan strategi tuning parameter. Tujuan dari pendekatan ini adalah untuk mengetahui konfigurasi terbaik dalam membangun model yang mampu melakukan prediksi secara akurat. Adapun skenario yang akan digunakan disajikan pada Tabel **II.**

TABEL III
SKENARIO PENGUJIAN

Skenario	Transformasi	Normalisasi	Ekstraksi Fitur	*Hyperparameter Tuning*
1	Ya	*Min-Max Normalization*	Tidak	Tidak
2	Ya	*Min-Max Normalization*	Tidak	Ya
3	Ya	*Standard Scaler*	Ya	Ya
4	Ya	*Standard Scaler*	Tidak	Ya

E. Evaluasi

Tahapan ini dilakukan evaluasi model menggunakan *Mean Squared Error* (MSE), *Root Mean Square Error* (RMSE), dan *Confusion Matrix*.

Mean Square Error (MSE) merupakan salah satu evaluasi model yang digunakan untuk melihat nilai error pada model yang dibangun. MSE merupakan fungsi risiko yang berkaitan dengan nilai harapan dari kehilangan kuadrat kesalahan. MSE juga menilai kualitas prediktor dalam konteks prediksi, mengukur seberapa jauh nilai-nilai yang diprediksi oleh model dari nilai-nilai yang sebenarnya [14]. Untuk memperoleh nilai pengukuran *Mean Square Error* (MSE) dapat menggunakan persamaan berikut:

$$MSE = \frac{1}{2}\sum_{i=1}^{n}(Y_i - \hat{y}_i)^2$$

(8)

Root Mean Square Error (RMSE) merupakan salah satu cara evaluasi model berdasarkan pada error estimasi. Semakin kecil nilai *error* yang dihasilkan maka hasil prediksi dari model tersebut semakin akurat. Untuk memperoleh nilai pengukuran *Root Mean Square Error* (RMSE) dapat menggunakan persamaan berikut:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

(9)

Confusion matrix adalah tabel yang digunakan untuk melakukan pengujian performa pada model klasifikasi yang telah dibuat. Dalam mengukur tingkat ketepatan prediksi adalah dengan mengukur akurasi, sensitivitas, dan spesifikasi menggunakan *confusion matrix* menggunakan *true positive* (TP), *false positive* (FP), *false negative* (FN), dan *true negative* (TN). *Confusion matrix* akan dipresentasikan pada Tabel III.

TABEL III
CONFUSION MATRIX

		True Class	
		Positive	***Negative***
Predicted Class	***Positive***	*True Positive*	*False Positive*
	Negative	*False Negative*	*True Negative*

Pada tahapan ini, nilai – nilai MSE, RMSE, dan akurasi dari *Confusion Matrix* yang didapat dari tahap pengujian akan dibandingkan untuk mendapatkan nilai akurasi terbaik yang nantinya akan digunakan dalam tahapan prediksi.

F. Prediksi

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang. Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan untuk prediksi diantaranya adalah yaitu pembangunan model, penerapan model, dan evaluasi [15].

Model dengan hasil akurasi terbaik yang didapat pada tahap evaluasi akan digunakan untuk melakukan prediksi mahasiswa yang berpotensi mengundurkan diri. Data yang digunakan pada tahapan ini merupakan data mahasiswa angkatan 2021 pada Fakultas Teknik Universitas Tanjungpura. Hasil prediksi akan dibandingkan dengan data aktual untuk mengetahui performa algoritma SVM dalam memprediksi mahasiswa yang mengundurkan diri dan menganalisis hasil prediksi tersebut.

III. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Dataset didapat dari Biro Akademik dan Kemahasiswaan (BAK) Universitas Tanjungpura. Data yang didapat merupakan data mahasiswa aktif dan data mahasiswa yang telah melakukan

pengunduran diri dari angkatan 2014 hingga angkatan 2021 di Fakultas Teknik Universitas Tanjungpura. Data yang akan digunakan adalah data mahasiswa angkatan 2014 hingga angkatan 2020 sebagai data untuk proses *training* dan *testing* pada model yang akan dibangun, sedangkan data untuk melakukan prediksi mahasiswa yang berpotensi mengundurkan diri yaitu menggunakan data mahasiswa angkatan 2021 dan nantinya akan dilakukan evaluasi hasil prediksi tersebut dengan data aktualnya. Terdapat total 5340 jumlah mahasiswa dengan jumlah mahasiswa yang tidak mengundurkan diri berjumlah 4000 mahasiswa dan jumlah mahasiswa yang mengundurkan diri berjumlah 1340 mahasiswa. Seluruh mahasiswa yang mengundurkan diri pada setiap program studi disajikan pada Tabel IV.

TABEL IV
DATA MAHASISWA MENGUNDURKAN DIRI BERDASARKAN PROGRAM STUDI

No	Program Studi	Jumlah Mahasiswa Mengundurkan Diri
1.	Teknik Sipil	421
2.	Teknik Elektro	259
3.	Informatika	157
4.	Arsitektur	80
5.	Teknik Kelautan	80
6.	Teknik Mesin	78
7.	Teknik Lingkungan	70
8.	Teknik Industri	60
9.	Teknik Pertambangan	58
10.	Teknik Kimia	43
11.	Perencanaan Wilayah dan Kota	34

B. Preprocessing

1) *Pembersihan Data:* Dataset yang didapat dari BAK Universitas Tanjungpura memiliki 32 atribut data Dalam penelitian ini atribut yang digunakan hanya 17 atribut diantaranya yaitu ID, jalur masuk, tahun akademik, jurusan SLTA, pekerjaan ayah, pekerjaan ibu, pendapatan ayah, pendapatan ibu, status keluar, sks semester 2, sks semester 3, sks semester 4, ips semester 1, ips semester 2, ips semester 3, ips semester 4, dan nama prodi.

Untuk mengatasi data yang hilang pada dataset, dilakukan proses *Logical Imputation*, yaitu mengisi nilai yang hilang pada suatu atribut data menggunakan nilai pada atribut data lainnya. Berikut adalah contoh sampel data yang akan diterapkan proses *logical imputation.*

	ID	Jalur Masuk	Tahun Akademik	Jurusan SLTA	Nama SLTA	Pekerjaan Ayah	Pekerjaan Ibu	Pendapatan Ayah	Pendapatan Ibu
0	97093	MANDIRI	2014	SMA/MA IPA	SMAN 1 BUNUT HILIR	Petani/Nelayan	Belum Bekerja	500000	500000
1	46334	MANDIRI	2014	SMA/MA IPA	SMAN 1 SANGGAU	Wiraswasta	Lainnya	2000000	500000
2	24529	MANDIRI	2014	NaN	SMKN 1 SEKADAU	Wiraswasta	Lainnya	2000000	500000
3	25403	MANDIRI	2014	SMA/MA IPA	SMKN 1 SINTANG	Almarhum (Sudah Meninggal)	Lainnya	500000	500000
4	56870	MANDIRI	2014	NaN	SMKN 8 KOTA PONTIANAK	Almarhum (Sudah Meninggal)	Lainnya	500000	500000

Gambar 4. Sampel Data Sebelum Proses *Logical Imputation*

Sampel data dengan atribut Jurusan SLTA mengandung *missing value*. Sedangkan, atribut disebelahnya yaitu Nama SLTA memiliki informasi yang dapat digunakan untuk mengisi atribut Jurusan SLTA yang hilang.

	ID	Jalur Masuk	Tahun Akademik	Jurusan SLTA	Nama SLTA	Pekerjaan Ayah	Pekerjaan Ibu	Pendapatan Ayah	Pendapatan Ibu
0	97093	MANDIRI	2014	SMA/MA IPA	SMAN 1 BUNUT HILIR	Petani/Nelayan	Belum Bekerja	500000	500000
1	46334	MANDIRI	2014	SMA/MA IPA	SMAN 1 SANGGAU	Wiraswasta	Lainnya	2000000	500000
2	24529	MANDIRI	2014	SMK	SMKN 1 SEKADAU	Wiraswasta	Lainnya	2000000	500000
3	25403	MANDIRI	2014	SMA/MA IPA	SMKN 1 SINTANG	Almarhum (Sudah Meninggal)	Lainnya	500000	500000
4	56870	MANDIRI	2014	SMK	SMKN 8 KOTA PONTIANAK	Almarhum (Sudah Meninggal)	Lainnya	500000	500000

Gambar 5. Sampel Data Seteleh Proses *Logical Imputation*

Proses *logical imputation* ini juga diterapkan pada atribut data lainnya yang masih memiliki data yang hilang seperti mengisi nilai IPS menggunakan jumlah SKS, mengisi Status Keluar, dan mengisi atribut SKS Semester dengan nilai pada status mahasiswa mengundurkan diri. Persentase data yang hilang setelah proses *logical imputation* dipresentasikan pada Gambar 6.

```
ID                  0.000000
Jalur Masuk         0.000000
Tahun Akademik      0.000000
Jurusan SLTA        1.011236
Pekerjaan Ayah      0.823970
Pekerjaan Ibu       0.842697
Pendapatan Ayah     0.000000
Pendapatan Ibu      0.000000
Status Keluar       0.000000
SKS Semester 2      0.000000
SKS Semester 3      0.000000
SKS Semester 4      0.000000
IPS Semester 1      0.000000
IPS Semester 2      0.000000
IPS Semester 3      0.000000
IPS Semester 4      0.000000
Nama Prodi         17.940075
dtype: float64
```

Gambar 6. Persentase *Missing Value*

Setelah melewati proses *Logical Imputation*, masih terdapat beberapa atribut data yang memiliki *missing value*. Nilai yang hilang dan tidak dapat diisi menggunakan atribut lain akan dihapus menggunakan teknik *Listwise Deletion*, yaitu baris data akan langsung dihapus jika terdapat atribut data yang hilang pada baris tersebut. Jumlah data yang hilang setelah dilakukan proses *listwise deletion* dipresentasikan pada Gambar 7.

```
ID                  0
Jalur Masuk         0
Tahun Akademik      0
Jurusan SLTA        0
Pekerjaan Ayah      0
Pekerjaan Ibu       0
Pendapatan Ayah     0
Pendapatan Ibu      0
Status Keluar       0
SKS Semester 2      0
SKS Semester 3      0
SKS Semester 4      0
IPS Semester 1      0
IPS Semester 2      0
IPS Semester 3      0
IPS Semester 4      0
Nama Prodi        957
dtype: int64
```

Gambar 7. Jumlah *Missing Value*

Setelah dilakukan proses pembersihan data, terdapat perubahan jumlah mahasiswa yang tidak mengundurkan diri dan mahasiswa yang mengundurkan diri. Total keseluruhan data mahasiswa berubah menjadi 5243 dari yang sebelumnya 5340 mahasiswa dengan mahasiswa yang tidak mengundurkan diri berjumlah 3949 dari yang sebelumnya 4000 mahasiswa dan mahasiswa yang mengundurkan diri berjumlah 1294 dari yang sebelumnya 1340 mahasiswa.

TABEL V
DATA MAHASISWA MENGUNDURKAN DIRI BERDASARKAN PROGRAM STUDI SETELAH PEMBERSIHAN DATA

No	Program Studi	Jumlah Mahasiswa Mengundurkan Diri
1.	Teknik Sipil	407
2.	Teknik Elektro	244
3.	Informatika	152
4.	Arsitektur	80
5.	Teknik Kelautan	76
6.	Teknik Mesin	72
7.	Teknik Lingkungan	69
8.	Teknik Industri	60
9.	Teknik Pertambangan	58
10.	Teknik Kimia	43
11.	Perencanaan Wilayah dan Kota	33

2) *Transformasi Data:* Selanjutnya akan dilakukan proses transformasi data dengan tujuan untuk menyamakan jenis data dan mengubah menjadi bentuk numerik.

	ID	Jalur Masuk	Tahun Akademik	Jurusan SLTA	Pekerjaan Ayah	Pekerjaan Ibu	Pendapatan Ayah	Pendapatan Ibu	Status Keluar	SKS Semester 2	SKS Semester 3	SKS Semester 4	IPS Semester 1	IPS Semester 2	IPS Semester 3
0	97093	MANDIRI	2014	SMA	Petani/Nelayan	Belum Bekerja	500000	500000	1.0	21	18.0	21.0	2.76	2.57	2.50
1	46334	MANDIRI	2014	SMA	Wiraswasta	Lainnya	2000000	500000	4.0	0	0.0	0.0	0.00	0.00	0.00
2	24529	MANDIRI	2014	SMK	Wiraswasta	Lainnya	2000000	500000	1.0	24	21.0	21.0	3.38	3.46	3.33
3	25403	MANDIRI	2014	SMK	Almarhum (Sudah Meninggal)	Lainnya	500000	500000	1.0	21	24.0	21.0	2.90	3.48	2.83
4	56870	MANDIRI	2014	SMK	Almarhum (Sudah Meninggal)	Lainnya	500000	500000	1.0	24	24.0	24.0	3.48	3.13	3.08
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5238	49939	SNMPTN	2020	SMA	Petani/Nelayan	Ibu Rumah Tangga	1500000	1000000	1.0	21	23.0	23.0	3.55	3.83	3.11
5239	84474	SNMPTN	2020	SMK	Buruh	Tidak bekerja	1000000	500000	1.0	23	22.0	19.0	3.03	3.54	3.02
5240	81993	SNMPTN	2020	SMK	PNS Lembaga Pemerintah	Ibu Rumah Tangga	3000000	1000000	1.0	23	0.0	0.0	3.10	2.11	0.00
5241	74964	SNMPTN	2020	SMK	Wiraswasta	Ibu Rumah Tangga	1500000	500000	1.0	23	22.0	19.0	3.40	3.59	2.95
5242	84044	SNMPTN	2020	SMA	Wiraswasta	Tidak bekerja	3000000	500000	1.0	23	22.0	19.0	3.48	3.52	3.23

5243 rows × 17 columns

Gambar 8. *Dataframe* Sebelum Transformasi Data

Setelah dilakukan proses transformasi data, sampel data tersebut telah berubah menjadi bentuk numerik.

	ID	Jalur Masuk	Tahun Akademik	Jurusan SLTA	Pekerjaan Ayah	Pekerjaan Ibu	Pendapatan Ayah	Pendapatan Ibu	Status Keluar	SKS Semester 2	SKS Semester 3
0	97093	3.0	2014	1.0	1.0	0.0	1.0	1.0	0.0	4.0	3.0
1	46334	3.0	2014	1.0	1.0	1.0	4.0	1.0	1.0	1.0	1.0
2	24529	3.0	2014	2.0	1.0	1.0	4.0	1.0	0.0	5.0	4.0
3	25403	3.0	2014	2.0	0.0	1.0	1.0	1.0	0.0	4.0	5.0
4	56870	3.0	2014	2.0	0.0	1.0	1.0	1.0	0.0	5.0	5.0
...	...	...	...	...	...	...	...	...	...	...	...
5238	49839	1.0	2020	1.0	1.0	0.0	3.0	2.0	0.0	4.0	5.0
5239	34474	1.0	2020	2.0	1.0	1.0	2.0	1.0	0.0	5.0	5.0
5240	81993	1.0	2020	2.0	1.0	0.0	5.0	2.0	0.0	5.0	1.0
5241	74964	1.0	2020	2.0	1.0	0.0	3.0	1.0	0.0	5.0	5.0
5242	84044	1.0	2020	1.0	1.0	1.0	5.0	1.0	0.0	5.0	5.0

Gambar 9. *Dataframe* Setelah Transformasi Data

3) Normalisasi Data

Normalisasi data pada penelitian ini akan menggunakan metode *Min-max normalization* dan metode *Standard Scaler* sesuai dengan skenario yang akan digunakan. Pada metode *Min-max normalization*, nilai skala pada dataset akan disamaratakan dengan rentang nilai 0 sampai dengan 1 menggunakan persamaan (1). Dataset yang telah dilakukan proses normalisasi menggunakan *Min-max Normalization* dipresentasikan pada Gambar 10.

	ID	Jalur Masuk	Tahun Akademik	Jurusan SLTA	Pekerjaan Ayah	Pekerjaan Ibu	Pendapatan Ayah	Pendapatan Ibu	Status Keluar	SKS Semester 2	SKS Semester 3
0	97093	1.0	2014	0.0	1.0	0.0	0.00	0.00	0.0	0.75	0.50
1	46334	1.0	2014	0.0	1.0	1.0	0.75	0.00	1.0	0.00	0.00
2	24529	1.0	2014	0.5	1.0	1.0	0.75	0.00	0.0	1.00	0.75
3	25403	1.0	2014	0.5	0.0	1.0	0.00	0.00	0.0	0.75	1.00
4	56870	1.0	2014	0.5	0.0	1.0	0.00	0.00	0.0	1.00	1.00
...	...	...	...	...	...	...	...	...	...	...	...
5238	49839	0.0	2020	0.0	1.0	0.0	0.50	0.25	0.0	0.75	1.00
5239	34474	0.0	2020	0.5	1.0	1.0	0.25	0.00	0.0	1.00	1.00
5240	81993	0.0	2020	0.5	1.0	0.0	1.00	0.25	0.0	1.00	0.00
5241	74964	0.0	2020	0.5	1.0	0.0	0.50	0.00	0.0	1.00	1.00
5242	84044	0.0	2020	0.0	1.0	1.0	1.00	0.00	0.0	1.00	1.00

Gambar 10. *Dataframe* Setelah Normalisasi Menggunakan *Min-max Normalization*

Sedangkan pada metode *Standard Scaler*, data dinormalisasi dengan cara mengubah distribusi nilai fitur sehingga memiliki rata-rata *(mean)* sebesar 0 dan standar deviasi sebesar

1. Proses ini dilakukan dengan menggunakan rumus pada persamaan (2). Dataset yang telah melalui proses normalisasi menggunakan *Standard Scaler* disajikan pada Gambar 11.

	Jalur Masuk	Jurusan SLTA	Pekerjaan Ayah	Pekerjaan Ibu	Pendapatan Ayah	Pendapatan Ibu	SKS Semester 2	SKS Semester 3
0	0.988059	-0.515186	0.269086	-0.940540	-1.749494	-0.633983	0.213109	-0.329575
1	0.988059	-0.515186	0.269086	1.063219	0.236060	-0.633983	-1.999425	-1.575853
2	0.988059	1.235377	0.269086	1.063219	0.236060	-0.633983	0.950620	0.293563
3	0.988059	1.235377	-3.716280	1.063219	-1.749494	-0.633983	0.213109	0.916702
4	0.988059	1.235377	-3.716280	1.063219	-1.749494	-0.633983	0.950620	0.916702
...	...	...	...	...	...	...	...	...
5238	-1.507313	-0.515186	0.269086	-0.940540	-0.425791	-0.024974	0.213109	0.916702
5239	-1.507313	1.235377	0.269086	1.063219	-1.087643	-0.633983	0.950620	0.916702
5240	-1.507313	1.235377	0.269086	-0.940540	0.897911	-0.024974	0.950620	-1.575853
5241	-1.507313	1.235377	0.269086	-0.940540	-0.425791	-0.633983	0.950620	0.916702
5242	-1.507313	-0.515186	0.269086	1.063219	0.897911	-0.633983	0.950620	0.916702

Gambar 11. *Dataframe* Setelah Normalisasi Menggunakan *Standard Scaler*

4) *Penyeimbangan Data:* Dalam penelitian ini dilakukan penyeimbangan data menggunakan metode Synthetic Minority Over-sampling Technique (SMOTE). SMOTE dilakukan secara khusus pada data pelatihan (training set) setelah proses pembagian data (train-test split), untuk menghindari kebocoran data (data leakage) dan menjaga validitas pengujian model. Implementasi SMOTE terintegrasi dalam pipeline bersama proses pelatihan model Support Vector Machine (SVM).

5) *Ekstraksi Fitur:* Pada penelitian ini, proses ekstraksi fitur dilakukan menggunakan metode Principal Component Analysis (PCA). Metode PCA digunakan untuk menyederhanakan kompleksitas data dengan mengubah sejumlah fitur awal menjadi beberapa komponen utama yang tetap mempertahankan sebagian besar varians dari data asli. Sampel data setelah dilakukan proses ekstraksi fitur dipresentasikan pada Gambar 12.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
0	0.100614	-0.898740	1.172487	0.151324	-1.171471	-0.174552	-1.229329	-0.672734	0.458747
1	4.766466	0.327678	-0.603211	-0.209866	-0.650935	-0.740710	-0.419004	0.386734	-1.187597
2	-1.589216	0.043451	-0.040123	1.757663	0.290773	0.269926	-0.437820	-0.051164	-1.197644
3	-1.007755	1.524137	3.791793	1.259463	-0.728511	0.057271	0.955183	0.915178	-1.138650
4	-2.038722	1.496850	3.795055	1.347460	-0.744387	0.456607	0.898657	0.287129	-1.345040
...	...	...	...	...	...	...	...	...	...
5238	-1.917063	-0.666123	0.303196	-1.195205	0.556963	-0.308467	-0.064599	0.048902	0.880194
5239	-2.048186	0.203632	1.066706	0.041113	2.002279	0.014110	-0.991838	0.236072	-0.546567
5240	1.984250	-0.903379	-0.462364	-0.526131	1.910271	2.854689	0.466733	-0.317077	0.411181
5241	-1.790343	-1.214550	0.839999	-0.052250	1.694495	0.442296	-0.020071	0.239139	0.576747
5242	-2.233921	0.111161	-0.740327	-1.083565	0.911898	0.178086	-0.083692	0.598880	-1.056269

Gambar 12. *Dataframe* Setelah Esktraksi Fitur Menggunakan PCA

C. Pengembangan Model

1) Pembagian data: Dataset dibagi menjadi data training dan data testing yang dipresentasikan pada Tabel VI.

TABEL VI
PEMBAGIAN DATA TRAINING DAN DATA TESTING

Dataset	Data *Training* (%)	Jumlah Data *Training*	Data *Testing* (%)	Jumlah Data *Testing*
1	60%	3145	40%	2098
2	70%	3670	30%	1573
3	80%	4194	20%	1049
4	90%	4718	10%	525

2) *Pemisahan Atribut Identitas:* Atribut data akan dipisahkan menjadi atribut identitas dan atribut prediktif. terdapat 3 atribut yang dikategorikan sebagai atribut identitas, yaitu ID, Tahun Akademik, dan Nama Prodi. Ketiga atribut tersebut tidak digunakan dalam proses pemodelan dan hanya berfungsi sebagai penanda atau informasi administratif. Sementara itu, sebanyak 14 atribut lainnya dikategorikan sebagai atribut prediktif.

3) *Pemisahan Fitur Target:* Fitur yang digunakan yaitu Jurusan SLTA, Pekerjaan Ayah, Pekerjaan Ibu, Pendapatan Ayah, Pendapatan Ibu, SKS Semester 2, SKS Semester 3, SKS Semester 4, IPS Semester 1, IPS Semester 2, IPS Semester 3, dan IPS Semester 4. Sedangkan target yang digunakan yaitu Status Keluar.

D. Evaluasi

Pada skenario pertama, proses diawali dengan tahap transformasi data, di mana data dikonversi ke dalam bentuk numerik. Selanjutnya, dilakukan proses normalisasi menggunakan metode *Min-Max Normalization*, yang bertujuan untuk menyamakan skala setiap atribut ke dalam rentang nilai antara 0 hingga 1. Uji coba menggunakan algoritma SVM dilakukan dengan 5243 sampe data training yang telah ditentukan komposisinya seperti yang tertera pada Tabel VI. Hasil evaluasi scenario 1 sesuai komposisi ditampilkan pada Tabel dibawah. Performa model dengan akurasi paling tinggi dipresentasikan dengan baris berwarna hijau, sedangkan model dengan akurasi paling rendah dipresentasikan dengan baris berwarna merah.

TABEL VII
HASIL EVALUASI SKENARIO 1 KOMPOSISI
60% : 40%

Ke-	Parameter				Akurasi		MSE	RMSE
	Kernel	C	Gamma	Iterasi	Training	Testing		
1	RBF	0.1	0.01	50	0.86900	0.86749	0.13251	0.36402
2	RBF	0.1	0.01	200	0.74213	0.73499	0.26501	0.51480
3	RBF	0.1	0.1	50	0.82289	0.80696	0.19304	0.43936
4	RBF	0.1	0.1	200	0.83370	0.83746	0.16254	0.40316
5	RBF	0.1	0.5	50	0.89285	0.89609	0.10391	0.32235
6	RBF	0.1	0.5	200	0.90715	0.90372	0.09628	0.31029
7	RBF	0.01	0.01	50	0.86900	0.86749	0.13251	0.36402
8	RBF	0.01	0.01	200	0.74213	0.73499	0.26501	0.51480
9	RBF	0.01	0.1	50	0.82289	0.80696	0.19304	0.43936
10	RBF	0.01	0.1	200	0.82989	0.83556	0.16444	0.40551
11	RBF	0.01	0.5	50	0.89285	0.89609	0.10391	0.32235
12	RBF	0.01	0.5	200	0.56248	0.56196	0.43804	0.66184

TABEL VIII
HASIL EVALUASI SKENARIO 1 KOMPOSISI
70% : 30%

Ke-	Parameter				Akurasi		MSE	RMSE
	Kernel	C	Gamma	Iterasi	Training	Testing		
1	RBF	0.1	0.01	50	0.82752	0.82899	0.17101	0.41353
2	RBF	0.1	0.01	200	0.44523	0.43357	0.56643	0.75262
3	RBF	0.1	0.1	50	0.88883	0.89765	0.10235	0.31993
4	RBF	0.1	0.1	200	0.67956	0.66306	0.33694	0.58046
5	RBF	0.1	0.5	50	0.72234	0.72600	0.27400	0.52345
6	RBF	0.1	0.5	200	0.80082	0.82009	0.17991	0.42416
7	RBF	0.01	0.01	50	0.82752	0.82899	0.17101	0.41353
8	RBF	0.01	0.01	200	0.44523	0.43357	0.56643	0.75262
9	RBF	0.01	0.1	50	0.88883	0.89765	0.10235	0.31993
10	RBF	0.01	0.1	200	0.87384	0.88048	0.11952	0.34571
11	RBF	0.01	0.5	50	0.72234	0.72600	0.27400	0.52345
12	RBF	0.01	0.5	200	0.61117	0.62428	0.37572	0.61296

TABEL IX
HASIL EVALUASI SKENARIO 1 KOMPOSISI
80% : 20%

Ke-	Parameter				Akurasi		MSE	RMSE
	Kernel	C	Gamma	Iterasi	Training	Testing		
1	RBF	0.1	0.01	50	0.86695	0.87131	0.12869	0.35874
2	RBF	0.1	0.01	200	0.55722	0.56149	0.43851	0.66220
3	RBF	0.1	0.1	50	0.78040	0.78170	0.21830	0.46723
4	RBF	0.1	0.1	200	0.85026	0.84557	0.15443	0.39298
5	RBF	0.1	0.5	50	0.87220	0.86368	0.13632	0.36922
6	RBF	0.1	0.5	200	0.78684	0.79218	0.20782	0.45587
7	RBF	0.01	0.01	50	0.86695	0.87131	0.12869	0.35874
8	RBF	0.01	0.01	200	0.55722	0.56149	0.43851	0.66220
9	RBF	0.01	0.1	50	0.78040	0.78170	0.21830	0.46723
10	RBF	0.01	0.1	200	0.84359	0.84271	0.15729	0.39660
11	RBF	0.01	0.5	50	0.87220	0.86368	0.13632	0.36922
12	RBF	0.01	0.5	200	0.65594	0.63775	0.36225	0.60187

TABEL X
HASIL EVALUASI SKENARIO 1 KOMPOSISI
90% : 10%

Ke-	Parameter				Akurasi		MSE	RMSE
	Kernel	C	Gamma	Iterasi	Training	Testing		
1	RBF	0.1	0.01	50	0.85312	0.84762	0.15238	0.39036
2	RBF	0.1	0.01	200	0.68398	0.67810	0.32190	0.56737
3	RBF	0.1	0.1	50	0.87919	0.88000	0.12000	0.34641
4	RBF	0.1	0.1	200	0.81496	0.80000	0.20000	0.44721
5	RBF	0.1	0.5	50	0.85608	0.84762	0.15238	0.39036
6	RBF	0.1	0.5	200	0.81306	0.82476	0.17524	0.41861
7	RBF	0.01	0.01	50	0.85312	0.84762	0.15238	0.39036
8	RBF	0.01	0.01	200	0.68398	0.67810	0.32190	0.56737
9	RBF	0.01	0.1	50	0.87919	0.88000	0.12000	0.34641
10	RBF	0.01	0.1	200	0.81496	0.80000	0.20000	0.44721
11	RBF	0.01	0.5	50	0.85608	0.84762	0.15238	0.39036
12	RBF	0.01	0.5	200	0.81878	0.82095	0.17905	0.42314

Pada skenario kedua dilakukan proses *hyperparameter tuning* terhadap beberapa nilai parameter seperti *C*, *gamma*, dan iterasi untuk menemukan kombinasi parameter terbaik dalam menghasilkan model dengan performa paling optimal.

TABEL XI
HASIL EVALUASI SKENARIO 2

Pembagian Dataset	Parameter	Jumlah	Hasil Evaluasi Pengujian			
			Training	*Testing*	*MSE*	*RMSE*
60% *Training* 40% *Testing*	Kernel	RBF	0.93672	0.93548	0.07317	0.27047
	C	1				
	Gamma	0.5				
	Iterasi	1000				
70% *Training* 30% *Testing*	Kernel	RBF	0.93198	0.92934	0.07829	0.27978
	C	1				
	Gamma	0.1				
	Iterasi	500				
80% *Training* 20% *Testing*	Kernel	RBF	0.93579	0.93352	0.08398	0.28980
	C	1				
	Gamma	0.5				
	Iterasi	1500				
90% *Training* 10% *Testing*	Kernel	RBF	0.92829	0.92551	0.08201	0.28639
	C	0.1				
	Gamma	0.1				
	Iterasi	500				

Pada skenario ketiga, proses diawali dengan tahap transformasi data, yaitu data diubah menjadi bentuk numerik. Selanjutnya, dilakukan normalisasi menggunakan metode *Standard Scaler* sehingga memiliki distribusi dengan rata-rata 0 dan standar deviasi 1. Selanjutnya, dilakukan tahap ekstraksi fitur menggunakan metode *Principal Component Analysis* (PCA). Hasil Evaluasi Skenario 3 ditampilkan pada Tabel XII.

TABEL XII
HASIL EVALUASI SKENARIO 3

Pembagian Dataset	Parameter	Jumlah	Hasil Evaluasi Pengujian			
			Training	*Testing*	*MSE*	*RMSE*
60% *Training* 40% *Testing*	Kernel	RBF	0.93439	0.93205	0.07693	0.27743
	C	1				
	Gamma	0.1				
	Iterasi	1000				
70% *Training* 30% *Testing*	Kernel	RBF	0.93164	0.92913	0.07751	0.27848
	C	10				
	Gamma	0.01				
	Iterasi	1000				
80% *Training* 20% *Testing*	Kernel	RBF	0.93322	0.93101	0.08398	0.28980
	C	1				
	Gamma	0.1				
	Iterasi	500				
90% *Training* 10% *Testing*	Kernel	RBF	0.93418	0.93132	0.07862	0.28034
	C	1				
	Gamma	0.1				
	Iterasi	1000				

Pada skenario keempat, proses diawali dengan tahap transformasi data, yaitu data diubah menjadi bentuk numerik. Selanjutnya, dilakukan normalisasi menggunakan metode Standard Scaler sehingga memiliki distribusi dengan ratarata 0 dan standar deviasi 1. Pada skenario ini

dilakukan proses hyperparameter tuning terhadap beberapa nilai parameter seperti C, gamma, dan iterasi. Hasil Evaluasi Skenario 4 disajikan dalam Tabel XIII.

TABEL XIII
HASIL EVALUASI SKENARIO 3

Pembagian Dataset	Parameter	Jumlah	Hasil Evaluasi Pengujian			
			Training	*Testing*	*MSE*	*RMSE*
60% *Training* 40% *Testing*	Kernel	RBF	0.93122	0.92879	0.07546	0.07546
	C	1				
	Gamma	0.1				
	Iterasi	1000				
70% *Training* 30% *Testing*	Kernel	RBF	0.92576	0.92362	0.08229	0.08229
	C	10				
	Gamma	0.01				
	Iterasi	1000				
80% *Training* 20% *Testing*	Kernel	RBF	0.92879	0.92633	0.08158	0.08158
	C	1				
	Gamma	0.1				
	Iterasi	500				
90% *Training* 10% *Testing*	Kernel	RBF	0.92729	0.92448	0.08201	0.08201
	C	1				
	Gamma	0.1				
	Iterasi	1000				

Setelah proses pengujian model *Support Vector Machine* telah selesai dilakukan, diperoleh hasil yang berbeda pada tiap scenario. Berdasarkan hasil tersebut, skenario 2 memberikan performa terbaik dengan akurasi training sebesar 0.93672 dan akurasi testing tertinggi sebesar 0.93548. Nilai MSE dan RMSE pada skenario ini juga merupakan yang terendah, masing-masing 0.07317 dan 0.27047, yang menunjukkan bahwa model memiliki kesalahan prediksi yang paling kecil dibandingkan skenario lainnya. Hasil evaluasi terbaik yang akan digunakan untuk prediksi mahasiswa berpotensi mengundurkan diri, yaitu menggunakan skenario 2 dengan parameter kernel RBF, C dengan nilai 1, gamma dengan nilai 0.5, dan iterasi sebanyak 1000.

E. Prediksi

Dataframe mahasiswa angkatan 2021 yang digunakan untuk prediksi. *Dataframe* tersebut dilakukan prediksi menggunakan menggunakan skenario 2 dengan parameter kernel RBF, *cost* sebesar 1, *gamma* sebesar 0.5, dan iterasi sebanyak 1000. *Dataframe* hasil prediksi dipresentasikan pada Gambar 13.

	ID	Jalur Masuk	Tahun Akademik	Jurusan SLTA	Prediksi Status Keluar	Pekerjaan Ayah	Pekerjaan Ibu	Pendapatan Ayah	Pendapatan Ibu	Status Keluar	SKS Semester 2
0	29868	1.0	2021	1.0	1.0	1	1	1.00	1.00	NaN	1.0
1	84495	1.0	2021	0.0	1.0	1	0	1.00	0.00	NaN	0.0
2	14971	1.0	2021	0.0	1.0	1	1	0.75	0.25	NaN	1.0
3	16433	1.0	2021	0.0	0.0	1	0	1.00	0.00	NaN	1.0
4	93426	1.0	2021	0.0	1.0	0	1	0.00	1.00	NaN	1.0
...	...	...	...	...	...	...	...	...	...	...	...
718	92984	0.0	2021	0.0	0.0	1	0	1.00	0.00	NaN	1.0
719	89276	0.0	2021	0.5	0.0	1	0	0.00	0.00	NaN	1.0
720	29233	0.0	2021	0.5	0.0	1	0	0.25	0.00	NaN	1.0
721	57434	0.0	2021	0.5	0.0	1	0	0.25	0.00	NaN	1.0
722	78529	0.0	2021	0.5	0.0	1	0	1.00	0.25	NaN	1.0

Gambar 13. *Dataframe* Hasil Prediksi

Hasil prediksi mahasiswa angkatan 2021 yang berpotensi mengundurkan diri. Prediksi dilakukan menggunakan 723 sampel data dan diperoleh bahwa 630 mahasiswa diprediksi tidak mengundurkan diri dan 93 mahasiswa diprediksi mengundurkan diri.

Evaluasi prediksi dilakukan dengan membandingkan hasil prediksi dan data aktual mahasiswa angkatan 2021 untuk menganalisis hasil prediksi algoritma SVM terhadap mahasiswa yang diprediksi berpotensi mengundurkan diri. Rekapitulasi data hasil prediksi dan data aktual dipresentasikan pada tabel berikut.

TABEL XIV
REKAPITULASI DATA AKTUAL DAN PREDIKSI MAHASISWA MENGUNDURKAN DIRI

Kategori	Tidak Mengundurkan Diri	Mengundurkan Diri
Data Aktual	640	83
Hasil Prediksi	630	93
Benar	621	74
Salah	19	9

Dengan demikian, hasil prediksi tersebut dapat direpresentasikan dalam bentuk *confusion matrix* seperti yang ditunjukkan pada Tabel XV.

TABEL XV
CONFUSION MATRIX HASIL PREDIKSI

		True Class	
		Positive	***Negative***
Predicted Class	***Positive***	74	19
	Negative	9	621

Berdasarkan tabel tersebut, dapat diperoleh nilai-nilai berikut:
- *True Positive* = 74
- *True Negative* = 621
- *False Positive* = 19
- *False Negative* = 9

Berdasarkan *confusion matrix* tersebut, diperoleh nilai akurasi sebagai berikut:

$$Akurasi = \frac{74 + 621}{74 + 621 + 19 + 9} x\ 100\% = \frac{695}{723} x\ 100\% = 96.13\%$$

Dengan demikian, dapat disimpulkan bahwa tingkat akurasi dari model dalam mengklasifikasikan data mahasiswa yang mengundurkan diri dan yang tidak adalah sebesar 96.13%.

IV. KESIMPULAN

1. Berdasarkan hasil evaluasi, algoritma Support Vector Machine (SVM) memiliki performa yang baik dalam memprediksi mahasiswa berpotensi mengundurkan diri dengan hasil akurasi data testing sebesar 93,54%, nilai MSE sebesar 0. 07317, dan RMSE sebesar 0.27047.
2. Data aktual dan hasil prediksi mahasiswa mengundurkan diri menunjukkan tingkat akurasi tinggi, yaitu 96,13%. Hal ini menunjukkan bahwa model memiliki performa sangat baik dalam mengklasifikasikan mahasiswa yang berpotensi mengundurkan diri.
3. Mahasiswa yang diterima melalui jalur mandiri memiliki rasio pengunduran diri yang lebih tinggi dibandingkan dengan mahasiswa yang masuk melalui jalur SNMPTN dan SBMPTN. Selain itu, mahasiswa lulusan SMA memiliki risiko mengundurkan diri yang lebih tinggi dibandingkan dengan lulusan SMK atau MA. Mahasiswa yang mengalami penurunan

nilai IPS dari semester 1 hingga semester 4 lebih dominan dalam data mahasiswa yang mengundurkan diri.

Adapun saran untuk penelitian selanjutnya diharapkan dapat menggunakan atribut yang lebih beragam serta melakukan perbandingan algortima untuk memperoleh model yang lebih optimal.

DAFTAR PUSTAKA

[1] Salam, A., Nugroho, F. B., & Zeniarja, J. (2020b). Implementasi Algoritma K Nearest Neighbor Berbasis Forward Selection Untuk Prediksi Mahasiswa Non Aktif Universitas Dian Nuswantoro Semarang. JOINS (Journal of Information System), 5(1), 69–76.

[2] Wahyudin, M. (2019). Penentuan Penjualan Produk Minuman Bernutrisi Terlaris dengan Algoritma C4.5 pada PT. Nutrifood Indonesia.

[3] Kusumo, D. S., Bijaksana, M. A., & Darmantoro, D. (2014). Data Mining Dengan Algoritma Apriori Pada Rdbms Oracle. Jurnal Penelitian dan Pengembangan Telekomunikasi, 8, 1-5.

[4] Wiyono, S., & Abidin, T. (2018). Perbandingan Algoritma Machine Learning SVM dan Decision Tree untuk Prediksi Keaktifan Mahasiswa. Jurnal & Penelitian Teknik Informatika, 3(1).

[5] Bode, A. (2019). Perbandingan Metode Prediksi Support Vector Machine dan Linear Regression Menggunakan Backward Elimination pada Produksi Minyak Kelapa. Jurnal Sistem Informasi dan Teknik Komputer, 4(2), 1–7.

[6] Mahanggara, A., & Laksito, A. D. (2019). Prediksi Pengunduran Diri Mahasiswa Universitas Amikom Yogyakarta Menggunakan Metode Naive Bayes. Jurnal SIMETRIS, 10(1).

[7] Sinanto, I., Nuraminah, A., & Studi, P. (2022). Komparasi Algoritma Feature Selection Pada Analisis Sentimen Review Film. JUITIK, 2(2).

[8] Monika Parapat, I., & Tanzil Furqon, M. (2018). Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 2(10), 3163–3169.

[9] Santoso, I., Gata, W., & Budi Paryanti, A. (2017). Penggunaan Feature Selection di Algoritma Support Vector Machine untuk Sentimen Analisis Komisi Pemilihan Umum. Jurnal Rekayasa Sistem dan Teknologi Informasi, 1(3), 364–370.

[10] Widayani, W., & Harliana, H. (2021). Analisis Support Vector Machine Untuk Pemberian Rekomendasi Penundaan Biaya Kuliah Mahasiswa. Jurnal Sains dan Informatika, 7(1), 20–27.

[11] Nugroho, A. S., Witarto, A. B., & Handoko, D. (2015). Support Vector Machine Untuk Prakiraan Beban Listrik Harian pada Sektor Industri. Jurnal Teknologi Elektro dan Kejuruan, 24(2), 1693 8739.

[12] Pratiwi, N., & Setyawan, Y. (2021). Analisis Akurasi dari Perbedaan Fungsi Kernel dan Cost Pada Support Vector Machine Studi Kasus Klasifikasi Curah Hujan di Jakarta. Journal of Fundamental Mathematics and Applications (JFMA), 4(2), 203–212.

[13] Fitriyah, N., Warsito, B., Asih, D., & Maruddani, I. (2020). Analisis Sentimen Gojek pada Media Sosial Twitter Dengan Klasifikasi Support Vector Machine (Svm). Jurnal Gaussian, 9(3), 376–390.

[14] Nicolson, A., & Paliwal, K. K. (2019). Deep Learning for Minimum Mean-square Error Approaches to Speech Enhancement. Speech Communication, 111(May), 44–55.

[15] Kumalasari, N. A. (2014). Implementasi Algoritma Modified K-Nearest Neighbor (MKNN) untuk Menentukan Tingkat Resiko Penyakit Lemak Darah (Profil Lipid). Universitas Brawijaya.