# TESIS

# OPTIMASI HYPERPARAMETER GRID SEARCH PADA PERBANDINGAN RANDOM FOREST DAN SUPPORT VECTOR MACHINE UNTUK PREDIKSI RETENSI MAHASISWA STRATA 1 BERBASIS DATASET PRIMER MULTIDIMENSI


disusun oleh

**DIMAS PRADIPTO**

**24.55.2702**

**Konsentrasi: Digital Transformation Intelligence**



**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS AMIKOM YOGYAKARTA**

**YOGYAKARTA**

**2026**

# TESIS

# OPTIMASI HYPERPARAMETER GRID SEARCH PADA PERBANDINGAN RANDOM FOREST DAN SUPPORT VECTOR MACHINE UNTUK PREDIKSI RETENSI MAHASISWA STRATA 1 BERBASIS DATASET PRIMER MULTIDIMENSI

# GRID SEARCH HYPERPARAMETER OPTIMIZATION IN THE COMPARISON OF RANDOM FOREST AND SUPPORT VECTOR MACHINE FOR PREDICTING BACHELOR'S DEGREE STUDENT RETENTION BASED ON A MULTIDIMENSIONAL PRIMARY DATASET

Diajukan untuk memenuhi salah satu syarat mencapai derajat Pascasarjana

Program Studi Teknik Informatika


Disusun oleh

**DIMAS PRADIPTO**

**24.55.2702**

**Konsentrasi: Digital Transformation Intelligence**



**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS AMIKOM YOGYAKARTA**

**YOGYAKARTA**


**2026**

## HALAMAN PERSETUJUAN

**OPTIMASI HYPERPARAMETER GRID SEARCH PADA PERBANDINGAN RANDOM FOREST DAN SUPPORT VECTOR MACHINE UNTUK PREDIKSI RETENSI MAHASISWA STRATA 1 BERBASIS DATASET PRIMER MULTIDIMENSI**

**GRID SEARCH HYPERPARAMETER OPTIMIZATION IN THE COMPARISON OF RANDOM FOREST AND SUPPORT VECTOR MACHINE FOR PREDICTING BACHELOR'S DEGREE STUDENT RETENTION BASED ON A MULTIDIMENSIONAL PRIMARY DATASET**

yang disusun dan diajukan oleh

**DIMAS PRADIPTO**

**24.55.2702**

telah disetujui oleh Dosen Pembimbing Tesis
pada tanggal 11 Juni 2026

**Dosen Pembimbing,**

**I Made Artha Agastya, S.T., M.Eng., Ph.D**
NIK. 190302352

## HALAMAN PENGESAHAN

**OPTIMASI HYPERPARAMETER GRID SEARCH PADA PERBANDINGAN RANDOM FOREST DAN SUPPORT VECTOR MACHINE UNTUK PREDIKSI RETENSI MAHASISWA STRATA 1 BERBASIS DATASET PRIMER MULTIDIMENSI**

**GRID SEARCH HYPERPARAMETER OPTIMIZATION IN THE COMPARISON OF RANDOM FOREST AND SUPPORT VECTOR MACHINE FOR PREDICTING BACHELOR'S DEGREE STUDENT RETENTION BASED ON A MULTIDIMENSIONAL PRIMARY DATASET**

yang disusun dan diajukan oleh

**DIMAS PRADIPTO**

**24.55.2702**

Telah dipertahankan di depan Dewan Penguji
pada tanggal 1 Juli 2026

**Susunan Dewan Penguji**

| Nama Penguji | Tanda Tangan |
|---|---|
| **Robert Marco, S.T., M.T., Ph.D.**<br>**NIK. 190302228** | |
| **Dhani Ariatmanto, S.Kom., M.Kom., Ph.D.**<br>**NIK. 190302197** | |
| **I Made Artha Agastya, S.T., M.Eng., Ph.D.**<br>**NIK. 190302352** | |

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer
Tanggal 1 Juli 2026

**DEKAN FAKULTAS ILMU KOMPUTER**

**Prof. Dr. Kusrini, M.Kom**
**NIK. 190302106**

## HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

**Nama mahasiswa : Dimas Pradipto**
**NIM : 24.55.2702**

Menyatakan bahwa Tesis dengan judul berikut:

**OPTIMASI HYPERPARAMETER GRID SEARCH PADA PERBANDINGAN RANDOM FOREST DAN SUPPORT VECTOR MACHINE UNTUK PREDIKSI RETENSI MAHASISWA STRATA 1 BERBASIS DATASET PRIMER MULTIDIMENSI**

Dosen Pembimbing : I Made Artha Agastya, S.T., M.Eng., Ph.D

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 1 Juli 2026

Yang Menyatakan,

Dimas Pradipto

## HALAMAN PERSEMBAHAN

*Dengan penuh rasa syukur dan kerendahan hati, karya ini kupersembahkan kepada:*

*Kedua orang tuaku tercinta,*

*yang tak pernah lelah mendoakan, mendukung, dan menjadi sumber kekuatan dalam setiap langkahku. Terima kasih atas kasih sayang yang tak terhingga dan pengorbanan yang tak ternilai.*

*Istri tercinta,*

*yang selalu setia mendampingi dalam setiap proses, memberikan doa, kesabaran, dan dukungan tanpa henti. Terima kasih telah menjadi tempat berbagi lelah, harapan, dan kebahagiaan.*

*Keluarga dan orang-orang terdekat,*

*yang senantiasa memberikan semangat dan kehangatan di sepanjang perjalanan ini.*

*Para dosen dan pembimbing,*

*yang dengan penuh kesabaran membimbing dan mengarahkan hingga karya ini dapat terselesaikan.*

*Sahabat dan teman seperjuangan,*

*yang telah menjadi bagian dari perjalanan ini, berbagi cerita, tawa, dan dukungan.*

*Serta untuk diriku sendiri,*

*yang telah berjuang dan bertahan hingga sampai di titik ini.*

## KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat Allah SWT atas segala limpahan rahmat, taufik, dan hidayah-Nya sehingga penulisan tesis ini dapat diselesaikan sebagaimana mestinya. Tesis berjudul " Optimasi Hyperparameter Grid Search Pada Perbandingan Random Forest Dan Support Vector Machine Untuk Prediksi Retensi Mahasiswa Strata 1 Berbasis Dataset Primer Multidimensi" diajukan sebagai salah satu persyaratan akademik dalam rangka memperoleh gelar Magister Komputer pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas AMIKOM Yogyakarta.

Penyelesaian tesis ini tidak terlepas dari kontribusi berbagai pihak yang telah memberikan dukungan, bimbingan, dan arahan kepada penulis. Untuk itu, penulis menyampaikan penghargaan dan terima kasih yang sebesar-besarnya kepada:

1. Prof. Dr. Kusrini, M.Kom, selaku Dekan Fakultas Ilmu Komputer Universitas AMIKOM Yogyakarta.
2. I Made Artha Agastya, S.T., M.Eng., Ph.D., selaku Dosen Pembimbing yang telah memberikan bimbingan, arahan, dan koreksi secara intensif selama proses penyusunan tesis ini.
3. Seluruh Dosen Penguji atas masukan konstruktif yang diberikan demi penyempurnaan tesis ini.
4. Pimpinan dan staf Universitas Ibnu Sina yang telah memberikan izin penelitian serta memfasilitasi proses pengumpulan data.

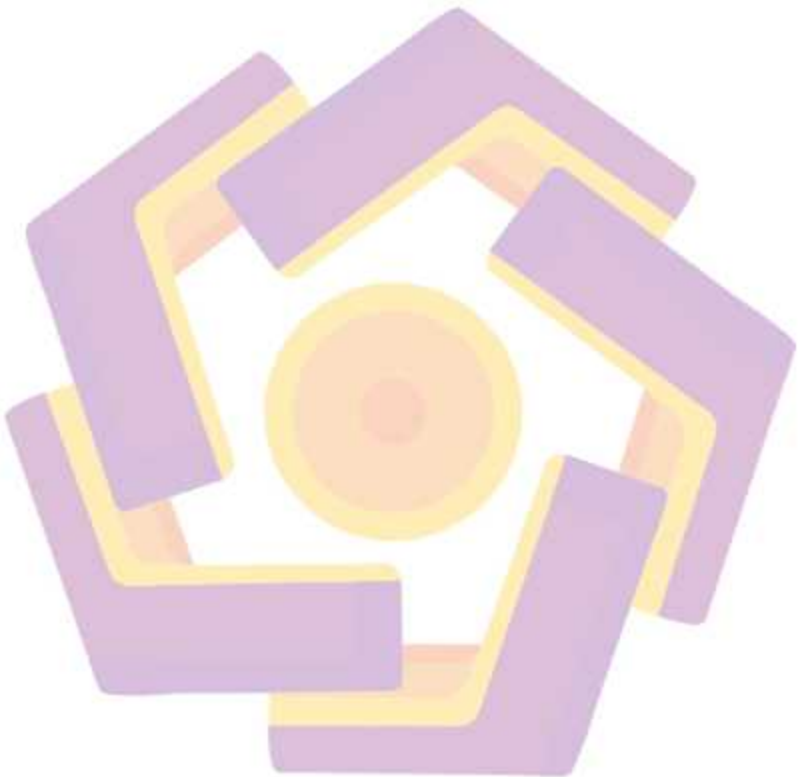
5. Mertua, istri, dan besti tercinta atas doa, dukungan moral, dan pengorbanan yang tidak ternilai.

Penulis menyadari bahwa tesis ini tidak luput dari keterbatasan dan kekurangan. Oleh karena itu, kritik serta saran yang bersifat konstruktif sangat penulis harapkan guna penyempurnaan lebih lanjut. Semoga tesis ini memberikan kontribusi yang berarti bagi pengembangan ilmu pengetahuan, khususnya dalam bidang machine learning dan manajemen pendidikan tinggi di Indonesia.

Yogyakarta, 1 Mei 2026
Penulis

**DAFTAR ISI**

## DAFTAR TABEL



## DAFTAR GAMBAR

## DAFTAR LAMPIRAN

# DAFTAR LAMBANG DAN SINGKATAN

## 1. Lambang Matematika dan Statistika

| Lambang | Keterangan |
|---|---|
| $\alpha_i$ | Lagrange multipliers dalam formulasi SVM; bobot untuk setiap support vector |
| $\gamma$ (gamma) | Parameter kernel RBF yang mengontrol lebar distribusi Gaussian |
| $\lambda$ | Bilangan acak dalam rentang [0, 1] digunakan pada algoritma SMOTE |
| $\|\cdot\|$ | Notasi norma Euclidean (jarak dalam ruang vektor) |
| $\Sigma$ | Operasi penjumlahan (sigma) |
| $\in$ | Notasi keanggotaan himpunan ("termasuk dalam") |
| $\hat{y}$ | Nilai prediksi kelas (y-hat) |
| argmax | Fungsi yang mengembalikan argumen dengan nilai terbesar |
| $I(\cdot)$ | Fungsi indikator; bernilai 1 jika kondisi terpenuhi, 0 jika tidak |
| $\exp(\cdot)$ | Fungsi eksponensial dengan basis bilangan natural $e \approx 2{,}718$ |
| $K(x_i, x_j)$ | Fungsi kernel yang mengukur kemiripan antara dua vektor fitur |
| $f(x)$ | Fungsi keputusan (decision function) pada SVM |
| $b$ | Nilai bias (intercept) dalam fungsi keputusan SVM |
| $T$ | Jumlah total pohon keputusan dalam ensemble Random Forest |
| $h_t(x)$ | Prediksi pohon ke-t untuk input x dalam Random Forest |
| $c$ | Label kelas target dalam klasifikasi |
| $n$ | Jumlah sampel data/titik data dalam dataset |
| $p$ | Jumlah fitur (dimensi) dalam dataset |
| $x_i, x_j$ | Vektor fitur data ke-i dan ke-j |
| $y_i$ | Label kelas data ke-i (+1 atau −1 untuk SVM biner) |
| $x'$ | Nilai fitur setelah normalisasi Min-Max |
| $x_{min}$ | Nilai minimum fitur dalam data training |
| $x_{max}$ | Nilai maksimum fitur dalam data training |
| $x_{new}$ | Sampel sintetis baru hasil SMOTE |
| $x_{nn}$ | Nearest neighbor dari sampel $x_i$ dalam SMOTE |
| $C$ | Parameter regularisasi pada SVM yang mengontrol trade-off margin vs kesalahan |
| $N$ | Total jumlah sampel dalam dataset |
| TP | True Positive – prediksi benar untuk kelas positif |
| TN | True Negative – prediksi benar untuk kelas negatif |
| FP | False Positive – prediksi salah (diprediksi positif, aktualnya negatif) |
| FN | False Negative – prediksi salah (diprediksi negatif, aktualnya positif) |

## 2. Singkatan dan Akronim

| Singkatan | Kepanjangan dan Keterangan |
|---|---|
| AUC | Area Under the Curve – luas area di bawah kurva ROC sebagai metrik evaluasi model |
| CART | Classification and Regression Trees – algoritma pohon keputusan untuk klasifikasi dan regresi |
| CV | Cross Validation – teknik evaluasi model dengan membagi data menjadi beberapa fold |
| DNN | Deep Neural Network – jaringan saraf tiruan berlapis banyak |
| DTI | Digital Transformation Intelligence – konsentrasi Program Studi dalam penelitian ini |

| | |
|---|---|
| EWS | Early Warning System — sistem peringatan dini untuk mendeteksi mahasiswa berisiko dropout |
| F1 | F1-Score — rata-rata harmonik antara Precision dan Recall |
| FEB | Fakultas Ekonomi dan Bisnis — salah satu fakultas di Universitas Ibnu Sina |
| FIKes | Fakultas Ilmu Kesehatan — salah satu fakultas di Universitas Ibnu Sina |
| FST | Fakultas Sains dan Teknologi — salah satu fakultas di Universitas Ibnu Sina |
| GI | Gini Impurity — metrik ketidakmurnian yang digunakan untuk evaluasi split pada pohon keputusan |
| IPK | Indeks Prestasi Kumulatif — rata-rata nilai akademik mahasiswa secara kesseluruhan (skala 0–4) |
| IPS | Indeks Prestasi Semester — rata-rata nilai akademik dalam satu semester (skala 0–4) |
| K3 | Keselamatan dan Kesehatan Kerja — nama program studi di Universitas Ibnu Sina |
| KNN | K-Nearest Neighbors — algoritma klasifikasi berbasis kemiripan tetangga terdekat |
| LMS | Learning Management System — platform manajemen pembelajaran daring (e.g., Moodle) |
| LR | Logistic Regression — algoritma klasifikasi berbasis regresi logistik |
| ML | Machine Learning — pembelajaran mesin; cabang kecerdasan buatan |
| MOOC | Massive Open Online Course — platform pembelajaran daring terbuka dan masif |
| MLP | Multi-Layer Perceptron — jenis jaringan saraf tiruan berlapis banyak |
| NB | Naïve Bayes — algoritma klasifikasi probabilistik berbasis teorema Bayes |
| NN | Neural Network — jaringan saraf tiruan |
| NPM | Nomor Pokok Mahasiswa — identitas unik setiap mahasiswa |
| OvO | One-vs-One — strategi klasifikasi multikelas dengan membuat pasangan kelas binary |
| OvR | One-vs-Rest — strategi klasifikasi multikelas dengan satu kelas vs semua kelas lainnya |
| PTS | Perguruan Tinggi Swasta — institusi pendidikan tinggi yang dikelola swasta |
| PTN | Perguruan Tinggi Negeri — institusi pendidikan tinggi milik pemerintah |
| RBF | Radial Basis Function — fungsi kernel yang digunakan dalam SVM |
| RF | Random Forest — algoritma ensemble berbasis pohon keputusan dengan bagging |
| ROC | Receiver Operating Characteristic — kurva yang menggambarkan trade-off TP rate vs FP rate |
| SIAKAD | Sistem Informasi Akademik — sistem pengelolaan data akademik universitas |
| SMOTE | Synthetic Minority Over-sampling Technique — teknik oversampling untuk kelas minoritas |
| SKS | Satuan Kredit Semester — satuan beban studi mahasiswa |
| SVM | Support Vector Machine — algoritma machine learning berbasis maximized margin hyperplane |
| SVC | Support Vector Classifier — implementasi SVM untuk tugas klasifikasi |
| T. Industri | Teknik Industri — nama program studi di Universitas Ibnu Sina |
| T. Informatika | Teknik Informatika — nama program studi di Universitas Ibnu Sina |
| UIS | Universitas Ibnu Sina — objek studi kasus dalam penelitian ini, berlokasi di Batam |

# DAFTAR ISTILAH

| Istilah | Definisi |
|---|---|
| **Accuracy (Akurasi)** | Metrik evaluasi yang mengukur proporsi prediksi benar terhadap total data. Dihitung sebagai (TP+TN)/(TP+TN+FP+FN). Cocok digunakan saat distribusi kelas seimbang. |
| **Algoritma** | Serangkaian langkah atau instruksi yang terstruktur dan terdefinisi dengan baik untuk menyelesaikan suatu permasalahan tertentu, khususnya dalam konteks komputasi dan machine learning. |
| **Bagging (Bootstrap Aggregating)** | Teknik ensemble learning yang membangun beberapa model (pohon keputusan) secara paralel dari subset data training yang diambil secara acak dengan penggantian (bootstrap sample), kemudian menggabungkan hasilnya melalui voting atau rata-rata. |
| **Class Imbalance (Ketidakseimbangan Kelas)** | Kondisi dalam dataset klasifikasi di mana satu atau lebih kelas memiliki jumlah sampel yang jauh lebih sedikit dibandingkan kelas lainnya. Pada penelitian ini: Aktif 75%, Berisiko 21,9%, Tidak Aktif 3,2%. |
| **Confusion Matrix** | Tabel ringkasan yang memvisualisasikan performa model klasifikasi dengan menampilkan jumlah prediksi benar (True Positive, True Negative) dan salah (False Positive, False Negative) untuk setiap kelas. |
| **Cross Validation** | Teknik evaluasi model yang membagi dataset menjadi k bagian (fold) secara bergilir sebagai data testing, sementara sisanya digunakan untuk training. Penelitian ini menggunakan Stratified 5-Fold Cross Validation. |
| **Data Cleaning** | Proses mengidentifikasi dan memperbaiki atau menghapus data yang tidak akurat, tidak lengkap, duplikat, atau tidak relevan dari dataset sebelum dilakukan analisis lebih lanjut. |
| **Dataset** | Kumpulan data terstruktur yang digunakan untuk melatih dan menguji model machine learning. Pada penelitian ini: 2.389 record mahasiswa dari 6 program studi dengan 32 variabel awal yang setelah melalui proses seleksi fitur dan pembersihan data menghasilkan 18 variabel prediktor dari empat dimensi (akademik, demografis, sosial-ekonomi, dan administratif) yang digunakan dalam pemodelan klasifikasi. |
| **Decision Tree (Pohon Keputusan)** | Algoritma machine learning berbasis pohon yang membuat keputusan melalui serangkaian pertanyaan biner (ya/tidak) berdasarkan nilai fitur. Merupakan komponen dasar dari algoritma Random Forest. |
| **Dropout (Putus Studi)** | Kondisi di mana mahasiswa meninggalkan atau menghentikan studi sebelum menyelesaikan program pendidikan yang sedang ditempuh, baik karena alasan akademik maupun non-akademik. |
| **Early Warning System (Sistem Peringatan Dini)** | Sistem berbasis data yang secara proaktif mengidentifikasi mahasiswa yang berisiko dropout pada tahap awal, sehingga memungkinkan intervensi akademik yang tepat waktu oleh pihak universitas. |
| **Encoding** | Proses mengkonversi variabel kategorik (seperti jenis kelamin, asal sekolah) menjadi representasi numerik agar dapat diproses oleh algoritma machine learning. |
| **Ensemble Learning** | Pendekatan machine learning yang menggabungkan prediksi dari beberapa model secara bersamaan untuk menghasilkan prediksi akhir yang lebih akurat dan stabil dibandingkan model tunggal. |

| | |
|---|---|
| **F1-Score** | Rata-rata harmonik antara Precision dan Recall. Sangat berguna untuk mengevaluasi model pada dataset yang tidak seimbang karena mempertimbangkan baik false positive maupun false negative. Nilai terbaik adalah 1,0. |
| **Feature Importance** | Skor yang mengukur kontribusi relatif setiap fitur/variabel dalam proses pengambilan keputusan model. Pada Random Forest, dihitung berdasarkan rata-rata penurunan Gini impurity di seluruh pohon. |
| **Gini Impurity** | Metrik yang mengukur seberapa sering elemen yang dipilih secara acak akan salah diklasifikasikan. Digunakan sebagai kriteria pemilihan split pada Decision Tree dan Random Forest. Nilai 0 berarti murni (satu kelas). |
| **Grid Search** | Metode optimasi hyperparameter yang secara sistematis menguji semua kombinasi nilai parameter yang telah ditentukan, kemudian memilih kombinasi yang menghasilkan performa model terbaik. |
| **Hyperparameter** | Parameter konfigurasi model machine learning yang nilainya ditentukan sebelum proses training dimulai dan tidak dapat dipelajari dari data training itu sendiri (misalnya jumlah pohon, kedalaman pohon, nilai C pada SVM). |
| **Hyperplane** | Bidang (subspace) berdimensi n-1 yang digunakan oleh SVM sebagai batas pemisah antar kelas dalam ruang fitur n-dimensi. Hyperplane optimal memaksimalkan margin antara kelas-kelas yang dipisahkan. |
| **IPK (Indeks Prestasi Kumulatif)** | Nilai rata-rata kumulatif dari seluruh mata kuliah yang telah ditempuh mahasiswa, dihitung pada skala 0 hingga 4,0. IPK di bawah 2,0 umumnya menjadi indikator risiko akademik. |
| **IPS (Indeks Prestasi Semester)** | Nilai rata-rata mata kuliah yang ditempuh mahasiswa dalam satu semester, dihitung pada skala 0 hingga 4,0. Penurunan IPS secara konsisten merupakan indikator risiko dropout. |
| **Kernel Trick** | Teknik matematis dalam SVM yang memungkinkan pemrosesan data non-linear dengan cara memetakan data ke ruang berdimensi lebih tinggi menggunakan fungsi kernel, tanpa perlu menghitung koordinat eksplisit dalam ruang tersebut. |
| **Klasifikasi** | Tugas machine learning supervised learning yang memprediksi kelas/kategori dari input data berdasarkan pola yang dipelajari dari data training berlabel. Penelitian ini menggunakan klasifikasi 3 kelas: Aktif, Berisiko, Tidak Aktif. |
| **Label Encoding** | Teknik encoding yang mengubah nilai kategori menjadi bilangan bulat berurutan (0, 1, 2, ...). Digunakan untuk variabel biner seperti jenis kelamin dan asal sekolah. |
| **Machine Learning** | Cabang kecerdasan buatan yang memungkinkan sistem komputer untuk belajar dan membuat keputusan dari data tanpa diprogram secara eksplisit, dengan mengidentifikasi pola dan membuat prediksi berdasarkan pengalaman sebelumnya. |
| **Majority Voting** | Strategi aggregasi prediksi dalam ensemble learning di mana kelas yang diprediksi oleh jumlah model terbanyak dipilih sebagai hasil akhir. Digunakan oleh Random Forest untuk menggabungkan prediksi dari seluruh pohon. |
| **Min-Max Normalization** | Teknik penskalaan fitur yang mengubah nilai asli menjadi rentang [0, 1] menggunakan rumus: $x' = (x - x\_min) / (x\_max - x\_min)$. Digunakan sebelum melatih model SVM. |
| **Overfitting** | Kondisi model machine learning yang terlalu 'hafal' pola pada data training sehingga performanya buruk pada data baru (testing). Ditandai dengan akurasi training yang sangat tinggi namun akurasi testing yang rendah. |

| | |
|---|---|
| **Precision** | Metrik evaluasi yang mengukur proporsi prediksi positif yang benar-benar positif: TP / (TP + FP). Tinggi artinya model jarang memprediksi positif secara salah. |
| **Recall (Sensitivity)** | Metrik evaluasi yang mengukur proporsi data positif yang berhasil teridentifikasi: TP / (TP + FN). Sangat penting dalam konteks EWS karena ingin mendeteksi sebanyak mungkin mahasiswa berisiko. |
| **Random Forest** | Algoritma ensemble berbasis pohon keputusan yang membangun banyak pohon menggunakan bagging dan random feature selection. Prediksi akhir diperoleh melalui majority voting dari seluruh pohon. |
| **Retensi Mahasiswa** | Kemampuan institusi pendidikan tinggi untuk mempertahankan mahasiswa agar tetap terdaftar dan aktif belajar hingga menyelesaikan program studi. Merupakan indikator penting keberhasilan universitas. |
| **RBF (Radial Basis Function)** | Fungsi kernel SVM berbasis jarak Euclidean: $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$. Cocok untuk data yang tidak dapat dipisahkan secara linear. Merupakan kernel terbaik yang digunakan dalam penelitian ini. |
| **SIAKAD (Sistem Informasi Akademik)** | Sistem informasi berbasis komputer yang digunakan oleh universitas untuk mengelola data akademik mahasiswa, termasuk nilai, jadwal, KRS, dan status keaktifan. |
| **SKS (Satuan Kredit Semester)** | Satuan yang digunakan untuk mengukur beban studi mahasiswa. Satu SKS umumnya setara dengan 50 menit perkuliahan tatap muka per minggu. Gelar S1 umumnya membutuhkan 144–160 SKS. |
| **SMOTE (Synthetic Minority Over-sampling Technique)** | Teknik penanganan class imbalance dengan membuat sampel sintetis baru untuk kelas minoritas melalui interpolasi antara sampel-sampel yang ada, menggunakan k-nearest neighbors. |
| **Stratified Sampling** | Teknik pembagian data yang memastikan proporsi kelas target pada data training dan testing sama dengan proporsi kelas pada dataset keseluruhan, sehingga representasi kelas terjaga. |
| **Support Vector** | Titik-titik data dari setiap kelas yang berada paling dekat dengan hyperplane pemisah. Titik-titik inilah yang secara kritis menentukan posisi dan orientasi hyperplane dalam SVM. |
| **Support Vector Machine (SVM)** | Algoritma supervised learning yang mencari hyperplane optimal dengan margin maksimum untuk memisahkan kelas dalam ruang fitur. Sangat efektif untuk data berdimensi tinggi dan mendukung kernel trick. |
| **Training Data (Data Latih)** | Subset dataset yang digunakan untuk melatih model machine learning agar dapat mempelajari pola dari data. Penelitian ini menggunakan 80% data (1.912 sampel) untuk training. |
| **Testing Data (Data Uji)** | Subset dataset yang digunakan untuk mengevaluasi performa model setelah training selesai. Penelitian ini menggunakan 20% data (478 sampel) untuk testing. |
| **Underfitting** | Kondisi model yang terlalu sederhana sehingga tidak mampu menangkap pola dalam data, ditandai dengan akurasi rendah baik pada data training maupun testing. |
| **Variabel Prediktor (Fitur)** | Variabel input yang digunakan model machine learning untuk membuat prediksi. Pada penelitian ini terdapat 18 variabel prediktor dari empat dimensi, yaitu akademik, demografis, sosial-ekonomi, dan administratif. |
| **Weighted Average** | Rata-rata yang memperhitungkan proporsi setiap kelas dalam dataset saat menghitung metrik evaluasi secara keseluruhan. Lebih representatif dari macro average untuk dataset yang tidak seimbang. |

## INTISARI

Retensi mahasiswa merupakan salah satu isu strategis yang paling kritis dalam pengelolaan pendidikan tinggi modern. Data dari National Center for Education Statistics (NCES) Amerika Serikat mencatat bahwa sekitar 40% mahasiswa yang mendaftar di perguruan tinggi tidak berhasil menyelesaikan gelarnya dalam enam tahun pertama, sementara di Indonesia, Kementerian Pendidikan dan Kebudayaan mencatat 602.208 mahasiswa atau sekitar 7% dari total 8,48 juta mahasiswa terdaftar berhenti kuliah pada tahun 2019, dengan 79,5% di antaranya berasal dari Perguruan Tinggi Swasta. Kondisi ini mendorong penelitian berbasis data di Universitas Ibnu Sina (UIS), perguruan tinggi swasta di Kota Batam dengan 2.389 mahasiswa aktif dari enam program studi pada tiga fakultas, yang mencatat variasi tingkat retensi signifikan antarprogram studi khususnya Teknik Informatika dengan persentase mahasiswa berisiko tertinggi sebesar 36,2%.

Penelitian ini bertujuan membangun dataset primer retensi mahasiswa UIS yang mencakup 18 variabel prediktor dari empat dimensi, yaitu akademik (IPS Semester 1–4, IPK Sementara, Jumlah SKS Gagal, Jumlah MK Mengulang, Jumlah Nilai D & E), demografis (Jenis Kelamin, Asal Sekolah, Usia Mendaftar, Asal Kota), sosial-ekonomi (Pekerjaan Orang Tua, Penghasilan Orang Tua, Jumlah Cuti), dan administratif (Program Studi, Periode Masuk, Sisa Masa Studi) dari 2.389 mahasiswa angkatan 2021/2022–2023/2024. Ketidakseimbangan kelas yang ekstrem (Aktif 75,0%, Berisiko 21,9%, Tidak Aktif 3,2%) ditangani menggunakan Synthetic Minority Over-sampling Technique (SMOTE) yang diterapkan eksklusif pada data training untuk mencegah data leakage. Kedua algoritma Random Forest (RF) dan Support Vector Machine (SVM) dengan kernel RBF dioptimalkan melalui Grid Search 5-Fold Cross Validation dan dievaluasi menggunakan akurasi, precision, recall, serta F1-Score weighted dan macro average.

Hasil penelitian menunjukkan Random Forest secara konsisten mengungguli SVM pada seluruh metrik evaluasi, dengan akurasi 92,24% vs. 87,63% dan F1-Score Macro Average 84,06% vs. 75,27%. Perbedaan paling signifikan terdapat pada deteksi kelas "Tidak Aktif" dengan F1-Score RF (72,73%) lebih tinggi 15,59 poin dibandingkan SVM (57,14%). Analisis feature importance mengungkap bahwa Jumlah SKS Gagal (0,2847) dan IPK Sementara (0,2134) merupakan prediktor terkuat, diikuti IPS Semester 1 (0,1023), sementara faktor non-akademik riwayat cuti (0,0428) dan penghasilan orang tua (0,0312) turut berkontribusi signifikan. Disimpulkan bahwa Random Forest merupakan algoritma terbaik untuk prediksi retensi mahasiswa UIS. Peneliti selanjutnya disarankan memperluas fitur dataset dengan variabel psikologis, membandingkan dengan algoritma boosting (XGBoost, LightGBM), serta mengeksplorasi teknik penanganan class imbalance lain seperti ADASYN atau SMOTE-Tomek Links.



**Kata Kunci:** Class Imbalance, Prediksi Dropout, Random Forest, Retensi Mahasiswa, SMOTE, Support Vector Machine.

## ABSTRACT

Student retention is one of the most critical strategic issues in modern higher education management. Data from the National Center for Education Statistics (NCES) in the United States indicates that approximately 40% of students who enroll in higher education institutions fail to complete their degrees within six years, while in Indonesia, the Ministry of Education and Culture recorded 602,208 students approximately 7% of the total 8.48 million enrolled students dropping out in 2019, with 79.5% originating from private universities. This condition has prompted data-driven research at Universitas Ibnu Sina (UIS), a private higher education institution in Batam City with 2,389 active students across six study programs in three faculties, which has recorded significant retention rate variations among programs particularly Informatics Engineering, which exhibits the highest proportion of at-risk students at 36.2%.

This study aims to construct a primary student retention dataset for UIS encompassing 18 predictor variables from four dimensions: academic performance (Semester GPA 1–4, Cumulative GPA, Failed Credits, Repeated Courses, D & E Grades), demographic characteristics (Gender, School of Origin, Age at Enrollment, City of Origin), socioeconomic factors (Parental Occupation, Parental Income, Leave Count), and administrative factors (Study Program, Enrollment Period, Remaining Study Period) from 2,389 students of the 2021/2022–2023/2024 cohorts. The extreme class imbalance (Active 75.0%, At-Risk 21.9%, Inactive 3.2%) was addressed using the Synthetic Minority Over-sampling Technique (SMOTE), applied exclusively to the training data to prevent data leakage. Both Random Forest (RF) and Support Vector Machine (SVM) with an RBF kernel were optimized through Grid Search 5-Fold Cross Validation and evaluated using accuracy, precision, recall, and both weighted and macro average F1-Score.

The results demonstrate that Random Forest consistently outperformed SVM across all evaluation metrics, achieving an accuracy of 92.24% vs. 87.63% and a Macro Average F1-Score of 84.06% vs. 75.27%. The most significant difference was observed in the detection of the "Inactive" class, where RF's F1-Score (72.73%) was 15.59 percentage points higher than SVM's (57.14%). Feature importance analysis revealed that the Number of Failed Credits (0.2847) and Cumulative GPA (0.2134) are the strongest predictors, followed by Semester 1 GPA (0.1023), while non-academic factors including leave history (0.0428) and parental income (0.0312) also contributed significantly. It is concluded that Random Forest is the optimal algorithm for student retention prediction at UIS. Future researchers are advised to expand the feature set with psychological variables, benchmark against boosting algorithms (XGBoost, LightGBM), and explore alternative class imbalance handling techniques such as ADASYN or SMOTE-Tomek Links.



**Keywords:** Class Imbalance, Dropout Prediction, Random Forest, Student Retention, SMOTE, Support Vector Machine.

# BAB 1

# PENDAHULUAN

## 1.1. Latar Belakang

Retensi mahasiswa merupakan salah satu isu strategis yang paling kritis dalam pengelolaan pendidikan tinggi modern. Berbagai kajian internasional menunjukkan bahwa fenomena putus studi (dropout) bukan sekadar persoalan individual, melainkan tantangan institusional yang berdampak luas secara sosial, ekonomi, dan akademik. Data dari *National Center for Education Statistics* (NCES) Amerika Serikat mencatat bahwa sekitar 40% mahasiswa yang mendaftar di perguruan tinggi tidak berhasil menyelesaikan gelarnya dalam enam tahun pertama [1]. Di tingkat global, laporan UNESCO (2022) memperkirakan lebih dari sepertiga mahasiswa di negara berkembang meninggalkan bangku kuliah sebelum menyelesaikan studi, dengan kerugian ekonomi yang ditanggung institusi mencapai miliaran dolar setiap tahunnya [2]. Di Indonesia, data Kementerian Pendidikan dan Kebudayaan mencatat sebanyak 602.208 mahasiswa atau sekitar 7% dari total 8,48 juta mahasiswa yang terdaftar memutuskan untuk berhenti kuliah pada tahun 2019, dengan 79,5% di antaranya setara 478.826 mahasiswa berasal dari Perguruan Tinggi Swasta (PTS) [3]. Kondisi ini menegaskan bahwa kemampuan sebuah perguruan tinggi dalam menjaga mahasiswanya tetap terdaftar dan aktif hingga menyelesaikan studiyang dikenal sebagai retensi mahasiswamerupakan indikator keberhasilan institusional yang tidak dapat diabaikan [4]. Tingginya angka dropout

tidak hanya merugikan mahasiswa secara personal, tetapi juga menurunkan akreditasi, reputasi, serta keberlanjutan finansial perguruan tinggi itu sendiri [5].

Universitas Ibnu Sina (UIS) adalah perguruan tinggi swasta yang berlokasi di Kota Batam, Provinsi Kepulauan Riau, dengan segmen mahasiswa yang sebagian besar berasal dari keluarga pekerja industri dan memiliki latar belakang sosial-ekonomi yang beragam. Karakteristik ini menjadikan UIS rentan terhadap permasalahan retensi yang multidimensional. Saat ini, Universitas Ibnu Sina memiliki 2.389 mahasiswa aktif yang tersebar di enam program studi dalam tiga fakultas, yakni Fakultas Ekonomi dan Bisnis (FEB), Fakultas Sains dan Teknologi (FST) dan Fakultas Ilmu Kesehatan (FIKes). Berdasarkan data administrasi akademik periode 2021/2022–2023/2024, terdapat variasi tingkat retensi yang signifikan antarprogram studi, dengan sebagian program studi mencatat mahasiswa tidak aktif atau cuti dalam jumlah yang cukup tinggi. Kondisi tersebut mengindikasikan bahwa setiap program studi menghadapi dinamika retensi yang unik, yang dipengaruhi oleh kombinasi faktor akademik seperti Indeks Prestasi Kumulatif (IPK), Indeks Prestasi Semester (IPS), dan jumlah SKS yang gagal maupun faktor non-akademik, seperti pekerjaan orang tua, penghasilan orang tua, serta riwayat cuti studi. Mengingat besarnya dampak putus studi bagi keberlanjutan institusi dan masa depan mahasiswa, Universitas Ibnu Sina memiliki urgensi yang nyata untuk mengembangkan sistem prediksi retensi yang mampu mendeteksi risiko secara dini dan berbasis data.

Perkembangan pesat bidang machine learning telah membuka peluang bagi institusi pendidikan untuk memanfaatkan data akademik secara lebih optimal dalam

memprediksi retensi mahasiswa. Sejumlah penelitian terdahulu telah mengeksplorasi berbagai algoritma klasifikasi untuk tujuan ini dan hasilnya menjanjikan. Novianto et al. (2024) melaporkan bahwa Random Forest mencapai akurasi 97,67% dalam memprediksi capaian studi mahasiswa, mengungguli Support Vector Machine (SVM) yang memperoleh akurasi 91,47% [6], sementara Park & Yoo (2021) dalam konteks pembelajaran daring menemukan F1-score sebesar 96,70% [7]. Meski demikian, kajian terhadap penelitian-penelitian tersebut mengungkap tiga celah mendasar yang membatasi relevansinya bagi konteks perguruan tinggi di Indonesia. Pertama, sebagian besar penelitian hanya menggunakan fitur akademik dasar seperti IPK dan nilai ujian, tanpa mengintegrasikan faktor non-akademik yang sesungguhnya turut memengaruhi keputusan mahasiswa untuk bertahan atau meninggalkan studi. Kondisi ini terjadi karena data non-akademikseperti kondisi ekonomi keluarga, status pekerjaan, dan keterlibatan organisasi umumnya tidak tersedia dalam sistem informasi akademik yang digunakan sebagai sumber data penelitian, sehingga peneliti cenderung membatasi analisis hanya pada variabel yang mudah diakses [8]. Kedua, persoalan ketidakseimbangan kelas (class imbalance) di mana jumlah mahasiswa aktif jauh lebih besar dibandingkan mahasiswa berisiko atau tidak aktifmasih jarang ditangani secara eksplisit. Hal ini terjadi karena banyak penelitian menggunakan akurasi sebagai metrik evaluasi tunggal, yang justru menyembunyikan kegagalan model dalam mendeteksi kelas minoritas yang paling kritis, yaitu mahasiswa yang berisiko putus studi [9]. Ketiga, fokus penelitian lebih banyak diarahkan pada prediksi kelulusan atau kinerja akademik, bukan pada keputusan objektif mahasiswa untuk

bertahan atau meninggalkan studi. Hal ini disebabkan oleh kemudahan dalam mengukur nilai dan kelulusan secara kuantitatif, sementara status retensi yang bersifat dinamis dan multifaktor lebih sulit dioperasionalkan sebagai variabel target dalam sebuah dataset [10].

Secara kuantitatif, tinjauan terhadap sembilan penelitian terdahulu yang dikaji dalam Tabel 2.1 memperkuat ketiga celah tersebut dengan temuan sebagai berikut. Lima dari sembilan penelitian (55,6%) hanya mengandalkan fitur dari satu dimensi akademik saja seperti IPK, nilai ujian, atau transkrip, sementara sisanya menggunakan data log LMS, kuesioner, atau variabel demografis dasar secara terpisah tanpa upaya integrasi lintas dimensi. Tidak satu pun dari sembilan penelitian tersebut yang mengintegrasikan faktor sosial-ekonomi seperti pekerjaan orang tua, penghasilan orang tua, atau riwayat cuti sebagai variabel prediktor dalam model klasifikasi. Selain itu, seluruh penelitian (100%) tidak menangani permasalahan ketidakseimbangan kelas secara eksplisit melalui teknik *resampling* seperti SMOTE, dan sebanyak empat penelitian (44,4%) masih memfokuskan prediksi pada kinerja akademik atau kepuasan mahasiswa, bukan pada keputusan objektif mahasiswa untuk bertahan atau meninggalkan studi. Fakta ini menegaskan bahwa celah penelitian yang telah diidentifikasi bukan sekadar asumsi naratif, melainkan pola sistematis yang terdokumentasi secara empiris dan menjadi dasar urgensi penelitian ini.

Berdasarkan uraian latar belakang di atas, teridentifikasi bahwa fitur-fitur yang digunakan dalam dataset retensi mahasiswa pada penelitian-penelitian sebelumnya masih sangat terbatas dan belum merepresentasikan kompleksitas

faktor yang sesungguhnya memengaruhi keputusan mahasiswa untuk bertahan atau meninggalkan studi. Oleh karena itu, penelitian ini hadir dengan pendekatan yang berbeda, yaitu membangun dataset primer baru dari data akademik dan non-akademik Universitas Ibnu Sina yang mencakup fitur-fitur dengan korelasi kuat terhadap status retensi, seperti tren IPS per semester, riwayat cuti, pekerjaan orang tua, dan penghasilan orang tua, yang belum banyak dieksplorasi secara terintegrasi dalam penelitian sebelumnya. Di samping itu, penelitian ini secara eksplisit menangani permasalahan ketidakseimbangan kelas menggunakan teknik Synthetic Minority Over-sampling Technique (SMOTE), guna memastikan model prediksi yang dihasilkan tidak hanya akurat secara global, tetapi juga sensitif terhadap kelas mahasiswa berisiko yang paling kritis. Dengan memanfaatkan dan membandingkan algoritma Random Forest dan Support Vector Machine (SVM) pada dataset yang dirancang secara komprehensif tersebut, penelitian ini diharapkan mampu menghasilkan model prediksi retensi yang lebih representatif, akurat, dan dapat diimplementasikan sebagai landasan pengambilan keputusan akademik di Universitas Ibnu Sina.

Pemilihan algoritma Random Forest (RF) dan Support Vector Machine (SVM) dalam penelitian ini didasarkan pada beberapa pertimbangan yang relevan dengan karakteristik permasalahan retensi mahasiswa. Random Forest merupakan algoritma ensemble berbasis pohon keputusan yang terbukti unggul dalam menangani data berdimensi tinggi dengan campuran variabel numerik dan kategorikal, mampu mengukur tingkat kepentingan tiap fitur (feature importance) secara langsung, serta bersifat robust terhadap outlier dan noise yang umum

ditemukan dalam data akademik nyata [4] [5]. Karakteristik ini sangat sesuai untuk dataset retensi mahasiswa yang melibatkan 18 variabel dari empat dimensi berbeda dengan skala pengukuran yang beragam. Di sisi lain, Support Vector Machine dipilih karena kemampuannya dalam menemukan batas keputusan yang optimal pada data berdimensi tinggi melalui penggunaan kernel Radial Basis Function (RBF), sehingga mampu menangkap hubungan non-linear antara variabel prediktor dan status retensi mahasiswa [6]. Tinjauan sistematis terhadap 70 artikel (2018-2023) mengkonfirmasi bahwa RF dan SVM merupakan dua algoritma yang paling dominan dan konsisten menghasilkan performa tinggi dalam penelitian prediksi akademik mahasiswa, sehingga perbandingan kedua algoritma ini akan memberikan landasan empiris yang kuat dalam menentukan pendekatan terbaik untuk sistem deteksi dini risiko putus studi di Universitas Ibnu Sina.

Berdasarkan hal tersebut, penelitian ini memberikan kontribusi pada dua isu utama. Pertama, pada isu dataset, penelitian ini membangun dataset primer retensi mahasiswa UIS yang mencakup 18 variabel prediktor dari empat dimensi yaitu akademik, demografis, sosial-ekonomi, dan administratif dari 2.389 mahasiswa angkatan 2021/2022 hingga 2023/2024, yang belum pernah dibangun sebelumnya dalam konteks perguruan tinggi swasta di Indonesia. Kedua, pada isu algoritma, penelitian ini menerapkan teknik SMOTE secara eksklusif pada data training untuk menangani ketidakseimbangan kelas yang ekstrem (Aktif 75,0%, Berisiko 21,9%, Tidak Aktif 3,2%), sekaligus membandingkan performa algoritma Random Forest dan Support Vector Machine menggunakan metrik F1-Score Macro Average dan Recall yang lebih representatif untuk data tidak seimbang.

### 1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, penelitian ini dirumuskan untuk menjawab permasalahan berikut:

1. Apakah dataset primer multidimensi yang dibangun dari data akademik, demografis, sosial-ekonomi, dan administratif Universitas Ibnu Sina memiliki fitur-fitur dengan korelasi yang cukup untuk memprediksi status retensi mahasiswa?
2. Apakah terdapat faktor-faktor dominan yang secara signifikan memengaruhi retensi mahasiswa di Universitas Ibnu Sina berdasarkan analisis *feature importance* pada model prediksi sehingga dapat dijadikan indikator risiko putus studi?
3. Apakah penerapan teknik SMOTE mampu menangani permasalahan ketidakseimbangan kelas pada dataset retensi mahasiswa, dan apakah algoritma Random Forest memiliki kinerja prediksi yang lebih baik dibandingkan Support Vector Machine dalam mengklasifikasikan status mahasiswa aktif, berisiko, dan tidak aktif?

### 1.3. Batasan Masalah

Dalam penelitian ini, penulis membatasi ruang lingkup agar kajian dapat dilakukan secara fokus, terarah, dan mendalam. Batasan masalah penelitian ini adalah sebagai berikut:

1. Penelitian hanya menggunakan data mahasiswa dari enam program studi yang tergabung dalam tiga fakultas Universitas Ibnu Sina, dengan total 2.389 mahasiswa angkatan 2021/2022–2023/2024, karena keterbatasan waktu dan

biaya untuk memperluas cakupan ke institusi lain maupun periode yang lebih panjang.

2. Variabel yang dianalisis mencakup 18 variabel prediktor dari empat dimensi, yaitu: faktor akademik (IPS Semester 1, IPS Semester 2, IPS Semester 3, IPS Semester 4, IPK Sementara, Jumlah SKS Gagal, Jumlah Mata Kuliah Mengulang, dan Jumlah Nilai D & E), faktor demografis (Jenis Kelamin, Asal Sekolah, Usia Mendaftar, dan Asal Kota), faktor sosial-ekonomi (Pekerjaan Orang Tua, Penghasilan Orang Tua, dan Jumlah Cuti), serta faktor administratif (Program Studi, Periode Masuk, dan Sisa Masa Studi). Variabel di luar cakupan data administrasi akademik SIAKAD tidak dianalisis.
3. Analisis klasifikasi dilakukan menggunakan metode Support Vector Machine (SVM) dengan optimasi parameter cost dan gamma, serta Random Forest dengan optimasi jumlah pohon dan jumlah variabel acak. Algoritma lain tidak digunakan, sehingga hasil penelitian berlaku khusus untuk perbandingan kedua metode tersebut.
4. Penanganan ketidakseimbangan kelas dilakukan menggunakan teknik SMOTE (Synthetic Minority Over-sampling Technique) yang diterapkan hanya pada data training untuk mencegah kebocoran informasi (data leakage) ke data testing.
5. Fokus penelitian adalah prediksi status retensi mahasiswa, yang diklasifikasikan ke dalam tiga kategori: Aktif, Berisiko, dan Tidak Aktif. Variabel kelulusan tepat waktu, indeks prestasi luar studi, atau aspek karier pasca-lulus tidak menjadi cakupan penelitian ini.

### 1.4. Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Membangun dataset primer retensi mahasiswa Universitas Ibnu Sina yang mencakup fitur-fitur akademik dan non-akademik dengan korelasi yang memadai terhadap status keberlanjutan studi, sehingga dapat digunakan sebagai basis data yang representatif untuk pemodelan prediksi.
2. Mengidentifikasi dan menganalisis faktor-faktor yang paling berpengaruh terhadap retensi mahasiswa di enam program studi Universitas Ibnu Sina berdasarkan analisis feature importance pada model machine learning yang dikembangkan.
3. Menerapkan teknik SMOTE untuk mengatasi ketidakseimbangan kelas pada dataset retensi, sekaligus membandingkan kinerja algoritma Random Forest dan Support Vector Machine dalam mengklasifikasikan status retensi mahasiswa, guna menentukan algoritma yang paling efektif dan dapat diandalkan untuk mendukung pengambilan keputusan akademik di Universitas Ibnu Sina.

### 1.5. Manfaat Penelitian

Bagian ini memuat penjelasan tentang manfaat penelitian yaitu:

1. Bagi Universitas Ibnu Sina

   Penelitian ini menghasilkan dataset retensi mahasiswa yang komprehensif serta model prediksi berbasis machine learning yang dapat digunakan sebagai dasar perancangan kebijakan akademik yang lebih tepat sasaran. Secara spesifik, hasil klasifikasi model Random Forest yang dikembangkan dapat

diimplementasikan sebagai sistem deteksi dini (Early Warning System) risiko putus studi mahasiswa di Universitas Ibnu Sina. Dengan sistem ini, pihak universitas dapat mengidentifikasi mahasiswa yang berisiko dropout sejak semester awal, sehingga intervensi pencegahan seperti bimbingan akademik, bantuan finansial, dan program konseling dapat dilakukan secara proaktif sebelum mahasiswa benar-benar memutuskan untuk meninggalkan studi. Pemahaman mendalam tentang faktor-faktor risiko dropout di setiap program studi juga memungkinkan universitas untuk mengalokasikan sumber daya secara lebih efisien dan tepat sasaran.

2. Bagi Program Studi dan Fakultas

   Model prediksi yang dikembangkan membantu program studi dan fakultas dalam mengidentifikasi indikator-indikator akademik dan non-akademik yang paling berpengaruh terhadap retensi mahasiswanya. Informasi ini dapat menjadi landasan perancangan program mentoring, konseling akademik, serta dukungan finansial yang lebih terarah bagi mahasiswa yang berisiko putus studi.

3. Bagi Pengembangan Sistem Akademik

   Penelitian ini memberikan kontribusi nyata dalam pengembangan sistem informasi akademik berbasis analisis data dan machine learning. Dataset dan model klasifikasi yang dihasilkan dapat diintegrasikan langsung ke dalam Sistem Informasi Akademik (SIAKAD) Universitas Ibnu Sina sebagai modul deteksi dini retensi mahasiswa. Secara teknis, setiap semester sistem dapat menjalankan model Random Forest terhadap data terbaru mahasiswa dan

menghasilkan label prediksi status retensi (Aktif, Berisiko, atau Tidak Aktif) beserta probabilitasnya. Label prediksi ini selanjutnya dapat ditampilkan kepada Dosen Pembimbing Akademik (DPA) atau bagian kemahasiswaan sebagai peringatan dini, sehingga tindakan pencegahan dapat segera diambil terhadap mahasiswa yang teridentifikasi berisiko tinggi untuk putus studi.

4. Bagi Komunitas Ilmiah dan Peneliti Selanjutnya

   Penelitian ini memberikan kontribusi metodologis berupa kerangka pengembangan dataset retensi dengan fitur-fitur yang belum banyak dieksplorasi dalam literatur sebelumnya, serta pendekatan penanganan class imbalance menggunakan SMOTE dalam konteks data akademik perguruan tinggi swasta di Indonesia. Hasil penelitian ini dapat menjadi referensi bagi peneliti lain yang ingin mengembangkan model prediksi serupa di perguruan tinggi lain dengan karakteristik yang berbeda.

# BAB 2

# TINJAUAN PUSTAKA

## 2.1. Tinjauan Pustaka

Sejumlah penelitian telah mengeksplorasi penerapan algoritma Random Forest (RF) dan Support Vector Machine (SVM) dalam konteks prediksi berbasis data akademik. Berikut adalah uraian masing-masing penelitian terdahulu yang relevan beserta dataset, metode, hasil, kekuatan, dan kelemahannya.

Sebuah penelitian melakukan perbandingan metode klasifikasi Random Forest dan Support Vector Machine dalam memprediksi capaian studi mahasiswa. Dataset yang digunakan bersumber dari sistem informasi akademik yang terdiri dari variabel IPK dan SKS. Metode yang diterapkan adalah RF dan SVM dengan evaluasi berbasis akurasi. Hasil penelitian menunjukkan bahwa RF mencapai akurasi 97,67% sedangkan SVM memperoleh akurasi 91,47%. Kekuatan penelitian ini terletak pada perbandingan langsung dua algoritma menggunakan data akademik riil. Namun demikian, penelitian ini memiliki kelemahan yaitu hanya menggunakan fitur akademik dasar (IPK dan SKS) tanpa mengintegrasikan faktor non-akademik, menggunakan akurasi sebagai metrik evaluasi tunggal yang tidak representatif untuk data tidak seimbang, serta berfokus pada prediksi kinerja akademik bukan pada keputusan retensi mahasiswa [4].

Penelitian lain meneliti prediksi keberhasilan akademik mahasiswa menggunakan algoritma machine learning. Dataset yang digunakan berupa data akademik dasar mahasiswa. Metode yang diterapkan meliputi SVM, RF, dan Logistic Regression. Hasil penelitian menunjukkan bahwa RF memperoleh akurasi

80%, SVM 78%, dan LR 76%. Kekuatan penelitian ini terletak pada penggunaan data akademik riil dari institusi pendidikan tinggi. Kelemahannya adalah tidak menekankan retensi sebagai variabel target, serta tidak menghubungkan hasil prediksi dengan strategi intervensi nyata yang dapat diterapkan oleh pihak universitas untuk mencegah dropout [5].

Pengembangan sistem peringatan dini di perguruan tinggi juga telah dilakukan menggunakan algoritma SVM dan RF. Dataset yang digunakan sangat terbatas, hanya mencakup variabel demografi mahasiswa berupa jenis kelamin dan usia. Metode yang diterapkan adalah SVM dan RF dengan evaluasi berbasis akurasi. Hasil penelitian menunjukkan bahwa SVM memperoleh akurasi 85% sedangkan RF memperoleh 80%. Kekuatan penelitian ini terletak pada konsep early warning system yang relevan dengan kebutuhan institusi pendidikan tinggi untuk deteksi dini mahasiswa berisiko. Kelemahannya adalah data yang sangat terbatas karena hanya mencakup variabel demografi tanpa data akademik, perilaku, atau konteks sosial, sehingga model tidak dapat menangkap kompleksitas faktor yang memengaruhi keputusan retensi mahasiswa [6].

Perbandingan algoritma Random Forest, Support Vector Machine, dan Neural Network juga telah dilakukan untuk klasifikasi kepuasan mahasiswa terhadap layanan akademik. Dataset yang digunakan berupa data kuesioner kepuasan mahasiswa. Metode yang diterapkan adalah RF, SVM, dan NN dengan evaluasi berbasis akurasi. Hasil penelitian menunjukkan bahwa RF memperoleh akurasi tertinggi sebesar 76,47% sedangkan SVM dan NN masing-masing memperoleh 74,12%. Kekuatan penelitian ini terletak pada perbandingan tiga

algoritma secara bersamaan. Kelemahannya adalah fokus pada persepsi subjektif (kepuasan) bukan perilaku objektif (retensi), serta tidak mengukur status retensi mahasiswa secara langsung sehingga tidak relevan dengan konstruk prediksi dropout [7].

Studi educational data mining juga telah dilakukan untuk memprediksi kinerja akademik mahasiswa menggunakan algoritma machine learning. Dataset yang digunakan berupa data nilai ujian tengah semester dan ujian akhir semester. Metode yang diterapkan meliputi RF, SVM, Logistic Regression, dan Neural Network. Hasil penelitian menunjukkan akurasi model berkisar antara 70% hingga 75%. Kekuatan penelitian ini terletak pada penggunaan data evaluasi akademik langsung. Kelemahannya adalah fokus pada prediksi nilai ujian bukan pada keputusan bertahan atau putus studi, sehingga tidak relevan dengan konstruk retensi yang bersifat multifaktor, serta tidak menangani permasalahan ketidakseimbangan kelas [8].

Prediksi dropout dini pada konteks pembelajaran online di perguruan tinggi juga telah diteliti. Dataset yang digunakan berupa data log aktivitas mahasiswa pada Learning Management System (LMS) Moodle. Metode yang diterapkan meliputi RF, SVM, dan Deep Neural Network (DNN). Hasil penelitian menunjukkan bahwa RF mencapai akurasi tertinggi sebesar 95% dengan F1-score 96,70%. Kekuatan penelitian ini terletak pada penggunaan data perilaku digital yang mencerminkan keterlibatan mahasiswa secara real-time dalam platform pembelajaran. Kelemahannya adalah konteks penelitian terbatas pada Massive Open Online Course (MOOC) yang tidak relevan untuk program S1 tatap muka, serta tidak

mempertimbangkan faktor sosial-ekonomi atau demografi mendalam yang turut memengaruhi keputusan dropout mahasiswa [9].

Sebuah framework machine learning juga telah dikembangkan untuk pengembangan kebijakan retensi mahasiswa. Dataset yang digunakan berupa data internal minimal yang terdiri dari variabel demografi dan akademik dasar. Metode yang diterapkan meliputi CatBoost, RF, SVM, dan Neural Network. Hasil penelitian menunjukkan bahwa CatBoost memperoleh F1-score tertinggi sebesar 82% diikuti RF dengan 81%. Kekuatan penelitian ini terletak pada orientasinya yang langsung mengarah pada pengembangan kebijakan retensi berbasis data. Kelemahannya adalah tidak menangani data imbalance secara eksplisit, hanya menggunakan data akademik dasar tanpa mengintegrasikan faktor sosial-ekonomi atau psikologis, serta tidak membandingkan RF dan SVM secara mendalam karena fokus utama pada CatBoost [10].

Prediksi dropout mahasiswa juga telah diteliti menggunakan data transkrip akademik sebagai dataset utama. Metode yang diterapkan meliputi CatBoost, RF, SVM, dan Neural Network. Hasil penelitian menunjukkan bahwa RF, NN, dan CatBoost memberikan akurasi yang tinggi. Kekuatan penelitian ini terletak pada penggunaan data longitudinal transkrip akademik yang mencerminkan perjalanan akademik mahasiswa. Kelemahannya adalah hanya menggunakan data akademik tanpa mengintegrasikan faktor sosial, ekonomi, dan perilaku yang turut memengaruhi keputusan dropout mahasiswa [11].

Prediksi dropout pada kursus online juga telah diteliti menggunakan algoritma SVM, RF, dan Logistic Regression. Dataset yang digunakan berupa data

interaksi mahasiswa dengan Learning Management System (LMS). Hasil penelitian menunjukkan bahwa RF mencapai akurasi tertinggi sebesar 84%, diikuti SVM dengan 80% dan LR dengan 75%. Kekuatan penelitian ini terletak pada perbandingan tiga algoritma klasifikasi secara langsung dalam konteks pembelajaran online. Kelemahannya adalah konteks kursus online singkat non-gelar yang tidak merepresentasikan dinamika program S1 penuh, serta tidak mencakup faktor sosial, emosional, atau finansial yang berpengaruh terhadap keputusan dropout [12].

Berdasarkan tinjauan terhadap sembilan penelitian di atas, dapat disimpulkan bahwa RF dan SVM merupakan dua algoritma yang paling dominan digunakan dalam prediksi berbasis data akademik, di mana RF secara konsisten menunjukkan performa yang lebih baik dibandingkan SVM pada mayoritas studi. Namun demikian, tiga celah metodologis teridentifikasi secara konsisten. Pertama, mayoritas penelitian hanya menggunakan fitur dari satu dimensi data (akademik saja atau demografi saja) tanpa integrasi lintas dimensi. Kedua, tidak satu pun penelitian yang menangani permasalahan ketidakseimbangan kelas secara eksplisit melalui teknik resampling. Ketiga, sebagian besar penelitian berfokus pada kinerja akademik atau kepuasan, bukan pada keputusan objektif mahasiswa untuk bertahan atau meninggalkan studi. Penelitian ini mengisi ketiga celah tersebut melalui pembangunan dataset primer multidimensi yang mengintegrasikan variabel dari empat dimensi (akademik, demografis, sosial-ekonomi, dan administratif), penerapan SMOTE untuk menangani ketidakseimbangan kelas, serta evaluasi berbasis F1-score dan Recall yang lebih representatif untuk data tidak seimbang.

Ringkasan posisi penelitian ini terhadap penelitian terdahulu disajikan secara detail pada Tabel 2.1..

### 2.2. Keaslian Penelitian

Tabel 2. 1 Matriks literatur review dan posisi penelitian

OPTIMASI HYPERPARAMETER GRID SEARCH PADA PERBANDINGAN RANDOM FOREST DAN SUPPORT VECTOR MACHINE UNTUK PREDIKSI RETENSI MAHASISWA STRATA 1 BERBASIS DATASET PRIMER MULTIDIMENSI

| No | Judul Penelitian | Nama Peneliti, Tahun, Index | Fokus Utama / Topik | Metode & Data yang Digunakan | Temuan Utama | Keterbatasan / Celah (Gap) | Kontribusi Keaslian Penelitian Ini |
|---|---|---|---|---|---|---|---|
| 1 | Perbandingan Metode Klasifikasi Random Forest Dan Support Vector Machine Dalam Memprediksi Capaian Studi Mahasiswa | Novianto et al., 2024 | Prediksi kinerja akademik (bukan retensi) | RF, SVM; data IPK dan SKS dari sistem akademik | Akurasi RF: 97,67%; SVM: 91,47% | Bukan fokus pada retensi; menggunakan akurasi sebagai metrik utama meskipun data tidak seimbang; tidak menangani data imbalance | Fokus eksplisit pada retensi mahasiswa (bertahan vs putus); evaluasi menggunakan F1-score dan Recall; menerapkan teknik SMOTE untuk menangani ketidakseimbangan kelas |

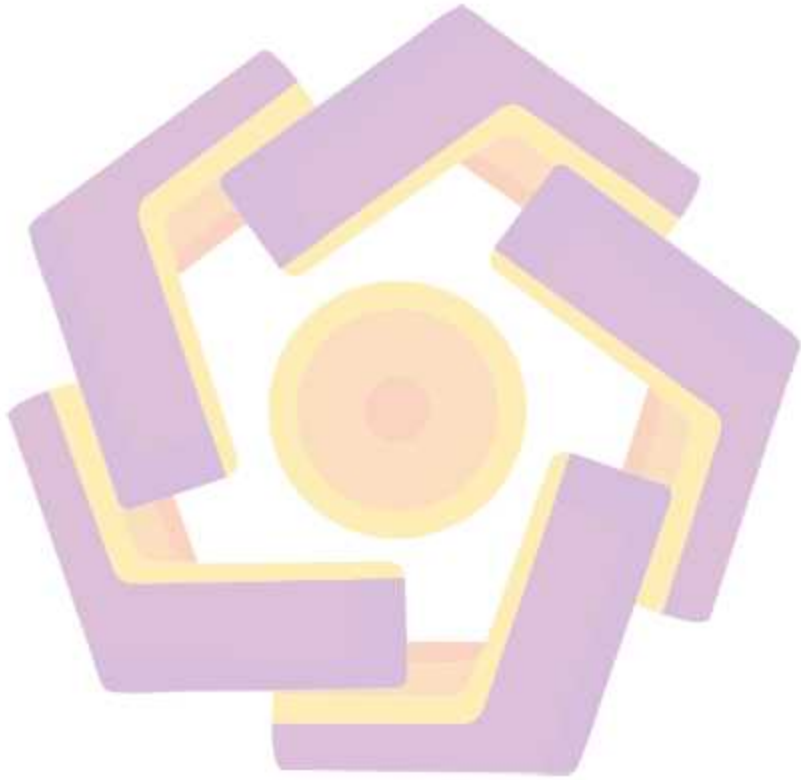
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2 | Early Dropout Prediction in Online Learning of University | Park & Yoo, 2021 | Prediksi dropout di lingkungan MOOC/online | RF, SVM, DNN; log aktivitas LMS (Moodle) | RF: akurasi 95%, F1-score 96,70% | Konteks pembelajaran online terbuka (MOOC), tidak relevan untuk program S1 tatap muka; tidak mempertimbangkan faktor sosial-ekonomi atau demografi mendalam | Studi kasus di program S1 tatap muka (Teknik Industri & Informatika); mengintegrasikan faktor non-akademik dari dimensi sosial-ekonomi (Pekerjaan Orang Tua, Penghasilan Orang Tua, Jumlah Cuti) dan administratif (Program Studi, Periode Masuk, Sisa Masa Studi) yang tidak tersedia dalam data log LMS |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 3 | Comparison of Random Forest Algorithm, Support Vector Machine and Neural Network for Classification of Student Satisfaction | Supriyadi et al., 2022 | Klasifikasi kepuasan layanan mahasiswa | RF, SVM, NN; data kuesioner kepuasan | Akurasi RF: 76,47%; SVM & NN: 74,12% | Tidak mengukur retensi secara langsung; fokus pada persepsi subjektif, bukan perilaku objektif (putus studi) | Menggunakan label objektif berdasarkan status administratif: "aktif", "cuti", "putus studi"; prediksi didasarkan pada data riil sistem akademik, bukan survei |
| 4 | Machine Learning Framework for Student Retention Policy Development | Hoca & Dimililer, 2025 | Prediksi risiko dropout untuk kebijakan retensi | CatBoost, RF, SVM, NN; data minimal internal (demografi & akademik dasar) | CatBoost: F1-score 82%; RF: 81% | Tidak menangani data imbalance; hanya menggunakan data akademik, mengabaikan konteks psikologis dan eksternal; tidak membandingkan RF vs SVM secara mendalam | Program S1 tatap muka reguler; variabel non-akademik dari dimensi sosial-ekonomi (Pekerjaan Orang Tua, Penghasilan Orang Tua, Jumlah Cuti) dan administratif (Program Studi, Periode Masuk, Sisa Masa Studi) yang tidak tersedia dalam |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | data log LMS; penanganan imbalance secara eksplisit |
| 5 | Educational Data Mining: Prediction of Students' Academic Performance Using Machine Learning Algorithms | Yağcı, 2022 | Prediksi nilai ujian akhir (kinerja akademik) | RF, SVM, LR, NN; data nilai ujian tengah & akhir semester | Akurasi model: 70–75% | Fokus pada prediksi nilai, bukan keputusan bertahan/putus studi; tidak relevan dengan konstruk retensi | Target prediksi adalah status retensi (bukan nilai); menggunakan data longitudinal periode 2022–2024 untuk menangkap dinamika putus studi |
| 6 | Predicting Student Dropout Using Machine Learning | Park, 2024 | Prediksi dini mahasiswa berisiko dropout | CatBoost, RF, SVM, NN; data transkrip akademik | RF, NN, CatBoost menunjukkan akurasi tinggi | Hanya menggunakan data akademik, mengabaikan faktor sosial, ekonomi, dan perilaku yang memengaruhi | Mengintegrasikan variabel komprehensif dari empat dimensi: akademik (IPS Semester 1–4, IPK Sementara, Jumlah SKS Gagal, Jumlah MK Mengulang, |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | keputusan putus studi | Jumlah Nilai D & E), demografis (Jenis Kelamin, Asal Sekolah, Usia Mendaftar, Asal Kota), sosial-ekonomi (Pekerjaan Orang Tua, Penghasilan Orang Tua, Jumlah Cuti), dan administratif (Program Studi, Periode Masuk, Sisa Masa Studi) |
| 7 | Predicting Dropout in Online Courses Using SVM, RF, and LR | Talamás-Carvajal, 2020 | Prediksi dropout di kursus online singkat | SVM, RF, LR; data interaksi LMS | Akurasi RF: 84%; SVM: 80%; LR: 75% | Konteks kursus online non-gelar, tidak merepresentasikan dinamika program S1 penuh; tidak mencakup faktor | Studi kasus di program S1 reguler tatap muka; memasukkan faktor sosial-ekonomi (Pekerjaan Orang Tua, Penghasilan Orang Tua, Jumlah Cuti) dan administratif (Program |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | sosial, emosional, atau finansial | Studi, Sisa Masa Studi) yang tidak tersedia dalam data interaksi LMS pada penelitian tersebut. |
| 8 | Predicting Student Success Using Machine Learning | Trussel, 2021 | Prediksi keberhasilan akademik umum | SVM, RF, LR; data akademik dasar | RF: 80%; SVM: 78%; LR: 76% | Tidak menekankan retensi; tidak menghubungkan hasil prediksi dengan strategi intervensi nyata | Hasil prediksi dikaitkan langsung dengan rekomendasi kebijakan intervensi (early warning system) untuk pihak universitas |
| 9 | Early Warning Systems in Higher Education Using SVM and RF | Selim & Rezk, 2020 | Sistem peringatan dini berbasis AI | SVM, RF; data demografi mahasiswa (jenis kelamin, usia) | Akurasi SVM: 85%; RF: 80% | Data sangat terbatas hanya mencakup variabel demografi, tanpa data akademik, perilaku, atau konteks sosial | Menggunakan data multi-dimensi: akademik (nilai, IPS), demografi, psikologis, dan eksternal, sehingga model lebih holistik dan representatif |



## 2.3. Landasan Teori

### 2.3.1. Pendidikan Tinggi

Pendidikan tinggi merupakan pendidikan jenjang untuk menuntut ilmu setelah pendidikan menengah (SMA/Sederajat). Pendidikan tinggi terdiri dari berbagai jenjang, yaitu jenjang diploma, sarjana, magister, doktor, dan program profesi. Satuan badan yang menjalankan pendidikan tinggi disebut perguruan tinggi dan dikenal dengan Perguruan Tinggi Negeri (PTN) dan Perguruan Tinggi Swasta (PTS). Bentuk perguruan tinggi ada bermacam-macam, seperti universitas, institute, sekolah tinggi, politeknik, spesialis, dan akademi. (Undang-Undang Republik Indonesia Nomor 12 pasal 1, 2012)[13].

Fungsi pendidikan tinggi berdasarkan Undang-Undang Republik Indonesia Nomor 12 pasal 4 tahun 2012 adalah sebagai berikut;

1. Mengembangkan kemampuan dan membentuk watak serta peradaban bangsa yang bermartabat dalam rangka mencerdaskan kehidupan bangsa.
2. Mengembangkan sivitas akademika yang inovatif, responsif, kreatif, terampil, berdaya saing, dan kooperatif melalui pelaksanaan tridharma, dan
3. Mengembangkan ilmu pengetahuan dan Teknologi dengan memperhatikan dan menerapkan nilai humaniora.

Adapun tujuan pendidikan tinggi berdasarkan Undang-Undang Republik Indonesia Nomor 12 pasal 5 tahun 2012 adalah;

1. Berkembangnya potensi Mahasiswa agar menjadi manusia yang beriman dan bertakwa kepada Tuhan Yang Maha Esa dan berakhlak mulia, sehat, berilmu, cakap, kreatif, mandiri, terampil, kompeten, dan berbudaya untuk kepentingan

bangsa.

2. Dihasilkannya lulusan yang menguasai cabang Ilmu Pengetahuan dan/atau Teknologi untuk memenuhi kepentingan nasional dan peningkatan daya saing bangsa.
3. Dihasilkannya Ilmu Pengetahuan dan Teknologi melalui Penelitian yang memperhatikan dan menerapkan nilai Humaniora agar bermanfaat bagi kemajuan bangsa, serta kemajuan peradaban dan kesejahteraan umat manusia.
4. Terwujudnya Pengabdian kepada Masyarakat berbasis penalaran dan karya Penelitian yang bermanfaat dalam memajukan kesejahteraan umum dan mencerdaskan kehidupan bangsa.

### 2.3.2. *Machine Learning*

Machine learning adalah cabang dari kecerdasan buatan (artificial intelligence) yang memungkinkan sistem komputer untuk belajar dari data dan meningkatkan kinerjanya secara otomatis tanpa diprogram secara eksplisit [14]. Berbeda dengan pemrograman konvensional yang mengandalkan aturan yang didefinisikan secara manual, machine learning membangun model matematis berdasarkan pola yang ditemukan dalam data pelatihan sehingga mampu membuat prediksi atau keputusan terhadap data baru yang belum pernah dilihat sebelumnya.

Dalam konteks penelitian ini, machine learning dimanfaatkan untuk membangun model prediktif yang mampu mengklasifikasikan status retensi mahasiswa ke dalam tiga kategori, yaitu Aktif, Berisiko, dan Tidak Aktif,

berdasarkan 18 variabel prediktor yang mencakup dimensi akademik, demografis, sosial-ekonomi, dan administratif. Pendekatan ini dipilih karena machine learning terbukti efektif dalam menangani data berdimensi tinggi, pola non-linear, serta ketidakseimbangan kelas yang umum ditemukan pada data retensi mahasiswa [15].

**1. Model *Supervised Learning***

Model ini digunakan untuk memprediksi hasil masa depan berdasarkan data historis untuk dipelajari menggunakan metode tertentu agar bisa memprediksi dengan akurat. Contoh penerapan metode *Supervised Learning* adalah untuk memprediksi kemungkinan terjadinya bahaya yang akan terjadi dengan melihat beberapa faktor sesuai dengan data historis yang telah dipelajari, seperti bencana gempa bumi, banjir, dan lain-lain[15].

*Supervised learning* adalah sebuah pendekatan dengan cara melatih data yang sudah ada dan terdapat variabel yang ditargetkan sehingga tujuan dari pendekatan ini adalah mengkelompokan suatu data ke data yang sudah ada. Metode analisis untuk *supervised learning* adalah *Decision tree, Random Forest, Nearest - Neighbor Classifier, Naive Bayes Classifier, Artificial Neural Network, Support Vector Machine, Fuzzy K-Nearest Neighbor*[16].

**2. *Model Unsupervised Learning***

*Unsupervised learning* tidak memiliki data latih, sehingga dari data yang ada, kita mengelompokan data tersebut menjadi 2 bagian atau 3 bagian dan seterusnya. Contoh metode ini adalah seseorang belum pernah membeli buku sama sekali, namun dalam suatu hari, orang tersebut membeli banyak buku dan ingin membaginya kedalam beberapa kategori agar nantinya mudah dicari namun belum

diketahui kategori dari buku tersebut. Maka pembagian buku dengan cara mengidentifikasi buku mana yang mirip berdasarkan isinya. Metode analisis *unsupervised learning* adalah *K-Means, Hierarchical Clustering, DBSCAN, Fuzzy C-Means, Self-Organizing Map*.

### 2.3.3. *Klasifikasi*

Klasifikasi adalah metode untuk memprediksi suatu kejadian atau keputusan yang akan datang berada di suatu titik. Klasifikasi merupakan suatu pekerjaan yang melakukan penilaian terhadap suatu objek data untuk masuk dalam suatu kelas tertentu dari sejumlah kelas yang tersedia[17]. Tahapan klasifikasi ada beberapa, yaitu;

**1. Data Collection**

*Data collection* adalah proses mengumpulkan dan memastikan informasi pada *variable of interest* (subjek yang akan dilakukan uji coba), dengan cara yang sistematis yang memungkinkan seseorang dapat menjawab pertanyaan dari permasalahan yang ada dan mengevaluasi hasil. Komponen pengumpulan data dari penelitian ini bersifat umum, bisa dilakukan untuk semua bidang studi termasuk ilmu fisik dan sosial, humaniora, bisnis, dan lainnya[18].

**2. Data Cleaning**

Data yang baik dan berkualitas adalah kunci dasar untuk menghasilkan data yang berkualitas, dengan cara menghapus data *error*, mengahapus data *incomplete* atau data yang nilai atributnya hilang atau datanya kosong, dan menyamakan satuan atau nilai dari data [9]*Data cleaning* adalah proses analisa kualitas dari suatu data dengan cara mengubah, mengoreksi, atau menghapus data-data yang salah, tidak

lengkap, tidak akurat, atau memiliki format yang salah dalam basis data guna menghasilkan data berkualitas tinggi Contoh;

Tabel 2. 2 Data Sebelum *Cleaning*

| No | Jenis Kimn | Prodi | Fakultas | Tempat Lahir | Kabupaten | Propinsi |
|---|---|---|---|---|---|---|
| 1 | P | Farmasi | MATEMATIKA DAN ILMU PENGETAHUA | WERU | KAB.INDRAMAYU | JAWA BARAT |
| 2 | L | Ilmu Komunikasi | PSIKOLOGI DAN ILMU SOSIAL BUDAYA | Bumi Agung | KAB.SLEMAN | D.I. YOGYAKARTA |
| 3 | L | Manajemen | EKONOMI | Tuban | KODYA YOGYAKARTA | D.I. YOGYAKARTA |
| 4 | | | TEKNIK SIPIL DAN PERENCANAAN | Lamongan | | |
| 5 | | | | Semarang | | |
| 6 | L | | TEKNIK SIPIL DAN PERENCANAAN | BUDRA | | |
| 7 | P | Psikologi | PSIKOLOGI DAN ILMU SOSIAL BUDAYA | Jakarta | KODYA JAKARTA UTARA | DKI JAKARTA |
| 8 | L | Teknik Sipil | TEKNIK SIPIL DAN PERENCANAAN | Bantul | KAB.BANTUL | D.I. YOGYAKARTA |
| 9 | L | Farmasi | MATEMATIKA DAN ILMU PENGETAHUA | Klaten | KAB.KLATEN | JAWA TENGAH |
| 10 | L | Teknik Industri | TEKNOLOGI INDUSTRI | Kediri | KAB.KEDIRI | JAWA TIMUR |
| 11 | L | Informatika | TEKNOLOGI INDUSTRI | Tuban | KAB.TUBAN | JAWA TIMUR |
| 12 | L | Teknik Kimia | TEKNOLOGI INDUSTRI | Kendal | KAB.KENDAL | JAWA TENGAH |
| 13 | P | Hukum | HUKUM | Madiun | | |
| 14 | P | Teknik Kimia | TEKNOLOGI INDUSTRI | Klangenan Cirebon | KAB.CIREBON | JAWA BARAT |
| 15 | P | Psikologi | PSIKOLOGI DAN ILMU SOSIAL BUDAYA | Serang | X kab.pandegelang | JAWA BARAT |
| 16 | L | Pendidikan Agama Islam | ILMU AGAMA ISLAM | Sleman | KAB.SLEMAN | D.I. YOGYAKARTA |
| 17 | L | Manajemen | EKONOMI | Serang | | |
| 18 | L | Manajemen | EKONOMI | Yogyakarta | KAB.KENDAL | JAWA TENGAH |
| 19 | L | Hukum | HUKUM | Sanggau | KAB.SANGGAU | KALIMANTAN BARAT |
| 20 | L | Hukum | HUKUM | Samarinda | KODYA YOGYAKARTA | D.I. YOGYAKARTA |

Tabel 2. 3 Data Setelah *Cleaning*

| No | Jenis Kimn | Prodi | Fakultas | Tempat Lahir | Kabupaten | Propinsi |
|---|---|---|---|---|---|---|
| 1 | P | Farmasi | MATEMATIKA DAN ILMU PENGETAHUA | WERU | KAB.INDRAMAYU | JAWA BARAT |
| 2 | L | Ilmu Komunikasi | PSIKOLOGI DAN ILMU SOSIAL BUDAYA | Bumi Agung | KAB.SLEMAN | D.I. YOGYAKARTA |
| 3 | L | Manajemen | EKONOMI | Tuban | KODYA YOGYAKARTA | D.I. YOGYAKARTA |
| 7 | P | Psikologi | PSIKOLOGI DAN ILMU SOSIAL BUDAYA | Jakarta | KODYA JAKARTA UTARA | DKI JAKARTA |
| 8 | L | Teknik Sipil | TEKNIK SIPIL DAN PERENCANAAN | Bantul | KAB.BANTUL | D.I. YOGYAKARTA |
| 9 | L | Farmasi | MATEMATIKA DAN ILMU PENGETAHUA | Klaten | KAB.KLATEN | JAWA TENGAH |
| 10 | L | Teknik Industri | TEKNOLOGI INDUSTRI | Kediri | KAB.KEDIRI | JAWA TIMUR |
| 11 | L | Informatika | TEKNOLOGI INDUSTRI | Tuban | KAB.TUBAN | JAWA TIMUR |
| 12 | L | Teknik Kimia | TEKNOLOGI INDUSTRI | Kendal | KAB.KENDAL | JAWA TENGAH |
| 14 | P | Teknik Kimia | TEKNOLOGI INDUSTRI | Klangenan Cirebon | KAB.CIREBON | JAWA BARAT |
| 15 | P | Psikologi | PSIKOLOGI DAN ILMU SOSIAL BUDAYA | Serang | X kab.pandegelang | JAWA BARAT |
| 16 | L | Pendidikan Agama Islam | ILMU AGAMA ISLAM | Sleman | KAB.SLEMAN | D.I. YOGYAKARTA |
| 18 | L | Manajemen | EKONOMI | Yogyakarta | KAB.KENDAL | JAWA TENGAH |
| 19 | L | Hukum | HUKUM | Sanggau | KAB.SANGGAU | KALIMANTAN BARAT |
| 20 | L | Hukum | HUKUM | Samarinda | KODYA YOGYAKARTA | D.I. YOGYAKARTA |

### 3. Data *Reduction*

Pada tahap data *reduction* adalah tahap dimana data yang telah terkumpul, kemudian di bersihkan, proses selanjutnya adalah memilih variabel atau atribut yang akan digunakan dalam penelitian. Tahap ini dilakukan untuk mengurangi atribut yang tidak digunakan akan tetapi tetap bersifat informatif[19].

Tabel 2. 4 *Data Sebelum Dan Sesudah Reduce Variabel*

| Variabel Asli |
|---|
| No Mhs |
| Nama Mhs |
| Jenis Klmn |
| Prodi |
| Propinsi |
| Pekerjaan Orang Tua |
| Penghasilan Orang Tua |
| IPK Sementara |
| Asal Sekolah |

| Variabel Setelah Reduce |
|---|
| Jenis Klmn |
| Prodi |
| Propinsi |
| Pekerjaan Orang Tua |
| Penghasilan Orang Tua |
| IPK Sementara |
| Asal Sekolah |

### 4. *Data Training* dan Data *Testing*

Data *training* digunakan oleh algoritma untuk membentuk sebuah model klasifikasi. Model ini merupakan representasi pengetahuan yang akan digunakan untuk mengukur sejauh mana klasifikasi berhasil melakukan prediski dengan benar. Karena itu, data yang ada yang ada pada data *testing* seharusnya tidak boleh ada pada data *training* sehingga dapat diketahui apakah model klasifikasi dapat melakukan klasifikasi dengan baik. Proporsi antara data *training* dan data *testing* tidak mengikat tetapi agar variasi dalam model tidak terlalu besar maka disarankan data *training* lebih besar dibandingkan data *testing*. Biasanya 3/4 dari total data dijadikan data *training* sedangkan sisanya dijadikan data *testing*. Selain itu, ada pula

penelitian yang menghasilkan keakuratan model klasifikasi optimum dengan proporsi 75 : 25 untuk data *training* dan data *testing*[20].

### 2.3.4. *Balancing* Data

*Balancing* data adalah merubah data yang tidak seimbang (*imbalance* data) menjadi data yang seimbang (*balance*). *Imbalance* data adalah kondisi ketidakseimbangan dalam jumlah data *training* antara dua kelas yang berbeda, salah satu kelasnya merepresentasikan jumlah data yang sangat besar (*majority class*) sedangkan kelas yang lainya merepresentasikan jumlah data yang sangat kecil (*minority class*)[21]. Metode ini secara luas dikenal sebagai 'Metode Sampling'. Umumnya, metode ini bertujuan untuk memodifikasi data yang tidak seimbang ke dalam distribusi yang seimbang menggunakan beberapa mekanisme. Modifikasi terjadi dengan mengubah ukuran kumpulan data asli dan menyediakan proporsi keseimbangan yang sama.

Menurut Tarid Wongvorachan (2023) menjelaskan bahwa *balancing* data ada beberapa metode, yaitu diantaranya adalah *under sampling*. Metode *under sampling* ini bekerja dengan kelas mayoritas (*majority class*), dengan mengurangi jumlah observasi dari kelas mayoritas untuk membuat kumpulan data dengan jumlah yang seimbang. Metode ini paling baik digunakan ketika kumpulan data sangat besar dan mengurangi jumlah sampel pelatihan yang membantu meningkatkan waktu operasi dan masalah penyimpanan. Metode *under sampling* secara acak memilih observasi dari kelas mayoritas yang dieliminasi sampai set data menjadi seimbang antar masing-masing kelas[22].

Selain metode under sampling yang bekerja dengan mengurangi jumlah observasi pada kelas mayoritas, terdapat pendekatan oversampling yang bekerja dengan cara sebaliknya, yaitu menambah jumlah observasi pada kelas minoritas hingga mencapai proporsi yang seimbang dengan kelas mayoritas. Pendekatan oversampling memiliki keunggulan dibandingkan under sampling karena tidak menghilangkan informasi yang terkandung dalam data kelas mayoritas, sehingga model klasifikasi tetap dapat mempelajari pola dari seluruh data yang tersedia. Namun demikian, teknik oversampling konvensional yang hanya menduplikasi data kelas minoritas secara identik memiliki kelemahan signifikan, yaitu model cenderung mengalami overfitting karena mempelajari pola yang sama secara berulang tanpa adanya variasi baru yang memperkaya representasi kelas minoritas dalam ruang fitur.

Untuk mengatasi kelemahan tersebut, Chawla et al. pada tahun 2002 memperkenalkan teknik Synthetic Minority Over-sampling Technique (SMOTE) yang merupakan salah satu teknik oversampling paling banyak digunakan dalam penelitian machine learning. Perbedaan mendasar antara SMOTE dengan teknik oversampling konvensional terletak pada mekanisme penambahan datanya. SMOTE tidak sekadar menduplikasi data yang sudah ada, melainkan menghasilkan sampel sintetis baru yang memiliki variasi nilai fitur melalui proses interpolasi antara sampel asli dari kelas minoritas dengan tetangga terdekatnya yang juga berasal dari kelas minoritas. Dengan demikian, setiap sampel sintetis yang dihasilkan merupakan titik data baru yang unik dan berbeda dari data aslinya,

namun tetap berada dalam distribusi yang realistis sesuai dengan karakteristik kelas minoritas.

Proses pembentukan sampel sintetis pada SMOTE dilakukan melalui beberapa tahapan. Pertama, algoritma memilih secara acak satu sampel dari kelas minoritas sebagai titik referensi. Kedua, algoritma mengidentifikasi sejumlah k tetangga terdekat (k-nearest neighbors) dari sampel tersebut yang juga berasal dari kelas minoritas menggunakan perhitungan jarak Euclidean dalam ruang fitur. Ketiga, salah satu dari k tetangga terdekat dipilih secara acak sebagai pasangan interpolasi. Keempat, sampel sintetis baru dibentuk pada titik yang terletak di sepanjang segmen garis yang menghubungkan sampel asli dengan tetangga yang terpilih, di mana posisi tepatnya ditentukan oleh bilangan acak λ. Secara matematis, proses pembentukan sampel sintetis SMOTE dirumuskan sebagai berikut:

$$x\_new = x\_i + \lambda \times (\hat{x}\_nn - x\_i) \quad \textbf{(3.20)}$$

Dimana x_new adalah sampel sintetis baru yang dihasilkan oleh algoritma SMOTE, x_i adalah sampel asli dari kelas minoritas yang dipilih secara acak sebagai titik referensi pembentukan sampel sintetis, $\hat{x}$_nn adalah salah satu dari k tetangga terdekat dari sampel x_i yang juga berasal dari kelas minoritas dan dipilih secara acak sebagai pasangan interpolasi, dan λ adalah bilangan acak yang diambil secara seragam dari rentang [0, 1] yang berfungsi mengontrol posisi relatif sampel sintetis pada segmen garis antara x_i dan $\hat{x}$_nn. Secara geometris, apabila λ bernilai 0 maka sampel sintetis akan berada tepat pada posisi x_i sehingga identik dengan sampel asli, sedangkan apabila λ bernilai 1 maka sampel sintetis akan berada tepat pada posisi $\hat{x}$_nn sehingga identik dengan tetangganya, dan nilai λ di antara

keduanya menghasilkan sampel sintetis pada posisi yang bervariasi di sepanjang segmen garis tersebut dalam ruang fitur.

Nilai parameter k pada umumnya ditetapkan sebesar 5, yang berarti setiap sampel kelas minoritas memiliki lima kandidat tetangga terdekat untuk proses interpolasi. Pemilihan nilai k memiliki implikasi yang signifikan terhadap kualitas sampel sintetis yang dihasilkan dan performa model klasifikasi secara keseluruhan. Nilai k yang terlalu kecil, misalnya k=1 atau k=2, dapat menghasilkan sampel sintetis yang terlalu mirip dengan data asli dan terkonsentrasi pada wilayah yang sangat sempit dalam ruang fitur, sehingga tidak memberikan variasi yang cukup bagi model untuk mempelajari pola yang lebih general. Sebaliknya, nilai k yang terlalu besar berpotensi menyebabkan sampel sintetis melampaui batas distribusi alami kelas minoritas dan tumpang tindih dengan wilayah kelas mayoritas, yang justru dapat meningkatkan ambiguitas pada batas keputusan (decision boundary) dan menurunkan performa klasifikasi. Oleh karena itu, nilai k=5 dianggap sebagai konfigurasi yang memberikan keseimbangan optimal antara variasi sampel sintetis dan konsistensi distribusi kelas minoritas.

Dalam penerapannya, terdapat satu aspek kritis yang harus diperhatikan, yaitu SMOTE harus diterapkan secara eksklusif pada data training saja, bukan pada keseluruhan dataset. Apabila SMOTE diterapkan sebelum pembagian data, maka sampel sintetis yang dihasilkan dari data training dapat mengandung informasi yang berasal dari data testing, sehingga terjadi data leakage yang menyebabkan evaluasi model menjadi terlalu optimistis dan tidak mencerminkan performa model pada data baru yang sesungguhnya. Keunggulan utama SMOTE dibandingkan metode

oversampling sederhana adalah kemampuannya memperluas wilayah keputusan kelas minoritas secara lebih natural tanpa sekadar memperkuat noise atau outlier yang sudah ada dalam data asli, sehingga model klasifikasi yang dilatih dengan data hasil SMOTE memiliki kemampuan generalisasi yang lebih baik terhadap data baru yang belum pernah dilihat sebelumnya.

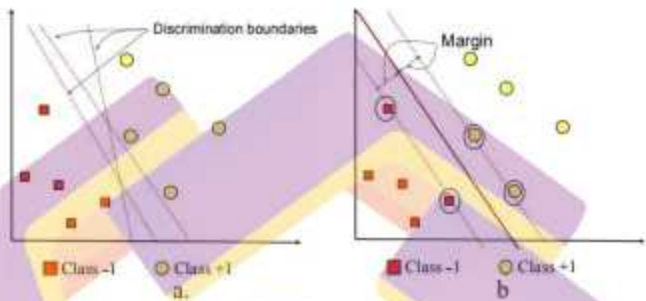
### 2.3.5. Support *Vector* Machine (SVM)

Vapnik memperkenalkan SVM untuk pertama kali pada tahun 1992 sebagai rangkaian konsep unggulan pada bidang *pattern recognition*. Usia SVM sebagai salah satu metode *pattern recognition* masih terbilang relatif muda. Dewasa ini SVM meruapakn salah satu metode yang berkembang pesat. SVM merupan salah satu metode *machine learning* yang bekerja atas prinsip *Structural Risk Minimization* (SRM) yang bertujuan untuk menemukan *hyperplane* terbaik untuk memisahkan dua kelas data. *SVM* bekerja dengan memaksimalkan *margin* yang merupakan jarak pemisah antara kedua kelas data tersebut. Pada dasarnya SVM memiliki prinsip linear, akan tetapi kini SVM telah berkembang sehingga dapat bekerja pada masalah *non*-linear. Cara kerja SVM pada masalah *non*-linear adalah dengan memasukkan konsep kernel pada ruang berdimensi tinggi. Pada ruang yang berdimensi ini, nantinya akan dicari pemisah atau yang sering disebut *hyperplane*.

*Hyperplane* dapat memaksimalkan jarak atau *margin* antara kelas data. *Hyperplane* terbaik antara kedua kelas dapat ditemukan dengan mengukur margin dan kemudian mencari titik maksimalnya. Usaha dalam mencari *hyperplane* yang terbaik sebagai pemisah kelas-kelas adalah inti dari proses pada metode SVM[23].

#### 1. Linear *Separable* Data

Metode SVM dengan *hyperplane* yang berbentuk garis lurus disebut dengan *linear separable*. Gambar 2.1 merupakan ilustrasi dari *hyperplane linear separable* data;




Gambar 2. 1 *Garis Linear Pemisah Dua Kelas*

Gambar 2.1 mengilustrasikan konsep dasar hyperplane pada algoritma Support Vector Machine. Hyperplane merupakan bidang pemisah berdimensi-n yang membagi ruang fitur menjadi dua wilayah yang masing-masing merepresentasikan kelas berbeda. Pada kasus dua dimensi sebagaimana ditunjukkan pada gambar, hyperplane berupa garis lurus yang memisahkan titik-titik data dari dua kelas. Posisi hyperplane ditentukan sedemikian rupa sehingga jarak antara hyperplane dengan titik data terdekat dari masing-masing kelas (yang disebut support vectors) menjadi maksimal, sehingga menghasilkan batas keputusan yang paling optimal dan memiliki kemampuan generalisasi terbaik terhadap data baru.

Dapat dilihat ilustrasi pada Gambar 2.1 adalah beberapa *pattern* yang merupakan anggota dari dua buah kelas yaitu kelas +1 dan kelas −1. Simbol untuk *pattern* pada kelas -1 adalah kotak yang berwarna merah, sedangkan simbol untuk

*pattern* pada kelas +1 adalah lingkaran dengan warna kuning. Dalam SVM yang telah disebutkan diatas menemukan garis (*hyperplane*) yang dapat memisahkan antara kedua kelompok tersebut. Berbagai macam garis pemisah *(discrimination boundaries) alternative* yang ditunjukkan pada gambar 2.1 bagian a dalam menemukan *hyperplane* yaitu dengan cara mengukur *Margin hyperplane* tersebut dan kemudian mencari titik maksimalnya. Jarak antara *hyperplane* dengan *pattern* pada masing-masing kelas biasa disebut dengan *margin*. Untuk *pattern* paling dekat disebut dengan *support vector*. Pada gambar 2.1 bagaian b garis yang berada di tengah menunjukkan *hyperplane* yang terbaik, karena terletak tepat pada tengah-tengah antar kelas, sedangkan *support vector* adalah titik merah dan kuning yang berada dalam lingkaran hitam. Usaha dalam mencari lokasi *hyperplane* ini merupakan proses inti dari SVM.

Pada penelitian ini data yang tersedia dapat dinotasikan sebagai $x \in$ Rd, sedangkan label untuk masing-masing kelas dinotasikan yi $\in \{-1, +1\}$ untuk i = 1,2,3, … n.

Pada kedua kelas dapat diasumsikan terpisah secara sempurna oleh hyperplane berdimensi d, yang didefinisikan sebagai berikut:

$$\vec{w}.\vec{x} + b = 0 \quad \text{..... (3.1)}$$

Untuk pattern $xi$ yang termasuk dalam kelas -1 dapat dirumuskan sebagai pattern yang memenuhi pertidaksamaan:

$$\vec{w}.\vec{x} + b \leq -1 \quad \text{..... (3.2)}$$

Sedangkan untuk pattern $xi$ yang termasuk kelas +1 dapat dirumuskan sebagai berikut:

$$\vec{w}.\vec{x} + b \leq +1 \quad \ldots (3.3)$$

Dimana:

$R^d$ = ruang vektor

d = dimensi

n = banyak data

$\vec{w}$ = vektor bobot

$\vec{x}$ = vektor data (input)

b = bias

Margin terbesar dapat diperoleh dengan cara memaksimalkan nilai jarak antara jarak dan titik terdekatnya, yaitu $\frac{1}{||w||}$. Hal ini dapat dirumuskan sebagai masalah Quadratic Programming (QP), yaitu mencari titik minimal persamaan 3.4, dengan memperlihatkan constraint persamaan 3.5.

$$min_w = \tau(w) = \frac{1}{2}||w||^2 \quad \ldots (3.4)$$

$$y_i(x_i.w + b) - 1 \geq 0 \quad \ldots (3.5)$$

Dimana ||w|| adalah vektor normal.

Salah satu teknis komputasi yaitu *lagrange multiplier* yang dapat memecahkan masalah ini dapat dinyatakan pada persamaan 3.6.

$$L(w, b, \alpha) = \frac{1}{2||w||^2} 2 - \sum \alpha_i \left( y_i \left( (\overrightarrow{X}.\overrightarrow{W} + b) - 1 \right) \right); i = 1,2 \ldots, n \quad \ldots (3.6)$$

Dimana α merupakan lagrange multiplier, yang bernilai $\alpha i \geq 0$. Nilai optimal pada persamaan 3.6 dapat dihitung dengan meminimalkan nilai L terhadap w dan

$b$, dan memaksimalkan nilai L terhadap $\alpha i$, dengan memperhatikan sifat bahwa pada titik optimal gradient L = 0. Pada persamaan 3.6 dapat dimodifikasi sebagai

maksimalisasi problem yang hanya mengandung $\alpha i$, seperti yang terlihat pada persamaan 3.7 dan persamaan 3.8 dibawah ini.

$\sum_i^1 = 1\ \alpha_i - \frac{1}{2}\sum_i^1 j = 1\ \alpha_i\ \alpha_j y_i y_j . \overrightarrow{x_i} . \vec{x}_i$ .....(3.7)

dimana $\alpha_i \geq 0$(i=1,2…,1) $\sum_i^1 = 1\ \alpha_i y_i = 0$.....(3.8)

Dengan demikian, maka akan diperoleh $\alpha i$ yang kebanyakan bernilai positif yang disebut sebagai support vector dan juga memperoleh persamaan 3.9 dan persamaan 3.10 sebagai solusi pemisah.

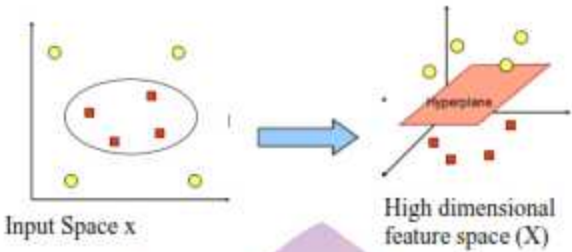
$w = \sum \alpha_i\ y_i x_i$ .....(3.9)

$b = \sum y_k - w^T x_k$ .....(3.10)

**2. NonLinear *Separable* Data**

Dalam dunia nyata (real world problem) pada umumnya masalah data yang diperoleh jarang yang bersifat linear, banyak yang bersifat non linear. Pada SVM terdapat sebuah fungsi kernel, yaitu fungsi yang digunakan untuk menyelesaikan problem non linear.

Kernel berfungsi memungkinkan untuk mengimplementasikan suatu model pada ruang dimensi lebih tinggi (ruang fitur).




Gambar 2. 2 Hyperplane

Gambar 2.2 menunjukkan konsep margin pada SVM, yaitu jarak antara hyperplane pemisah dengan support vectors terdekat dari masing-masing kelas. Margin yang lebih lebar mengindikasikan bahwa model memiliki kepercayaan yang lebih tinggi dalam memisahkan kedua kelas dan lebih robust terhadap variasi data baru. Support vectors yang ditandai pada gambar merupakan titik-titik data yang paling dekat dengan hyperplane dan menjadi penentu utama posisi serta orientasi hyperplane. Apabila support vectors berubah, maka posisi hyperplane juga akan berubah, sedangkan titik data lain yang bukan support vectors tidak memengaruhi posisi hyperplane.

a. Kernel Radian Basis Function (RBF)

$$K(\overrightarrow{X_i}, \overrightarrow{X_j}) = exp\left(\frac{\|\overrightarrow{X_i} - \overrightarrow{X_j}\|}{2\sigma^2}\right) \quad \text{.....(3.11)}$$

b. Kernel Sigmoid

$$K(\overrightarrow{X_i}, \overrightarrow{X_j}) = \tan(\alpha X_i^t X_j) \quad \text{.....(3.12)}$$

### 2.3.6. *Random Forest*

Random Forest pertama kali dikenalkan oleh Breiman pada Tahun 2001. Dalam penelitiannya menunjukkan kelebihan Random Forest antara lain dapat

menghasilkan error yang lebih rendah, memberikan hasil yang bagus dalam klasifikasi, dapat mengatasi data training dalam jumlah sangat besar secara efisien, dan metode yang efektif untuk mengestimasi missing data[24].

Secara sederhana, algoritma pembentukan RF dapat disebutkan sebagai berikut. Andaikan gugus data training yang kita miliki berukuran n dan terdiri atas d peubah penjelas (predictor). Tahapan penyusunan dan pendugaan menggunakan RF adalah:

a. (Tahapan bootstrap) tarik contoh acak dengan pemulihan berukuran n dari gugus data training.
b. (Tahapan random sub-setting) susun pohon berdasarkan data tersebut, namun pada setiap proses pemisahan pilih secara acak jumlah variable prediktor (m) < d peubah penjelas, dan lakukan pemisahan terbaik.
c. Ulangi langkah a-b sebanyak k kali sehingga diperoleh k buah pohon acak
d. Lakukan pendugaan gabungan berdasarkan k buah pohon tersebut (misal menggunakan majority vote untuk kasus klasifikasi, atau rata-rata untuk kasus regresi)

Perhatikan bahwa pada setiap kali pembentukan pohon, kandidat peubah penjelas yang digunakan untuk melakukan pemisahan bukanlah seluruh peubah yang terlibat namun hanya sebagian saja hasil pemilihan secara acak. Bisa dibayangkan bahwa proses ini menghasilkan kumpulan pohon tunggal dengan ukuran dan bentuk yang berbeda-beda. Hasil yang diharapkan adalah kumpulan pohon tunggal memiliki korelasi yang kecil antar pohonnya. Korelasi kecil ini

mengakibatkan ragam dugaan hasil RF menjadi kecil dan lebih kecil dibandingkan ragam dugaan hasil bagging[25].

Jika melihat secara seksama algoritma pembentukan RF, salah satu yang bisa diubah adalah nilai m, yaitu banyaknya peubah penjelas yang digunakan sebagai kandidat pemisah dalam pembentukan pohon. Nilai m yang semakin besar akan menyebabkan korelasi ($\rho$) semakin besar. Contoh ekstrim adalah jika kita gunakan m = d yang menyebabkan setiap kali pengulangan akan menghasilkan pohon yang sama sehingga nilai korelasi akan menjadi maksimum yaitu sebesar 1. Namun, jika nilai m kita buat sekecil mungkin yaitu hanya 1 peubah penjelas saja yang dijadikan kandidat pemisah, maka pohon yang diperoleh akan menjadi pohon dengan akurasi yang sangat rendah. Dengan demikian jelas bahwa pemilihan m memegang peranan dalam menentukan kebaikan RF yang dihasilkan.

Pohon keputusan dalam Random Forest dibentuk dengan menentukan kriteria pemisahan (splitting criterion) pada setiap node menggunakan Gini Impurity, yaitu metrik yang mengukur tingkat ketidakmurnian suatu himpunan kasus. Gini Impurity dirumuskan sebagai berikut.

$$Gini(Y) = 1 - \sum p(c|Y)^2 \quad (3.13)$$

Dimana Y adalah himpunan kasus pada suatu node, c adalah kelas target, dan p(c|Y) merupakan proporsi kasus pada node Y yang termasuk ke dalam kelas c. Nilai Gini Impurity berkisar antara 0 (node sepenuhnya murni, seluruh kasus berasal dari satu kelas) hingga mendekati 1 (distribusi kelas pada node tersebut sangat heterogen).

Pada setiap proses pemisahan node, algoritma memilih fitur dan nilai ambang (threshold) yang menghasilkan penurunan Gini Impurity terbesar antara node induk dan node anak, yang dirumuskan sebagai:

Gini = Gini(Y) - $\sum v \in \{kiri, kanan\}(|Yv|/Y)\ x\ Gini(Yv)$ ----------(3.14)

Dimana Yv adalah subset kasus pada node anak (kiri atau kanan) hasil pemisahan, dan |Yv|/|Y| adalah proporsi jumlah kasus pada node anak tersebut terhadap node induk. Fitur dengan nilai Gini terbesar pada setiap node dipilih sebagai dasar pemisahan, sehingga setiap pohon tumbuh secara rekursif hingga mencapai kondisi pemberhentian tertentu.

Setelah seluruh pohon keputusan dalam ensemble selesai dibangun melalui proses bootstrapping dan pemilihan fitur acak pada setiap node, tahap selanjutnya yang menjadi inti dari mekanisme Random Forest adalah menggabungkan prediksi dari seluruh pohon untuk menghasilkan keputusan klasifikasi akhir. Tahap agregasi ini merupakan hal yang membedakan Random Forest dari model pohon keputusan tunggal, karena prediksi akhir tidak bergantung pada satu pohon yang mungkin memiliki bias atau varians tinggi, melainkan merupakan hasil konsensus dari seluruh pohon yang masing-masing dibangun dari perspektif data dan fitur yang berbeda.

Pada permasalahan klasifikasi, Random Forest menggunakan mekanisme majority voting sebagai strategi agregasi prediksi. Dalam mekanisme ini, setiap pohon keputusan yang telah dibangun dalam ensemble memberikan satu suara berupa prediksi kelas untuk setiap observasi baru yang dimasukkan ke dalam model. Setelah seluruh pohon memberikan prediksinya, kelas yang memperoleh

jumlah suara terbanyak dari seluruh pohon ditetapkan sebagai prediksi final model Random Forest. Mekanisme majority voting ini berbeda dengan pendekatan rata-rata aritmatika yang digunakan pada permasalahan regresi, karena output klasifikasi bersifat kategorikal sehingga pemungutan suara mayoritas merupakan pendekatan yang paling sesuai dan intuitif. Secara matematis, mekanisme majority voting pada Random Forest dirumuskan sebagai berikut:

$$\hat{y} = \text{argmax}_c \sum_{t=1}^{T} I(h_t(x) = c) \quad \textbf{(3.15)}$$

Dimana $\hat{y}$ adalah kelas prediksi akhir yang dihasilkan oleh model Random Forest untuk suatu observasi x yang diklasifikasikan, T adalah jumlah total pohon keputusan yang dibangun dalam ensemble yang nilainya ditentukan oleh hyperparameter n_estimators, $h_t(x)$ adalah prediksi kelas yang dihasilkan oleh pohon keputusan ke-t untuk observasi x berdasarkan aturan pemisahan yang telah dipelajari dari subset data training, c adalah salah satu kelas target yang tersedia dalam dataset klasifikasi, dan $I(\cdot)$ adalah fungsi indikator yang bernilai 1 apabila prediksi pohon ke-t sesuai dengan kelas c dan bernilai 0 apabila tidak sesuai. Fungsi $\text{argmax}_c$ beroperasi dengan menjumlahkan seluruh nilai fungsi indikator untuk setiap kelas c yang tersedia, kemudian memilih kelas yang menghasilkan jumlah total suara terbesar dari seluruh T pohon sebagai prediksi akhir model.

Sebagai ilustrasi, pada penelitian ini yang menggunakan tiga kelas target yaitu Aktif, Berisiko, dan Tidak Aktif, apabila dari 200 pohon keputusan terdapat 150 pohon yang memprediksi suatu observasi sebagai kelas Aktif, 40 pohon memprediksi Berisiko, dan 10 pohon memprediksi Tidak Aktif, maka prediksi akhir model Random Forest untuk observasi tersebut adalah kelas Aktif karena

memperoleh suara terbanyak. Mekanisme ini memastikan bahwa keputusan klasifikasi final bersifat demokratis dan tidak didominasi oleh satu pohon keputusan tunggal yang mungkin mengalami overfitting terhadap subset data tertentu.

Kekuatan utama mekanisme majority voting terletak pada prinsip yang dikenal sebagai wisdom of crowds, di mana prediksi kolektif dari sejumlah besar model yang masing-masing memiliki perspektif berbeda karena dibangun dari subset data dan fitur yang berbeda cenderung menghasilkan keputusan yang lebih akurat dan stabil dibandingkan prediksi dari satu model tunggal. Setiap pohon keputusan secara individual mungkin memiliki kesalahan prediksi pada observasi-observasi tertentu, namun karena kesalahan tersebut bersifat acak dan tidak terkorelasi antar pohon berkat mekanisme bootstrapping dan pemilihan fitur acak, maka ketika prediksi dari seluruh pohon diagregasikan melalui voting, kesalahan individual cenderung saling mengeliminasi dan menghasilkan prediksi akhir yang lebih robust.

Semakin besar jumlah pohon (T) yang digunakan dalam ensemble, semakin stabil dan reliabel prediksi yang dihasilkan oleh model Random Forest karena variabilitas hasil voting semakin berkurang seiring bertambahnya jumlah suara. Namun demikian, peningkatan jumlah pohon di atas ambang tertentu akan memberikan perbaikan performa yang semakin marginal karena mayoritas pola dalam data sudah berhasil ditangkap oleh pohon-pohon yang ada, sementara biaya komputasi baik dalam hal waktu pelatihan maupun penggunaan memori meningkat secara linear seiring bertambahnya jumlah pohon. Oleh karena itu, pemilihan jumlah pohon yang optimal perlu mempertimbangkan keseimbangan antara

performa prediksi dan efisiensi komputasi, yang dalam penelitian ini ditentukan melalui proses optimasi hyperparameter menggunakan Grid Search.

### 2.3.7. *Confusion Matrix*

Berikut ini merupakan hasil dari confusion matrix[26]

Tabel 2. 5 Tabel Confusion Matrix

| Prediksi | aktual | |
|---|---|---|
| | True | False |
| True | TP | FN |
| False | FP | TN |

Keterangan:

TP = Jumlah prediksi yang tepat bersifat positif (True Positive).

TN = jumlah prediksi yang tepat bersifat negatif (True Negative).

FP = jumlah prediksi yang salah bersifat positif (False Positive).

FN = jumlah prediksi yang salah bersifat negatif (False Negative).

Accuracy merupakan proporsi jumlah prediksi benar. Rumus akurasi adalah:

$$Akaurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad (3.15)$$

### 2.3.8. F1-Score dan Recall

Selain akurasi, terdapat metrik evaluasi lain yang lebih informatif terutama pada kondisi data yang tidak seimbang (imbalanced data). Metrik-metrik tersebut juga diturunkan dari nilai TP, TN, FP, dan FN pada confusion matrix[27].

Recall (juga dikenal sebagai Sensitivity) mengukur seberapa baik model dalam mendeteksi seluruh data yang benar-benar positif. Recall didefinisikan sebagai proporsi prediksi positif yang benar terhadap seluruh data yang sesungguhnya positif [R17]. Rumus Recall adalah sebagai berikut:

$$Recall = \frac{TP}{TP+FN} \quad \text{.....(3.16)}$$

Precision mengukur ketepatan prediksi positif yang dilakukan model, yaitu seberapa besar proporsi prediksi positif yang benar-benar positif dari seluruh prediksi positif yang dihasilkan model. Rumus Precision adalah:

$$Precision = \frac{TP}{TP+FP} \quad \text{.....(3.17)}$$

Recall dan Precision memiliki hubungan yang saling berlawanan (trade-off), di mana peningkatan salah satu nilai cenderung menurunkan nilai yang lain. Untuk menyeimbangkan kedua metrik tersebut, digunakan F1-Score yang merupakan rata-rata harmonik (harmonic mean) antara Precision dan Recall. F1-Score memberikan gambaran menyeluruh tentang performa model, terutama ketika distribusi kelas tidak seimbang. Rumus F1-Score adalah sebagai berikut:

$$F1 - Scoree = 2 \times \frac{Precision \times Recall}{Precision+Recall} = \frac{2TP}{2TP+FP+FN} \quad \text{.....(3.18)}$$

Nilai F1-Score berkisar antara 0 hingga 1, di mana nilai 1 menunjukkan performa model yang sempurna dan nilai 0 menunjukkan performa terburuk. Dalam penelitian prediksi retensi mahasiswa yang memiliki distribusi kelas tidak seimbang antara mahasiswa yang bertahan dan yang keluar, penggunaan F1-Score lebih disarankan dibandingkan akurasi semata karena mampu mengukur performa model secara lebih adil terhadap kedua kelas.

Normalisasi data merupakan salah satu tahap pra-pemrosesan yang fundamental dalam pipeline machine learning yang bertujuan mentransformasikan nilai fitur numerik ke dalam skala yang seragam agar seluruh fitur memiliki

kontribusi yang proporsional dalam proses pembelajaran model. Tahap ini menjadi sangat krusial khususnya bagi algoritma yang berbasis perhitungan jarak atau optimasi numerik seperti Support Vector Machine, di mana perbedaan skala antar fitur dapat memberikan pengaruh yang tidak proporsional terhadap proses pembentukan hyperplane pemisah dan menghasilkan model yang tidak optimal. Permasalahan ini muncul karena algoritma berbasis jarak menghitung kedekatan antar titik data menggunakan metrik jarak seperti jarak Euclidean, yang secara matematis memberikan bobot lebih besar kepada fitur dengan rentang nilai yang lebih lebar.

Sebagai ilustrasi konkret dalam konteks penelitian ini, apabila sebuah dataset memiliki fitur Total SKS dengan rentang nilai 0 hingga 150 dan fitur IPK Kumulatif dengan rentang 0 hingga 4, maka selisih 10 satuan pada fitur Total SKS (misalnya dari 100 menjadi 110) akan memberikan kontribusi yang jauh lebih besar dalam perhitungan jarak Euclidean dibandingkan selisih 1 satuan pada fitur IPK Kumulatif (misalnya dari 2,5 menjadi 3,5), meskipun secara substansial perubahan 1 poin IPK mungkin memiliki signifikansi yang lebih tinggi terhadap status retensi mahasiswa. Tanpa normalisasi, kernel Radial Basis Function (RBF) pada SVM yang menghitung jarak antar titik data dalam ruang fitur akan didominasi oleh fitur berskala besar, sehingga variasi pada fitur berskala kecil yang sesungguhnya informatif menjadi terabaikan dan hyperplane pemisah yang dihasilkan tidak mencerminkan kontribusi sebenarnya dari masing-masing fitur.

Terdapat beberapa teknik normalisasi yang umum digunakan dalam penelitian machine learning, masing-masing dengan karakteristik dan keunggulan

yang berbeda. Z-Score Standardization mentransformasikan data sehingga memiliki rata-rata 0 dan standar deviasi 1, yang cocok untuk data dengan distribusi mendekati normal. Decimal Scaling menormalisasi data dengan membagi setiap nilai dengan pangkat 10 tertentu. Dalam penelitian ini, teknik yang dipilih adalah Min-Max Normalization yang mentransformasikan setiap nilai fitur ke dalam rentang [0, 1] dengan mempertahankan hubungan proporsional antar nilai aslinya. Pemilihan teknik Min-Max Normalization didasarkan pada beberapa pertimbangan, yaitu menghasilkan rentang nilai yang terbatas dan konsisten [0, 1] untuk seluruh fitur sehingga memudahkan interpretasi, tidak mengasumsikan distribusi data tertentu sehingga cocok diterapkan pada dataset dengan distribusi yang beragam, serta mempertahankan hubungan antar nilai asli tanpa mengubah bentuk distribusi data. Rumus Min-Max Normalization adalah sebagai berikut:

$$x' = (x - x\_min) / (x\_max - x\_min) \quad \textbf{(3.21)}$$

Dimana x' adalah nilai fitur setelah dinormalisasi yang berada dalam rentang [0, 1], x adalah nilai fitur asli sebelum proses normalisasi dilakukan, x_min adalah nilai minimum fitur yang diperoleh dari data training yang merepresentasikan batas bawah rentang fitur, dan x_max adalah nilai maksimum fitur yang juga diperoleh dari data training yang merepresentasikan batas atas rentang fitur. Hasil transformasi ini menjadikan nilai minimum fitur dalam data training menjadi 0 (karena x − x_min = 0) dan nilai maksimum menjadi 1 (karena x − x_min = x_max − x_min), sedangkan nilai-nilai di antaranya tersebar secara proporsional sesuai posisi relatifnya terhadap rentang asli. Sebagai contoh, apabila fitur Total SKS memiliki x_min = 0 dan x_max = 150, maka mahasiswa dengan Total SKS = 75

akan memiliki nilai normalisasi sebesar (75 – 0) / (150 – 0) = 0,5, yang secara tepat mencerminkan posisi tengah dalam rentang fitur tersebut.

Aspek penting yang harus diperhatikan dalam penerapan Min-Max Normalization adalah bahwa parameter x_min dan x_max harus dihitung secara eksklusif dari data training saja, bukan dari keseluruhan dataset yang mencakup data training dan data testing secara bersamaan. Hal ini bertujuan mencegah terjadinya data leakage, yaitu kondisi di mana informasi dari data testing yang seharusnya tidak diketahui oleh model selama proses pelatihan bocor ke dalam proses transformasi fitur. Apabila x_min dan x_max dihitung dari keseluruhan dataset, maka proses normalisasi akan menggunakan informasi statistik yang berasal dari data testing, sehingga model secara tidak langsung telah melihat karakteristik data testing sebelum tahap evaluasi dan menyebabkan estimasi performa model menjadi terlalu optimistis. Pada tahap pengujian, data testing dinormalisasi menggunakan parameter x_min dan x_max yang sama dari data training untuk menjaga konsistensi transformasi. Konsekuensi dari pendekatan ini adalah bahwa nilai fitur pada data testing setelah normalisasi mungkin sedikit keluar dari rentang [0, 1] apabila terdapat nilai yang lebih kecil dari x_min atau lebih besar dari x_max pada data training, namun hal ini merupakan kondisi yang wajar dan tidak memengaruhi proses klasifikasi secara signifikan.

Perlu dicatat bahwa tidak semua algoritma klasifikasi memerlukan tahap normalisasi data. Algoritma berbasis pohon keputusan seperti Random Forest melakukan pemisahan node berdasarkan nilai threshold pada setiap fitur secara independen tanpa melibatkan perhitungan jarak antar titik data dalam ruang fitur.

Pada setiap node, algoritma hanya membandingkan nilai fitur tertentu dengan threshold pemisahan untuk menentukan apakah suatu observasi masuk ke cabang kiri atau cabang kanan, sehingga perbedaan skala antar fitur tidak memengaruhi proses pembentukan pohon maupun hasil klasifikasi akhir. Oleh karena itu, dalam penelitian ini normalisasi Min-Max hanya diterapkan pada data yang digunakan untuk melatih dan menguji model SVM, sedangkan model Random Forest menggunakan data tanpa normalisasi.

# BAB 3

# METODE PENELITIAN

## 3.1.Jenis, Sifat, dan Pendekatan Penelitian

Penelitian ini dilaksanakan di Universitas Ibnu Sina (UIS), sebuah perguruan tinggi swasta yang berlokasi di Kota Batam, Provinsi Kepulauan Riau. Pemilihan Universitas Ibnu Sina sebagai objek penelitian didasarkan pada beberapa pertimbangan yang bersifat strategis dan metodologis. Pertama, UIS memiliki karakteristik mahasiswa yang heterogen, dengan latar belakang sosial-ekonomi yang beragam sebagian besar berasal dari keluarga pekerja industri di kawasan Batam dan sekitarnyasehingga menjadikannya representasi yang relevan bagi perguruan tinggi swasta di kota industri Indonesia yang rentan terhadap permasalahan retensi multidimensional. Kedua, UIS memiliki enam program studi yang tergabung dalam tiga fakultas berbeda, yaitu Fakultas Ekonomi dan Bisnis (FEB), Fakultas Sains dan Teknologi (FST) dan Fakultas Ilmu Kesehatan (FIKes), dengan total 2.389 mahasiswa angkatan 2021/2022–2023/2024. Keragaman program studi ini memungkinkan analisis perbandingan pola retensi lintas bidang ilmu dalam satu institusi yang sama, sehingga menghasilkan temuan yang lebih komprehensif. Ketiga, Universitas Ibnu Sina memiliki Sistem Informasi Akademik (SIAKAD) yang telah mencatat data akademik dan administratif mahasiswa secara terstruktur, mencakup variabel akademik seperti IPK, IPS per semester, dan jumlah SKS, serta variabel non-akademik seperti pekerjaan orang tua, penghasilan keluarga, dan riwayat cuti. Ketersediaan dan aksesibilitas data primer ini merupakan prasyarat utama dalam pembangunan dataset retensi yang komprehensif sebagaimana menjadi tujuan penelitian ini. Keempat, berdasarkan data administrasi

akademik UIS, terdapat variasi tingkat retensi yang signifikan antarprogram studi, yang mengindikasikan adanya dinamika dan pola risiko putus studi yang nyata dan mendesak untuk dianalisis secara ilmiah guna mendukung pengambilan keputusan akademik institusi.

Penelitian ini termasuk penelitian kuantitatif dengan sifat deskriptif-analitik, yang bertujuan untuk menganalisis data akademik mahasiswa Universitas Ibnu Sina dan mengevaluasi performa algoritma Support Vector Machine (SVM) serta Random Forest (RF) dalam memprediksi retensi mahasiswa. Penelitian ini dikategorikan sebagai penelitian kuantitatif karena menggunakan data numerik dari variabel akademik dan non-akademik mahasiswa yang dapat diukur secara objektif dan dianalisis menggunakan metode statistik serta algoritma klasifikasi. Pendekatan deskriptif digunakan untuk menjelaskan karakteristik dan distribusi data mahasiswa, sedangkan pendekatan analitik diterapkan untuk menguji dan membandingkan kinerja kedua algoritma dalam memprediksi status retensi mahasiswa pada tiga kategori: Aktif, Berisiko, dan Tidak Aktif.

Penelitian ini bersifat eksplanatif karena tidak hanya mendeskripsikan pola data, tetapi juga menjelaskan hubungan antara variabel-variabel akademik dan non-akademik dengan kemungkinan mahasiswa bertahan atau putus studi. Dengan pendekatan ini, penelitian termasuk dalam kategori penelitian kuantitatif deskriptif-analitik dengan penerapan metode klasifikasi berbasis machine learning, yang hasilnya dapat digunakan untuk memberikan rekomendasi kebijakan berbasis data bagi Universitas Ibnu Sina dalam upaya meningkatkan angka retensi dan mengurangi risiko putus studi mahasiswanya.

### 3.3.1 Metode Pengumpulan Data

Variabel penelitian yang digunakan dalam penelitian ini adalah sebanyak 18 variabel. Berikut ini adalah tabel berisi variabel penelitian dan penjelasan definisi operasional masing-masing variabelnya;

Tabel 3. 1 Variabel Penelitian dan Penjelasan Operasional

| Variabel | Definisi Penjelasan Operasional | Tipe Data |
|---|---|---|
| IPS Semester 1 | Indeks Prestasi Semester mahasiswa pada semester 1, digunakan untuk menilai perkembangan akademik awal mahasiswa. | Numerik |
| IPS Semester 2 | Indeks Prestasi Semester mahasiswa pada semester 2, sebagai kelanjutan penilaian akademik. | Numerik |
| IPS Semester 3 | Indeks Prestasi Semester mahasiswa pada semester 3, digunakan untuk menganalisis konsistensi akademik. | Numerik |
| IPS Semester 4 | Indeks Prestasi Semester mahasiswa pada semester 4, sebagai indikator stabilitas akademik mahasiswa. | Numerik |
| IPK Sementara | Indeks Prestasi Kumulatif sementara mahasiswa selama beberapa semester awal, digunakan untuk menilai pengaruh prestasi akademik terhadap retensi. | Numerik |
| Jml SKS Gagal | Jumlah SKS yang tidak lulus atau gagal oleh mahasiswa, digunakan sebagai indikator kesulitan akademik. | Numerik |

| Variabel | Definisi Penjelasan Operasional | Tipe Data |
|---|---|---|
| Jml MK Mengulang | Jumlah mata kuliah yang diulang oleh mahasiswa, mencerminkan tingkat kesulitan dalam menyelesaikan studi. | Numerik |
| Jml Nilai D & E | Jumlah mata kuliah yang memperoleh nilai D dan E, sebagai indikator performa akademik rendah mahasiswa. | Numerik |
| Jenis Kelamin | Jenis kelamin mahasiswa, yaitu: laki-laki dan perempuan. | Kategorikal |
| Asal Sekolah | Jenis sekolah asal mahasiswa sebelum masuk universitas (Negeri/Swasta), digunakan untuk menganalisis pengaruh kualitas pendidikan sebelumnya terhadap retensi mahasiswa. | Kategorikal |
| Usia Mendaftar | Usia mahasiswa saat pertama kali mendaftar ke universitas, digunakan untuk menganalisis pengaruh usia terhadap retensi. | Numerik |
| Asal Kota | Asal kota mahasiswa, yaitu: Lokal atau Luar Pulau, digunakan untuk menganalisis pengaruh jarak geografis terhadap retensi. | Kategorikal |
| Pekerjaan OT | Pekerjaan orang tua mahasiswa, yaitu: PNS, Wiraswasta, Pegawai Swasta, Pensiun, Petani, dan Lain-lain. | Kategorikal |

| Variabel | Definisi Penjelasan Operasional | Tipe Data |
|---|---|---|
| Penghasilan OT | Tingkat penghasilan orang tua mahasiswa alumni dalam 5 tingkat ordinal, digunakan untuk menganalisis dampak kondisi ekonomi keluarga terhadap kemampuan mahasiswa bertahan di kuliah. | Ordinal |
| Jumlah Cuti | Jumlah semester cuti yang pernah diambil oleh mahasiswa selama masa studi. | Numerik |
| Program Studi | Program studi yang diambil mahasiswa, yaitu: Manajemen, Akuntansi, Teknik Industri, Teknik Informatika, K3, dan Kesling (6 prodi). | Kategorikal |
| Periode Masuk | Tahun angkatan masuk mahasiswa ke universitas (3 angkatan), digunakan untuk menganalisis tren retensi per angkatan. | Kategorikal |
| Sisa Masa Studi | Sisa masa studi mahasiswa yang tersisa dalam satuan semester, digunakan untuk menilai posisi mahasiswa dalam perjalanan studinya. | Numerik |

### 3.2.Metode Analisis Data

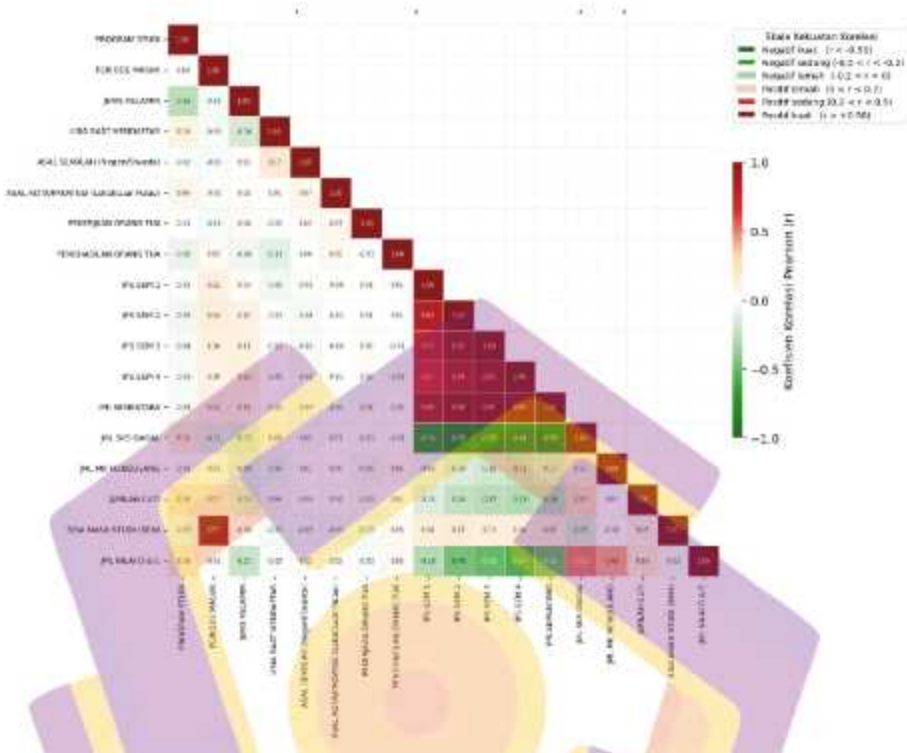
Penelitian ini menggunakan analisis deskriptif untuk mengetahui gambaran umum data dan perbandingan hasil analisis klasifikasi Support Vector Machine (SVM) dan Random Forest. Optimasi hyperparameter kedua algoritma dilakukan menggunakan metode Grid Search yang dikombinasikan dengan Stratified 5-Fold Cross Validation pada data training.

Untuk metode SVM, peneliti melakukan pencarian pada kernel Linear, Radial Basis Function (RBF), Polynomial, dan Sigmoid, dengan parameter regularisasi C = 0,1; 1; 10; 100; 1.000 dan parameter gamma = 0,001; 0,01; 0,1; 1; 'scale'; 'auto'. Untuk metode Random Forest, peneliti melakukan pencarian pada jumlah pohon (n_estimators) sebanyak 50, 100, 200, 300, dan 500, kedalaman maksimum pohon (max_depth) sebesar 5, 10, 15, 20, dan None, serta parameter max_features, min_samples_split, min_samples_leaf, dan class_weight.

### 3.2.1 Analisis Korelasi Antar Fitur Prediktor

Sebelum dilakukan pemodelan, analisis korelasi Pearson dilakukan terhadap seluruh 18 variabel prediktor untuk mengetahui hubungan linear antar fitur. Analisis ini bertujuan mengidentifikasi pasangan fitur yang memiliki korelasi kuat, mendeteksi potensi multikolinearitas, serta memahami pola hubungan antar variabel akademik, demografis, sosial-ekonomi, dan administratif sebelum dimasukkan ke dalam model klasifikasi.

Hasil analisis korelasi divisualisasikan dalam bentuk heatmap korelasi Pearson sebagaimana ditampilkan pada Gambar 3.1. Heatmap tersebut menampilkan matriks korelasi antar fitur prediktor menggunakan skema warna dengan warna merah gelap menunjukkan korelasi positif kuat ($r > +0,50$), warna putih menunjukkan tidak ada korelasi ($r \approx 0$), dan warna hijau gelap menunjukkan korelasi negatif kuat ($r < -0,50$).




Gambar 3. 1 Heatmap Korelasi Pearson Antar 18 Fitur Prediktor

Berdasarkan heatmap pada Gambar 3.1, terdapat beberapa temuan penting mengenai korelasi antar fitur. Variabel akademik menunjukkan korelasi positif yang kuat satu sama lain, khususnya antara IPS Semester 1 hingga IPS Semester 4 dan IPK Sementara dengan nilai korelasi berkisar antara r = 0,67 hingga r = 0,97. Hal ini mengindikasikan bahwa performa akademik per semester memiliki konsistensi yang tinggi dan saling berkaitan erat.

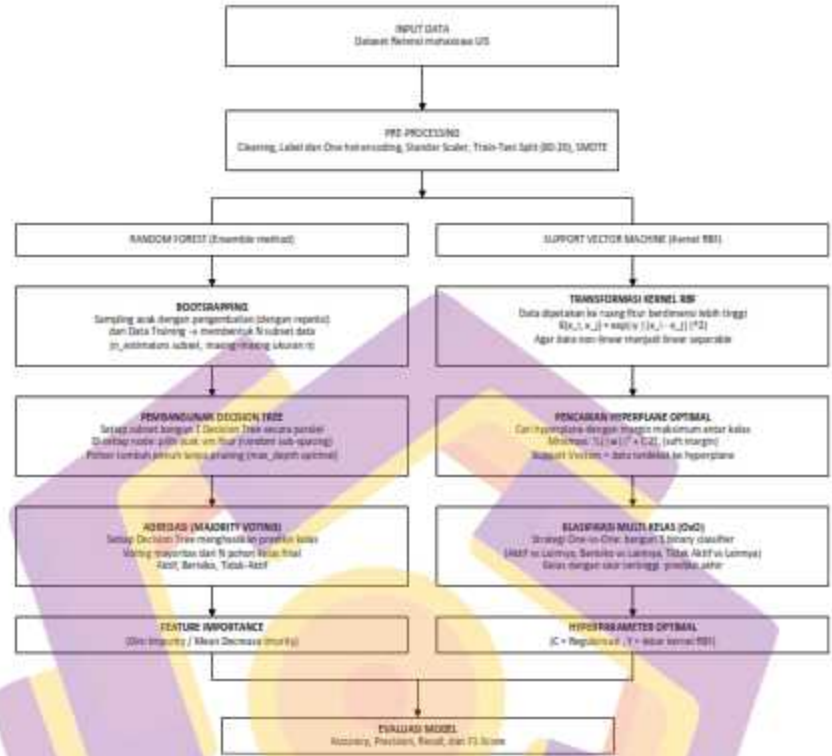
Sebaliknya, variabel Jumlah SKS Gagal dan Jumlah Nilai D & E menunjukkan korelasi negatif yang kuat terhadap variabel IPS dan IPK Sementara, dengan nilai korelasi berkisar antara r = -0,74 hingga r = -0,89. Artinya, semakin

rendah performa akademik (IPS/IPK) maka semakin tinggi jumlah SKS gagal dan nilai rendah yang diperoleh mahasiswa. Sementara itu, variabel demografis dan sosial-ekonomi seperti Jenis Kelamin, Asal Sekolah, Pekerjaan Orang Tua, dan Penghasilan Orang Tua menunjukkan korelasi yang lemah terhadap variabel akademik ($r < 0,20$), yang mengindikasikan bahwa faktor non-akademik bersifat independen terhadap performa akademik dan berfungsi sebagai prediktor komplementer dalam model klasifikasi retensi.

Potensi multikolinearitas teridentifikasi pada kelompok variabel IPS antar semester dan IPK Sementara yang memiliki korelasi sangat tinggi ($|r| > 0,70$). Meskipun demikian, seluruh variabel tersebut tetap dipertahankan dalam analisis karena masing-masing variabel merepresentasikan informasi per periode yang berbeda dan memiliki relevansi teoritis yang kuat terhadap prediksi retensi mahasiswa. Algoritma Random Forest dan SVM yang digunakan bersifat robust terhadap multikolinearitas sehingga tidak mengganggu performa model.

### 3.2.2 Arsitektur Proses Algoritma Klasifikasi

Setelah analisis korelasi antar fitur dilakukan pada sub-bab 3.3.1, tahap selanjutnya adalah merancang arsitektur proses algoritma klasifikasi yang akan digunakan. Untuk memberikan pemahaman yang komprehensif mengenai mekanisme kerja kedua algoritma, Gambar 3.2 menyajikan diagram arsitektur proses Random Forest dan Support Vector Machine (SVM) secara berdampingan. Diagram ini menggambarkan secara teknis bagaimana masing-masing algoritma memproses data input mulai dari tahap pra-pemrosesan hingga menghasilkan prediksi kelas retensi mahasiswa.




Gambar 3. 2 Diagram Arsitektur Proses Random Forest dan SVM untuk Prediksi Retensi Mahasiswa

Gambar 3.2 menunjukkan bahwa kedua algoritma memanfaatkan data yang telah melalui tahap pra-pemrosesan yang sama, mencakup data cleaning, encoding variabel kategorik menggunakan Label Encoding dan One-Hot Encoding, normalisasi fitur numerik menggunakan Standard Scaler, pembagian data dengan proporsi 80% training dan 20% testing melalui Stratified Sampling, serta penanganan ketidakseimbangan kelas menggunakan SMOTE yang diterapkan secara eksklusif pada data training. Setelah pra-pemrosesan selesai, data training diproses secara paralel oleh dua jalur algoritma yang berbeda secara fundamental.

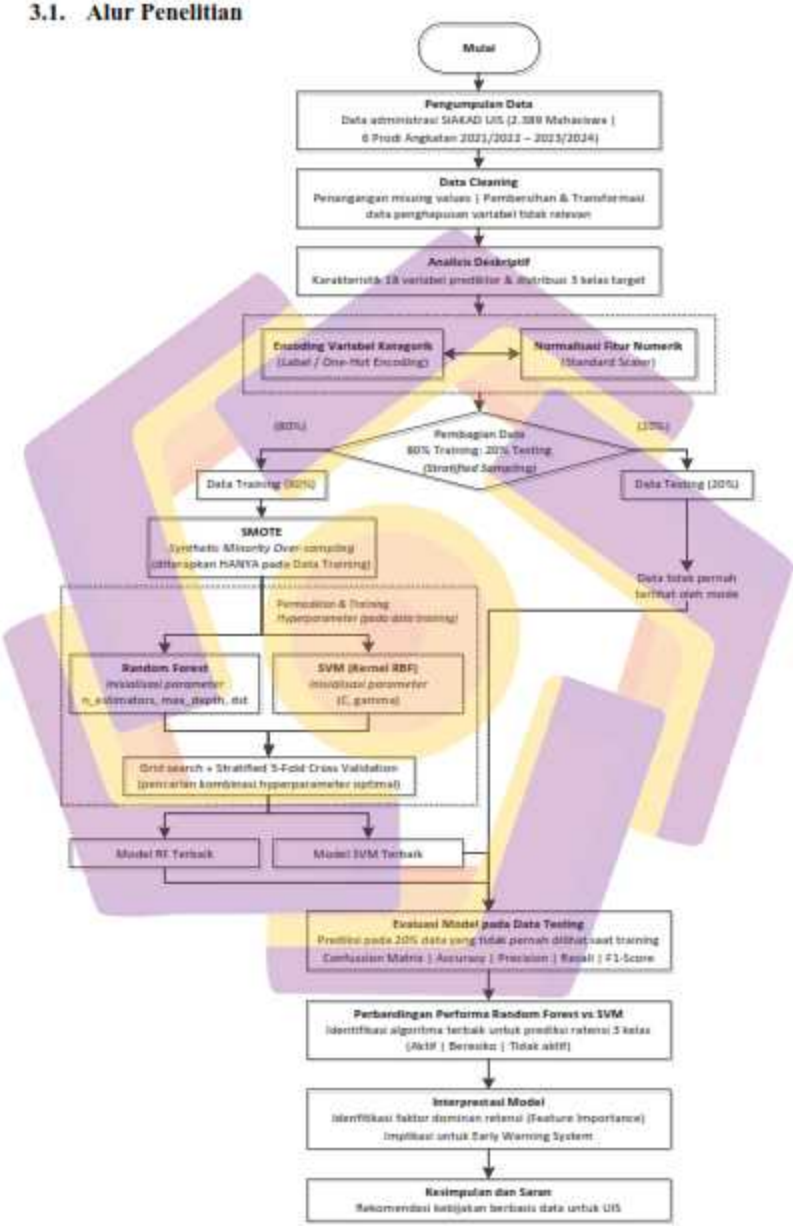
Pada jalur Random Forest, proses dimulai dengan tahap bootstrapping, yaitu pengambilan sampel acak dengan pengembalian dari data training untuk membentuk sejumlah n_estimators subset data. Setiap subset kemudian digunakan untuk membangun satu Decision Tree secara independen dan paralel, di mana pada setiap node pemisahan dipilih secara acak sejumlah akar kuadrat dari total fitur yang tersedia (random sub-spacing). Pendekatan ini bertujuan mengurangi korelasi antar pohon sehingga menghasilkan model ensemble yang lebih robust dan mampu mengatasi overfitting. Prediksi akhir ditentukan melalui mekanisme majority voting dari seluruh pohon yang telah dibangun. Di samping prediksi kelas, Random Forest juga menghasilkan nilai Feature Importance berdasarkan Gini Impurity yang digunakan untuk mengidentifikasi faktor-faktor dominan yang memengaruhi retensi mahasiswa.

Pada jalur Support Vector Machine (SVM), proses diawali dengan transformasi data menggunakan fungsi kernel Radial Basis Function (RBF) dengan formula $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$, yang memetakan data ke ruang fitur berdimensi lebih tinggi sehingga data yang tidak dapat dipisahkan secara linear menjadi linear separable. Selanjutnya algoritma mencari hyperplane optimal yang memaksimalkan margin antarkelas dengan meminimalkan fungsi objektif $\frac{1}{2}\|w\|^2 + C \cdot \Sigma\xi_i$, di mana parameter C mengontrol keseimbangan antara margin maksimum dan toleransi terhadap kesalahan klasifikasi (soft margin). Karena penelitian ini merupakan permasalahan klasifikasi tiga kelas (Aktif, Berisiko, dan Tidak Aktif), SVM menerapkan strategi One-vs-One (OvO) dengan membangun tiga buah binary classifier yang masing-masing memisahkan sepasang kelas (Aktif vs Berisiko, Aktif vs Tidak Aktif, dan Berisiko vs Tidak Aktif). Prediksi akhir ditentukan

melalui mekanisme voting mayoritas (max-wins voting) dari ketiga binary classifier tersebut, di mana kelas yang paling banyak memenangkan perbandingan berpasangan ditetapkan sebagai kelas prediksi final.

Kedua model melalui proses optimasi hyperparameter menggunakan metode Grid Search yang dikombinasikan dengan Stratified 5-Fold Cross Validation pada data training. Pendekatan ini memastikan pencarian hyperparameter optimal berlangsung secara sistematis dan menghasilkan estimasi performa model yang reliabel. Hasil dari kedua jalur tersebut kemudian dibandingkan secara komprehensif berdasarkan metrik Accuracy, Precision, Recall, dan F1-Score (Macro) pada data testing (20%) yang tidak pernah digunakan selama proses pelatihan maupun tuning hyperparameter. Hasil perbandingan ini menjadi dasar penentuan algoritma terbaik yang selanjutnya digunakan sebagai pondasi pengembangan sistem peringatan dini (Early Warning System) retensi mahasiswa di Universitas Ibnu Sina

## 3.1. Alur Penelitian




Gambar 3.3 Diagram Alur Penelitian

Gambar 3.1 menyajikan diagram alur penelitian secara menyeluruh, mulai dari tahap pengumpulan data hingga penarikan kesimpulan. Tahap awal penelitian adalah pengumpulan data administratif mahasiswa Universitas Ibnu Sina yang diperoleh melalui Sistem Informasi Akademik (SIAKAD). Dataset terdiri dari 2.389 mahasiswa yang berasal dari 6 program studi pada 3 fakultas, mencakup angkatan tahun akademik 2021/2022 hingga 2023/2024, dengan variabel yang meliputi faktor akademik, demografis, dan sosial-ekonomi yang relevan dengan keputusan retensi mahasiswa.

Setelah dataset terkumpul, dilakukan tahap data cleaning yang meliputi pemeriksaan dan penanganan missing values, pembersihan serta transformasi data, dan penghapusan variabel yang tidak relevan dengan tujuan penelitian. Tahap ini bertujuan menjamin kualitas dan konsistensi data sebelum dilakukan analisis lebih lanjut. Selanjutnya dilakukan analisis deskriptif untuk memberikan gambaran umum mengenai karakteristik 18 variabel prediktor serta distribusi tiga kelas target, yaitu "Aktif", "Berisiko", dan "Tidak Aktif". Analisis deskriptif juga berfungsi mengidentifikasi pola awal data dan tingkat ketidakseimbangan kelas yang perlu ditangani pada tahap berikutnya.

Pada tahap pra-pemrosesan data dilakukan dua proses transformasi, yaitu encoding variabel kategorik menggunakan teknik Label Encoding atau One-Hot Encoding agar variabel non-numerik dapat diproses oleh algoritma machine learning, dan normalisasi fitur numerik menggunakan Standard Scaler untuk menyeragamkan skala antarvariabel. Tahap normalisasi sangat krusial khususnya bagi algoritma Support Vector Machine yang sensitif terhadap perbedaan skala antarfitur.

Dataset yang telah melalui pra-pemrosesan selanjutnya dibagi menggunakan teknik Stratified Sampling dengan proporsi 80% untuk data training dan 20% untuk data testing. Penggunaan stratifikasi bertujuan memastikan proporsi setiap kelas target terjaga pada kedua subset, sehingga representasi ketiga kelas tetap proporsional baik pada tahap pelatihan maupun pengujian model. Mengingat distribusi kelas yang sangat tidak seimbang, dengan kelas Aktif sebesar 75,0%, kelas Berisiko 21,9%, dan kelas Tidak Aktif hanya 3,2%, maka teknik Synthetic Minority Over-sampling Technique (SMOTE) diterapkan secara eksklusif pada data training. Penerapan SMOTE hanya pada data training bertujuan mencegah terjadinya data leakage, sedangkan data testing tetap dipertahankan dalam distribusi aslinya agar mampu merefleksikan kondisi data pada situasi nyata.

Tahap pemodelan dilakukan secara paralel menggunakan dua algoritma klasifikasi, yaitu Random Forest dan Support Vector Machine. Kedua algoritma ini dipilih berdasarkan pertimbangan yang telah diuraikan pada Bab I, di mana Random Forest merepresentasikan pendekatan ensemble berbasis pohon keputusan yang robust terhadap noise dan overfitting, sedangkan Support Vector Machine merepresentasikan pendekatan berbasis optimasi margin yang unggul dalam menangani data berdimensi tinggi. Penggunaan kedua algoritma secara bersamaan memungkinkan perbandingan performa antara dua paradigma machine learning yang berbeda secara fundamental dalam menangani permasalahan klasifikasi retensi mahasiswa.

Untuk memperoleh kombinasi hyperparameter optimal pada masing-masing algoritma, digunakan metode Grid Search yang dipadukan dengan Stratified 5-Fold Cross Validation. Grid Search bekerja dengan mengevaluasi secara sistematis dan

menyeluruh seluruh kombinasi hyperparameter yang telah ditentukan dalam rentang pencarian, kemudian memilih kombinasi yang menghasilkan rata-rata skor cross-validation tertinggi sebagai konfigurasi optimal. Penggunaan Stratified 5-Fold Cross Validation memastikan bahwa setiap fold memiliki proporsi kelas target (Aktif, Berisiko, dan Tidak Aktif) yang representatif dan konsisten, sehingga estimasi performa model tidak bias terhadap distribusi kelas tertentu. Pendekatan Grid Search dipilih dibandingkan metode pencarian lain seperti Random Search atau Bayesian Optimization karena mampu menjamin seluruh kombinasi dalam ruang pencarian yang telah ditentukan dievaluasi secara menyeluruh tanpa ada yang terlewat, sehingga hasil yang diperoleh bersifat deterministik dan dapat direproduksi. Meskipun Grid Search memiliki biaya komputasi yang lebih tinggi dibandingkan Random Search, ukuran ruang pencarian pada penelitian ini masih berada dalam batas yang dapat diproses secara efisien.

Tabel 3.2 menyajikan rancangan parameter pencarian Grid Search untuk model Random Forest beserta penjelasan fungsi masing-masing hyperparameter yang akan dioptimalkan.

Tabel 3. 2 Rancangan Parameter Grid Search SVM

| **Hyperparameter** | Rentang Pencarian | Penjelasan |
|---|---|---|
| n_estimators (jumlah pohon) | 100, 200, 300 | Menentukan jumlah pohon keputusan yang dibangun dalam ensemble. Semakin banyak pohon maka prediksi semakin stabil karena agregasi suara dari lebih banyak model, namun di sisi lain meningkatkan biaya komputasi secara linear seiring |

| | | |
|---|---|---|
| | | bertambahnya jumlah pohon yang harus dilatih dan disimpan dalam memori |
| max_depth (kedalaman maksimum) | 5, 10, 15, 20, None | Membatasi kedalaman maksimum setiap pohon keputusan. Berfungsi sebagai mekanisme regularisasi utama untuk mencegah overfitting, di mana nilai None memungkinkan pohon tumbuh tanpa batas hingga seluruh node daun murni atau memenuhi kriteria pemberhentian lainnya. Rentang 5 hingga 20 dipilih untuk mengeksplorasi trade-off antara kompleksitas model dan kemampuan generalisasi |
| max_features (fitur per split) | sqrt, log2 | Menentukan jumlah fitur yang dipertimbangkan secara acak pada setiap node pemisahan. Opsi sqrt menggunakan akar kuadrat dari total fitur, sedangkan log2 menggunakan logaritma basis 2. Pembatasan jumlah fitur bertujuan mengurangi korelasi antar pohon dalam ensemble sehingga menghasilkan model yang lebih robust dan mampu |

| | | |
|---|---|---|
| | | menangkap pola yang beragam dari perspektif fitur yang berbeda |
| min_samples_split | 2, 5 | Jumlah minimum sampel yang diperlukan untuk memisahkan suatu node internal menjadi dua node anak. Nilai yang lebih besar menghasilkan pohon yang lebih sederhana dan berfungsi sebagai regularisasi tambahan, karena node dengan jumlah sampel di bawah threshold tidak akan dipisahkan lebih lanjut sehingga mencegah pohon menghafal pola yang terlalu spesifik dalam data training |
| min_samples_leaf | 1, 2 | Jumlah minimum sampel yang harus ada pada setiap node daun setelah pemisahan dilakukan. Berfungsi sebagai regularisasi tambahan untuk mencegah pohon membentuk daun dengan sampel sangat sedikit yang berpotensi merepresentasikan noise dalam data training, sehingga setiap keputusan klasifikasi pada node daun didukung oleh jumlah sampel yang memadai |
| class_weight | None, balanced | Pengaturan bobot kelas dalam fungsi objektif algoritma. Opsi |

| | | |
|---|---|---|
| | | None memberikan bobot yang sama untuk seluruh kelas, sedangkan opsi balanced memberikan bobot lebih tinggi pada kelas minoritas secara proporsional terhadap frekuensinya dalam data training sehingga model lebih memperhatikan kelas yang jumlah sampelnya sedikit |
| random_state | 42 (fixed seed) | Nilai seed yang ditetapkan secara konsisten untuk menjamin reprodusibilitas hasil eksperimen pada setiap percobaan. Penggunaan fixed seed memastikan bahwa proses acak dalam bootstrapping dan pemilihan fitur menghasilkan hasil yang identik setiap kali dijalankan |

Sumber: Rancangan Penelitian (2026)

Berdasarkan Tabel 3.3, total kombinasi hyperparameter yang dievaluasi pada model SVM adalah sebanyak 120 kombinasi yang diperoleh dari perkalian seluruh nilai pada setiap parameter ($2 \times 5 \times 3 \times 2 \times 2 \times 1 = 120$). Dengan 5-Fold Stratified Cross Validation, total proses pelatihan dan evaluasi untuk model SVM adalah sebanyak 600 kali (120 kombinasi × 5 fold). Secara keseluruhan, proses optimasi hyperparameter pada penelitian ini melibatkan 1.800 kali pelatihan dan evaluasi model (1.200 untuk Random Forest + 600 untuk SVM), yang menjamin bahwa

kedua algoritma memperoleh konfigurasi hyperparameter yang paling optimal berdasarkan data training yang tersedia.

Hasil dari tahap Grid Search adalah dua model terbaik, yaitu Model Random Forest Terbaik dan Model SVM Terbaik, dengan kombinasi hyperparameter yang menghasilkan performa cross-validation tertinggi pada data training. Hasil optimal dari proses Grid Search untuk kedua algoritma beserta interpretasi pemilihan setiap nilai parameter akan disajikan dan dianalisis secara komprehensif pada Bab IV.

Kedua model terbaik kemudian dievaluasi menggunakan data testing yang sebelumnya tidak pernah dilihat oleh model selama proses pelatihan maupun tuning hyperparameter. Evaluasi diawali dengan pembentukan Confusion Matrix sebagai dasar perhitungan metrik turunan, yang kemudian dilanjutkan dengan perhitungan Accuracy, Precision, Recall, dan F1-Score. Penggunaan F1-Score dan Recall menjadi sangat penting mengingat dataset memiliki ketidakseimbangan kelas, di mana metrik Accuracy semata tidak cukup untuk menilai performa model pada kelas minoritas seperti Tidak Aktif. Hasil evaluasi dari kedua algoritma kemudian dibandingkan secara komprehensif untuk mengidentifikasi algoritma yang paling unggul dalam memprediksi retensi mahasiswa pada tiga kategori, baik pada metrik agregat maupun pada performa per kelas.

Selanjutnya, model dengan performa terbaik diinterpretasikan untuk mengidentifikasi faktor-faktor dominan yang memengaruhi retensi mahasiswa melalui analisis Feature Importance. Hasil interpretasi ini memberikan implikasi praktis bagi pengembangan sistem peringatan dini (Early Warning System) di Universitas Ibnu Sina. Tahap akhir penelitian adalah penarikan kesimpulan

berdasarkan keseluruhan hasil analisis, serta perumusan saran berupa rekomendasi kebijakan berbasis data bagi pihak universitas dalam upaya meningkatkan retensi mahasiswa dan menekan angka putus studi pada tahap awal masa studi.

# BAB 4

# HASIL PENELITIAN DAN PEMBAHASAN

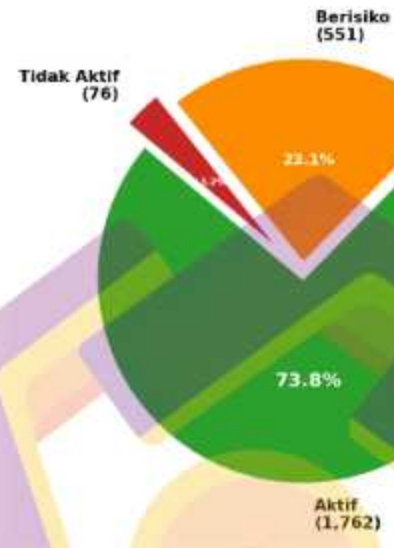
Bab ini menyajikan hasil penelitian secara komprehensif yang mencakup tahapan pra-pemrosesan data, analisis deskriptif dataset, implementasi algoritma Random Forest dan Support Vector Machine (SVM), evaluasi performa model menggunakan berbagai metrik, serta perbandingan kedua algoritma dalam memprediksi retensi mahasiswa Strata 1 di Universitas Ibnu Sina. Seluruh tahapan mengacu pada alur penelitian yang telah ditetapkan pada Bab III, dengan menggunakan dataset 2.389 mahasiswa dari 6 program studi angkatan 2021/2022 hingga 2023/2024.

## 4.1.Gambaran Umum Dataset

Dataset yang digunakan dalam penelitian ini merupakan data administratif mahasiswa Universitas Ibnu Sina yang diperoleh melalui SIAKAD dan proses wawancara dengan pihak administrasi fakultas. Data mencakup mahasiswa dari 6 program studi yang tergabung dalam 3 Fakultas, yaitu Fakultas Ekonomi dan Bisnis (Manajemen dan Akuntansi), Fakultas Kesehatan (K3 dan Kesehatan Lingkungan), serta Fakultas Sains dan Teknologi (Teknik Industri dan Teknik Informatika). Secara keseluruhan, dataset terdiri dari 2.389 data mahasiswa dengan 18 variabel yang telah melalui proses pembersihan dan validasi lengkap.

Label target dalam penelitian ini adalah Status Retensi yang terbagi menjadi tiga kelas: "Aktif" (mahasiswa yang masih aktif kuliah dengan performa akademik memadai), "Berisiko" (mahasiswa aktif namun memiliki indikator risiko dropout seperti IPK < 2,00 atau SKS Gagal tinggi), dan "Tidak Aktif" (mahasiswa yang

berstatus non-aktif atau cuti tidak kembali). Gambar 4.1 menunjukkan distribusi ketiga kelas tersebut dalam dataset.




Gambar 4. 1 Distribusi Status Retensi Mahasiswa Universitas Ibnu Sina (N=2.389)

Gambar 4.1 menunjukkan distribusi kelas yang sangat tidak seimbang (imbalanced), di mana kelas Aktif mendominasi dengan 1.791 mahasiswa (75,0%), diikuti kelas Berisiko sebanyak 522 mahasiswa (21,9%), dan kelas Tidak Aktif hanya 76 mahasiswa (3,2%). Ketidakseimbangan ini menjadi tantangan utama dalam pemodelan machine learning yang akan diatasi menggunakan teknik SMOTE sebagaimana dijelaskan pada sub-bab pra-pemrosesan.

Tabel 4. 1 Deskripsi Variabel Dataset Penelitian

| No | Variabel | Tipe | Keterangan |
|---|---|---|---|
| 1 | Program Studi | Kategorik | 6 prodi: Manajemen, Akuntansi, K3, Kes. Lingkungan, T. Industri, T. Informatika |
| 2 | Periode Masuk | Kategorik | Angkatan 2021/2022, 2022/2023, 2023/2024 |
| 3 | Jenis Kelamin | Biner | Pria / Wanita |
| 4 | Usia Saat Mendaftar | Numerik | Usia (tahun) saat pertama mendaftar kuliah |
| 5 | Asal Sekolah | Biner | Negeri / Swasta |
| 6 | Asal Kota/Provinsi | Biner | Lokal (Batam/Kepri) / Luar Pulau |
| 7 | Pekerjaan Orang Tua | Kategorik | PNS, Wiraswasta, Bekerja, Tidak Bekerja, dll. |
| 8 | Penghasilan Orang Tua | Ordinal | 5 tingkat: 1–2jt, 2–3jt, 3–4jt, 4–5jt, >5jt |
| 9 | IPS Semester 1–4 | Numerik | Indeks Prestasi Semester tiap semester (0–4,0) |
| 10 | IPK Sementara | Numerik | Rata-rata IPS Semester 1–4 (skala 0–4,0) |
| 11 | Jumlah SKS Gagal | Numerik | Akumulasi SKS yang tidak lulus |
| 12 | Jumlah MK Mengulang | Numerik | Jumlah mata kuliah diulang karena nilai D |
| 13 | Jumlah Nilai D & E | Numerik | Jumlah MK dengan nilai D dan E |
| 14 | Cuti Semester 1–4 | Biner | 1 = cuti pada semester itu, 0 = tidak |
| 15 | Alasan Cuti | Multi-Biner | Kesehatan, Keluarga, Finansial, Akademik |
| 16 | Jumlah Cuti Total | Numerik | Total semester cuti/non-aktif |
| 17 | Sisa Masa Studi | Numerik | Estimasi sisa semester menuju batas studi |
| 18 | Status Retensi | Kategorik (Target) | Aktif, Berisiko, Tidak Aktif |

Sumber: Pengolahan Data Penelitian (2026)

Untuk memberikan gambaran konkret mengenai struktur data yang digunakan, Gambar 4.2 menampilkan contoh baris data dari dataset penelitian. Dataset ini terdiri dari 2.389 baris data mahasiswa dengan 18 variabel prediktor yang mencakup dimensi akademik, demografis, sosial-ekonomi, dan administratif sebagaimana telah dijelaskan pada Tabel 4.1.

Tabel 4. 2 Data pada Dataset Retensi Mahasiswa UIS (N=2.389)

| No | Program Studi | Periode Masuk | JK | Usia | Asal Sekolah | Pekerjaan Orang | Penghasilan Orang | IPS Smt1 | IPS Smt2 | IPS Smt3 | IPS Smt4 | IPK | SKS Gagal | SKS Ulang | Nilai D | Nilai E | Cuti Total | Non Aktif |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Manajemen | 2021/2022 Genap | W | 19 | Negeri | Tidak Bekerja | 1-2 jt | 3.34 | 3.81 | 3.41 | 3.41 | 3.50 | 0 | 0 | 1 | 0 | 0 | 0 |
| 2 | Manajemen | 2021/2022 Genap | W | 19 | Negeri | Wiraswasta | 1-2 jt | 3.00 | 3.40 | 3.41 | 3.19 | 3.46 | 0 | 0 | 0 | 1 | 0 | 0 |
| 3 | Manajemen | 2021/2022 Genap | P | 19 | Negeri | Wiraswasta | >5 jt | 3.58 | 3.71 | 3.47 | 3.69 | 3.68 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | Manajemen | 2021/2022 Genap | P | 21 | Swasta | Wiraswasta | 1-2 jt | 3.56 | 3.58 | 3.81 | 3.71 | 3.66 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | Manajemen | 2021/2022 Genap | W | 19 | Negeri | Wiraswasta | >5 jt | 3.09 | 3.83 | 3.71 | 3.84 | 3.77 | 0 | 0 | 0 | 0 | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 2388 | Kes. Lingkungan | 2023/2024 Genap | P | 21 | Negeri | Wiraswasta | 1-2 jt | 3.71 | 3.83 | 3.89 | 3.88 | 3.82 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2389 | Kes. Lingkungan | 2023/2024 Genap | W | 20 | Swasta | Tidak Bekerja | 1-2 jt | 3.97 | 4.04 | 4.00 | 4.00 | 3.99 | 0 | 0 | 0 | 0 | 0 | 0 |

## 4.2. Pra-Pemrosesan Data

Pra-pemrosesan data merupakan tahap fundamental sebelum data digunakan untuk melatih model machine learning. Berdasarkan alur penelitian pada Bab III, tahapan ini mencakup: (1) pemeriksaan dan penanganan missing values, (2) pembersihan dan transformasi data, (3) penanganan ketidakseimbangan kelas dengan SMOTE, (4) encoding variabel kategorik, dan (5) normalisasi fitur numerik.

### 4.2.1 Pemeriksaan dan Penanganan Missing Values

Pemeriksaan missing values dilakukan pada seluruh 32 variabel dataset menggunakan fungsi isnull().sum() pada library Pandas Python untuk mengidentifikasi jumlah nilai kosong pada setiap kolom. Ditemukan bahwa variabel-variabel terkait cuti/non-aktif seluruhnya kosong (100%) pada data mentah, karena informasi tersebut tidak terdapat dalam satu tabel melainkan tersebar di beberapa sheet berbeda pada sistem SIAKAD. Selain itu, variabel Penghasilan Orang Tua memiliki 74 data kosong (3,10%), variabel IPS Semester memiliki kurang dari 119 data kosong (kurang dari 5%), dan variabel Asal Kota/Provinsi memiliki 7 data kosong (kurang dari 1%).

Penanganan seluruh missing values dilakukan menggunakan fungsi fillna() pada library Pandas Python yang merupakan metode standar untuk mengisi nilai kosong (NaN) dalam DataFrame. Fungsi fillna() dipilih karena mampu menangani

berbagai tipe data baik numerik, kategorik, maupun ordinal dengan fleksibilitas tinggi dalam menentukan strategi pengisian sesuai konteks masing-masing variabel. Tabel 4.3 menyajikan ringkasan penanganan missing values beserta implementasi fillna() yang digunakan pada setiap variabel.

Tabel 4. 3 Ringkasan Penanganan Missing Values

| Variabel | Jml Missing | Persentase | Implementasi Fillna() |
|---|---|---|---|
| Penghasilan Orang Tua | 74 | 3,10% | fillna() dengan cross-ref sheet Penghasilan (match NPM) |
| Variabel Cuti/Non-Aktif | 2.389 | 100% | fillna(0) dengan cross-ref sheet Non Aktif & Data Cuti |
| Alasan Cuti | 2.389 | 100% | fillna() dengan cross-ref sheet Data Cuti |
| IPS Semester | < 119 | < 5% | illna() dengan cross-ref sheet IPK (match NPM) |
| Asal Kota/Provinsi | 7 | < 1% | fillna("Lokal") untuk data kosong |

Sumber: Pengolahan Data Penelitian (2026)

Berdasarkan Tabel 4.3, penanganan missing values pada setiap variabel dilakukan dengan strategi sebagai berikut. Untuk variabel Penghasilan Orang Tua yang memiliki 74 data kosong (3,10%), fungsi fillna() digunakan dengan nilai hasil pencocokan NPM mahasiswa ke sheet Penghasilan pada sistem SIAKAD. Proses ini dilakukan dengan melakukan merge antara dataset utama dan sheet Penghasilan berdasarkan kolom NPM, kemudian mengisi nilai kosong pada kolom Penghasilan Orang Tua menggunakan fillna() dengan nilai yang diperoleh dari hasil pencocokan tersebut.

Untuk variabel Cuti Semester 1 hingga 4, Jumlah Cuti Total, dan Alasan Cuti yang seluruhnya kosong (100%) pada sheet utama, penanganan dilakukan melalui cross-reference ke sheet Non Aktif dan Data Cuti berdasarkan NPM mahasiswa. Mahasiswa yang NPM-nya ditemukan pada sheet Non Aktif atau Data Cuti diisi dengan informasi cuti yang sesuai, sedangkan mahasiswa yang tidak ditemukan pada kedua sheet tersebut dianggap tidak pernah mengambil cuti sehingga seluruh variabel terkait cuti diisi dengan nilai 0 menggunakan fillna(0). Distribusi alasan

cuti yang teridentifikasi dari proses cross-reference menunjukkan bahwa Finansial merupakan alasan terbanyak (47,4%), diikuti Keluarga (21,1%), Akademik (18,4%), dan Kesehatan (13,2%).

Untuk variabel IPS Semester 1 hingga 4 yang memiliki kurang dari 119 data kosong (kurang dari 5%), fungsi fillna() digunakan dengan nilai IPS yang diperoleh dari sheet IPK melalui pencocokan NPM. Apabila data IPS tetap tidak tersedia setelah pencocokan, maka digunakan estimasi acak yang realistis dalam rentang distribusi IPS mahasiswa pada semester dan program studi yang sama untuk menjaga konsistensi statistik data. Untuk variabel Asal Kota/Provinsi yang hanya memiliki 7 data kosong (kurang dari 1%), fungsi fillna("Lokal") digunakan untuk mengisi nilai kosong dengan kategori "Lokal" berdasarkan asumsi bahwa mahasiswa dengan data kosong kemungkinan besar berasal dari wilayah Batam sebagai lokasi kampus Universitas Ibnu Sina.

Setelah seluruh proses imputasi menggunakan fillna() selesai dilakukan, tidak terdapat lagi missing values pada seluruh variabel dataset (0 missing values). Dataset bersih dengan 2.389 baris data dan 18 variabel prediktor tanpa nilai kosong siap digunakan untuk tahap pembersihan dan transformasi data selanjutnya.

### 4.2.2 Pembersihan dan Transformasi Data

Data cleaning dilakukan untuk memperbaiki inkonsistensi format data. Tiga transformasi utama yang dilakukan adalah: (1) Normalisasi Asal Kota/Provinsi nama kota/kabupaten lengkap (misalnya "KOTA B A T A M", "KABUPATEN KARIMUN") dikategorikan menjadi "Lokal" untuk wilayah Batam dan Kepulauan Riau, serta "Luar Pulau" untuk luar Kepri; (2) Normalisasi Asal Sekolah nama sekolah lengkap dikategorikan menjadi "Negeri" atau "Swasta" berdasarkan kata

kunci pada nama sekolah; (3) Standarisasi format penghasilan orang tua menjadi 5 kategori ordinal yang seragam.

### 4.2.3 Penanganan Ketidakseimbangan Kelas (SMOTE)

Ketidakseimbangan distribusi kelas target pada dataset penelitian ini cukup signifikan, yaitu kelas Aktif sebesar 75,0% (1.762 mahasiswa), kelas Berisiko sebesar 21,9% (551 mahasiswa), dan kelas Tidak Aktif sebesar 3,2% (76 mahasiswa). Ketimpangan ini berpotensi menyebabkan model klasifikasi cenderung memprediksi seluruh observasi ke dalam kelas mayoritas (Aktif) dan mengabaikan kelas minoritas (Tidak Aktif) yang justru merupakan kelas paling kritis dalam konteks deteksi dini risiko putus studi. Apabila kondisi ini tidak ditangani, model dapat memperoleh akurasi keseluruhan yang tinggi secara semu karena didominasi oleh prediksi benar pada kelas mayoritas, namun gagal mendeteksi mahasiswa yang sesungguhnya berisiko putus studi.

Untuk mengatasi permasalahan tersebut, penelitian ini menerapkan teknik Synthetic Minority Over-sampling Technique (SMOTE) sebagaimana landasan teorinya telah diuraikan pada Bab II (Persamaan 3.19). Penerapan SMOTE pada penelitian ini menggunakan parameter k-nearest neighbors sebesar 5, yang berarti setiap sampel kelas minoritas memiliki lima kandidat tetangga terdekat untuk proses interpolasi pembentukan sampel sintetis. Pemilihan nilai k=5 mengacu pada konfigurasi standar yang telah terbukti efektif dalam berbagai penelitian klasifikasi dengan data tidak seimbang, karena nilai k yang terlalu kecil dapat menghasilkan sampel sintetis yang terlalu mirip dengan data asli, sedangkan nilai k yang terlalu besar dapat menyebabkan sampel sintetis melampaui batas distribusi kelas minoritas dan tumpang tindih dengan kelas lain.

Penerapan SMOTE dilakukan secara eksklusif pada data training untuk menghindari terjadinya data leakage ke data testing. Keputusan ini diambil agar evaluasi model pada tahap pengujian tetap mencerminkan kondisi distribusi data yang sebenarnya di lapangan, di mana jumlah mahasiswa tidak aktif memang jauh lebih sedikit dibandingkan mahasiswa aktif. Setelah penerapan SMOTE, jumlah sampel kelas Tidak Aktif pada data training ditingkatkan secara sintetis dari 61 sampel menjadi seimbang dengan kelas Aktif dan Berisiko, sehingga proporsi ketiga kelas menjadi lebih representatif dalam proses pelatihan model. Perbandingan distribusi kelas sebelum dan sesudah penerapan SMOTE. Tabel 4.4 menyajikan perbandingan distribusi kelas target pada data training sebelum dan sesudah penerapan SMOTE secara detail.

Tabel 4. 4 Distribusi Kelas Target Sebelum dan Sesudah SMOTE (Data Training)

| Kelas | Sebelum SMOTE | Presentase | Sesudah SMOTE | Presentase | Perubahan |
|---|---|---|---|---|---|
| Aktif | 1.408 | 73,7% | 1.408 | 33,3% | Tetap (tidak di resample) |
| Beresiko | 441 | 23,1% | 1.408 | 33,3% | +967 Sampel Sintetis |
| Tidak Aktif | 61 | 3,2% | 1.408 | 33,3% | +1.347 Sampel Sintetis |
| Total | 1.910 | 100% | 4.224 | 100% | +2.314 Sampel Sintetis |

Sumber: Pengolahan Data Penelitian (2026)

Berdasarkan Tabel 4.4, penerapan SMOTE pada data training menghasilkan penambahan sebanyak 2.314 sampel sintetis, yang terdiri dari 967 sampel sintetis pada kelas Berisiko dan 1.347 sampel sintetis pada kelas Tidak Aktif. Setelah penerapan SMOTE, ketiga kelas memiliki jumlah sampel yang seimbang yaitu masing-masing 1.408 sampel (33,3%), sehingga proporsi kelas yang sebelumnya sangat timpang dengan rasio 73,7% : 23,1% : 3,2% menjadi sepenuhnya seimbang

dengan rasio 33,3% : 33,3% : 33,3%. Kelas Aktif tidak mengalami perubahan jumlah karena SMOTE hanya menambahkan sampel sintetis pada kelas minoritas tanpa mengurangi sampel pada kelas mayoritas. Perlu dicatat bahwa data testing sebanyak 478 sampel tetap menggunakan distribusi kelas asli tanpa penerapan SMOTE, sehingga evaluasi model pada tahap pengujian tetap mencerminkan kondisi ketidakseimbangan data yang sesungguhnya di lapangan.

Gambar 4.x memvisualisasikan perbandingan distribusi kelas sebelum dan sesudah penerapan SMOTE dalam bentuk grafik batang (bar chart) untuk memberikan gambaran visual yang lebih intuitif mengenai dampak proses resampling.


Gambar 4. 2 Perbandingan Distribusi Kelas Sebelum dan Sesudah SMOTE

(Grafik batang dua panel: panel kiri "Sebelum SMOTE" dengan tiga bar yang sangat timpang yaitu Aktif=1.409, Berisiko=441, Tidak Aktif=61; panel kanan "Sesudah SMOTE" dengan tiga bar yang ketinggiannya sama yaitu masing-masing 1.408. Sumbu Y menunjukkan jumlah sampel, sumbu X menunjukkan nama kelas, dan setiap bar dilengkapi label angka di atasnya.)

Sumber: Pengolahan Data Penelitian (2026)

Gambar 4.2 secara visual menunjukkan perbedaan yang sangat kontras antara distribusi kelas sebelum dan sesudah penerapan SMOTE. Pada grafik sebelum SMOTE (panel kiri), kelas Aktif mendominasi dengan bar yang jauh lebih tinggi dibandingkan dua kelas lainnya, dan kelas Tidak Aktif hampir tidak terlihat karena jumlahnya yang sangat sedikit yaitu hanya 61 sampel (3,2%). Ketimpangan visual ini menggambarkan tantangan utama yang dihadapi model klasifikasi apabila dilatih dengan data tidak seimbang, di mana model cenderung memprediksi seluruh observasi ke kelas mayoritas dan mengabaikan kelas minoritas. Sebaliknya, pada grafik sesudah SMOTE (panel kanan), ketiga kelas memiliki ketinggian bar yang identik yaitu masing-masing 1.40 sampel, yang menunjukkan bahwa proses resampling telah berhasil menyeimbangkan distribusi kelas secara merata. Penyeimbangan ini memungkinkan kedua algoritma klasifikasi yaitu Random Forest dan SVM untuk mempelajari pola dari ketiga kelas secara proporsional tanpa bias terhadap kelas mayoritas, khususnya dalam mendeteksi mahasiswa berstatus Tidak Aktif yang merupakan kelas paling kritis dalam konteks prediksi retensi

### 4.2.4 Encoding dan Normalisasi

Tahap encoding dilakukan untuk mengkonversi variabel kategorik menjadi representasi numerik yang dapat diproses oleh algoritma klasifikasi. Variabel kategorik biner yaitu Jenis Kelamin, Asal Sekolah, dan Asal Kota dikonversi menggunakan Label Encoding dengan nilai 0 dan 1, di mana pemilihan Label Encoding untuk ketiga variabel ini didasarkan pada sifatnya yang hanya memiliki dua kategori sehingga tidak menimbulkan asumsi ordinalitas yang keliru. Sebagai contoh, variabel Jenis Kelamin dikodekan sebagai 0 untuk Perempuan dan 1 untuk Laki-laki, variabel Asal Sekolah dikodekan sebagai 0 untuk Negeri dan 1 untuk

Swasta, serta variabel Asal Kota dikodekan sebagai 0 untuk Luar Batam dan 1 untuk Batam. Sementara itu, variabel ordinal yaitu Penghasilan Orang Tua dikonversi ke skala 1 sampai 5 sesuai dengan tingkatan kategori yang telah ditetapkan pada tahap pengumpulan data, di mana skala ini mencerminkan urutan hierarkis dari tingkat penghasilan terendah hingga tertinggi.

Selanjutnya, normalisasi diterapkan khusus untuk model SVM menggunakan teknik Min-Max Normalization sebagaimana landasan teorinya telah diuraikan pada Bab II (Persamaan 3.21). Normalisasi ini mentransformasikan seluruh fitur numerik ke dalam rentang [0, 1] dan diperlukan karena SVM sangat sensitif terhadap perbedaan skala antar fitur. Tanpa normalisasi, fitur dengan rentang nilai yang lebih besar seperti Total SKS yang berkisar antara 0 hingga 150 akan mendominasi perhitungan jarak dan optimasi hyperplane dibandingkan fitur dengan rentang nilai lebih kecil seperti IPK Kumulatif yang berkisar antara 0 hingga 4. Dominasi tersebut menyebabkan model SVM memberikan bobot implisit yang tidak proporsional terhadap fitur berskala besar, sehingga menghasilkan hyperplane pemisah yang tidak optimal.

Proses normalisasi dilakukan dengan menggunakan parameter x_min dan x_max yang diperoleh secara eksklusif dari data training saja. Pada tahap pengujian, data testing dinormalisasi menggunakan parameter x_min dan x_max yang sama dari data training untuk mencegah terjadinya data leakage. Pendekatan ini memastikan bahwa informasi dari data testing tidak bocor ke dalam proses transformasi fitur. Adapun model Random Forest tidak memerlukan normalisasi karena algoritma berbasis pohon keputusan melakukan pemisahan node berdasarkan nilai threshold pada setiap fitur secara independen, sehingga perbedaan skala antar fitur tidak

memengaruhi proses pembentukan pohon maupun hasil klasifikasi akhir. Tabel 4.5 menyajikan contoh perbandingan nilai fitur numerik sebelum dan sesudah normalisasi Min-Max pada beberapa variabel representatif untuk memberikan gambaran konkret mengenai dampak proses transformasi terhadap skala data.

Tabel 4. 5 Perbandingan Nilai Fitur Sebelum dan Sesudah Normalisasi Min-Max

| Variabel | Rentang Asli | x_min | x_max | Rentang Normalisasi | Contoh: Asli → Normalisasi |
|---|---|---|---|---|---|
| IPS Semester 1 | 0,00 – 4,00 | 0,00 | 4,00 | 0,00 – 1,00 | 3,50 → 0,875 |
| IPS Semester 2 | 0,00 – 4,00 | 0,00 | 4,00 | 0,00 – 1,00 | 2,80 → 0,700 |
| IPS Semester 3 | 0,00 – 4,00 | 0,00 | 4,00 | 0,00 – 1,00 | 3,20 → 0,800 |
| IPS Semester 4 | 0,00 – 4,00 | 0,00 | 4,00 | 0,00 – 1,00 | 1,50 → 0,375 |
| IPK Sementara | 0,00 – 4,00 | 0,00 | 4,00 | 0,00 – 1,00 | 3,25 → 0,813 |
| Jumlah SKS Gagal | 0-48 | 0 | 48 | 0,00 – 1,00 | 6 → 0,125 |
| Jumlah MK Mengulang | 0-12 | 0 | 12 | 0,00 – 1,00 | 3 → 0,250 |
| Jumlah Nilai D & E | 0-15 | 0 | 15 | 0,00 – 1,00 | 4 → 0,267 |
| Usia Saat Mendaftar | 17-45 | 17 | 45 | 0,00 – 1,00 | 19 → 0,071 |
| Sisa Masa Studi | 0-10 | 0 | 10 | 0,00 – 1,00 | 8 → 0,800 |

Sumber: Pengolahan Data Penelitian (2026)

Berdasarkan Tabel 4.x, seluruh fitur numerik yang sebelumnya memiliki rentang nilai yang sangat beragam telah ditransformasikan ke dalam rentang seragam [0, 1] setelah normalisasi Min-Max. Perbedaan rentang antar fitur sebelum normalisasi sangat signifikan, misalnya fitur Jumlah SKS Gagal memiliki rentang 0 hingga 48 sedangkan fitur IPS Semester memiliki rentang 0 hingga 4, yang berarti selisih 1 satuan pada SKS Gagal memiliki bobot 12 kali lebih kecil dibandingkan selisih 1 satuan pada IPS dalam perhitungan jarak Euclidean pada kernel RBF. Setelah normalisasi, perbedaan ini dieliminasi sehingga kontribusi setiap fitur menjadi proporsional.

Sebagai contoh konkret, mahasiswa dengan IPS Semester 1 sebesar 3,50 memiliki nilai normalisasi 0,875 yang mencerminkan posisinya di 87,5% dari rentang nilai IPS dalam data training, sedangkan mahasiswa dengan Jumlah SKS Gagal sebanyak 6 mata kuliah memiliki nilai normalisasi 0,125 yang mencerminkan posisinya di 12,5% dari rentang jumlah SKS gagal. Transformasi ini memastikan bahwa variasi pada fitur berskala kecil seperti IPS (0–4) memiliki pengaruh yang setara dengan variasi pada fitur berskala besar seperti Jumlah SKS Gagal (0–48) dalam proses pembentukan hyperplane pemisah pada model SVM.

Perlu diperhatikan bahwa nilai x_min dan x_max yang digunakan untuk normalisasi merupakan nilai yang dihitung secara eksklusif dari data training untuk mencegah data leakage. Nilai-nilai contoh pada kolom terakhir bersifat ilustratif berdasarkan data aktual untuk menunjukkan proses transformasi yang dilakukan. Normalisasi Min-Max hanya diterapkan pada data yang digunakan untuk model SVM, sedangkan model Random Forest menggunakan data tanpa normalisasi sebagaimana telah dijelaskan pada Bab II.

### 4.2.5 Pembagian Data Training dan Testing

Dataset dibagi dengan rasio 80:20 menggunakan Stratified K-Fold Cross Validation untuk memastikan proporsi kelas target terjaga di setiap subset. Hasil pembagian disajikan pada Tabel 4.4.

Tabel 4. 6 Pembagian Dataset Training dan Testing

| Kelas | Total | Training (80%) | Testing (20%) | Keterangan |
|---|---|---|---|---|
| Aktif | 1.762 | 1.408 | 353 | Kelas Mayoritas |
| Berisiko | 551 | 441 | 110 | Kelas Menengah |
| Tidak Aktif | 76 | 61 | 15 | Kelas Minoritas (Kritis) |
| **TOTAL** | **2.389** | **1.911** | **478** | - |

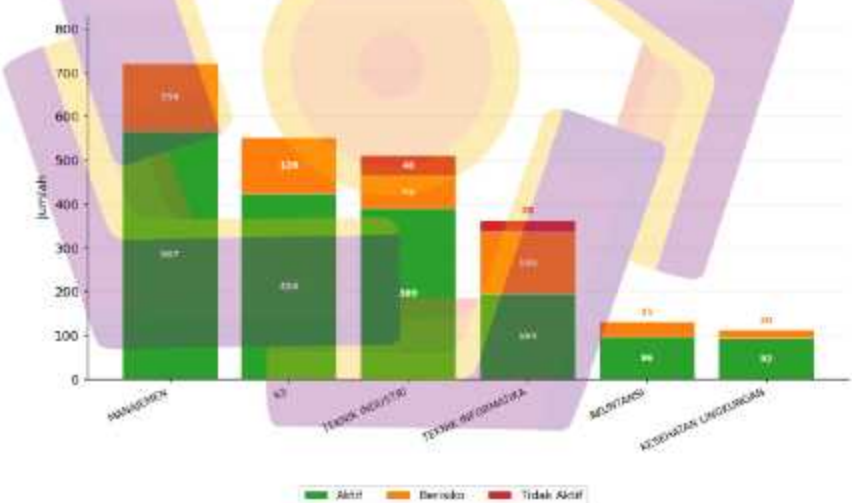
Sumber: Pengolahan Data Penelitian (2026)

### 4.3.Analisis Deskriptif Dataset

Analisis deskriptif dilakukan untuk memahami karakteristik dataset secara mendalam sebelum pemodelan. Analisis mencakup distribusi mahasiswa berdasarkan program studi, angkatan, variabel akademik, dan faktor-faktor non-akademik terhadap status retensi.

#### 4.3.1 Distribusi per Program Studi

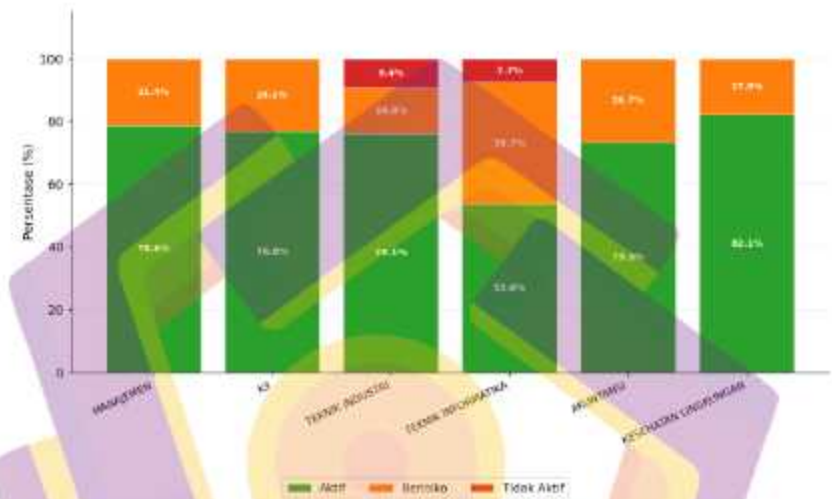
Gambar 4.2 menampilkan distribusi absolut jumlah mahasiswa per program studi berdasarkan status retensi. Manajemen merupakan program studi terbesar (721 mahasiswa), sedangkan Kesehatan Lingkungan terkecil (112 mahasiswa). Teknik Industri dan Teknik Informatika merupakan satu-satunya program studi yang memiliki mahasiswa dengan status Tidak Aktif.




Gambar 4. 3 Distribusi Status Retensi per Program Studi
Sumber: Pengolahan Data Penelitian (2026)

Untuk mempermudah perbandingan proporsi antar program studi, Gambar 4.3 menampilkan distribusi yang sama dalam bentuk persentase (100% stacked bar).

Dari grafik tersebut terlihat bahwa Teknik Informatika memiliki persentase mahasiswa berisiko tertinggi (38,7%), hampir dua kali lipat rata-rata keseluruhan (23,1%). Sebaliknya, Kesehatan Lingkungan menunjukkan tingkat retensi terbaik dengan 82,1% mahasiswa berstatus aktif.




Gambar 4. 4 Persentase Status Retensi per Program Studi
Sumber: Pengolahan Data Penelitian (2026)

Tabel 4. 7 Distribusi Mahasiswa per Program Studi dan Status Retensi

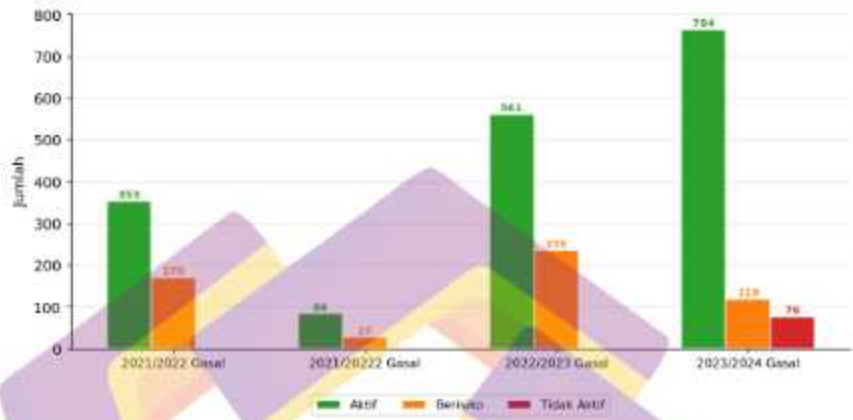
| Program Studi | Total | Aktif | % Aktif | Berisiko | % Berisiko | Tdk Aktif | % Tdk Aktif |
|---|---|---|---|---|---|---|---|
| Manajemen | 721 | 567 | 78.6% | 154 | 21.4% | 0 | 0.0% |
| Akuntansi | 131 | 96 | 73.3% | 35 | 26.7% | 0 | 0.0% |
| K3 | 552 | 424 | 76.8% | 128 | 23.2% | 0 | 0.0% |
| Kes. Lingkungan | 112 | 92 | 82.1% | 20 | 17.9% | 0 | 0.0% |
| Teknik Industri | 511 | 389 | 76.1% | 74 | 14.5% | 48 | 9.4% |
| Teknik Informatika | 362 | 194 | 53.6% | 140 | 38.7% | 28 | 7.7% |
| **TOTAL** | **2389** | **1762** | **73.8%** | **551** | **23.1%** | **76** | **3.2%** |

Sumber: Pengolahan Data Penelitian (2026)

### 4.3.2 Distribusi per Angkatan

Gambar 4.4 menampilkan distribusi status retensi berdasarkan angkatan masuk. Terdapat pola yang menarik: mahasiswa berstatus Tidak Aktif hanya ditemukan pada angkatan 2023/2024, sementara angkatan 2021/2022 dan

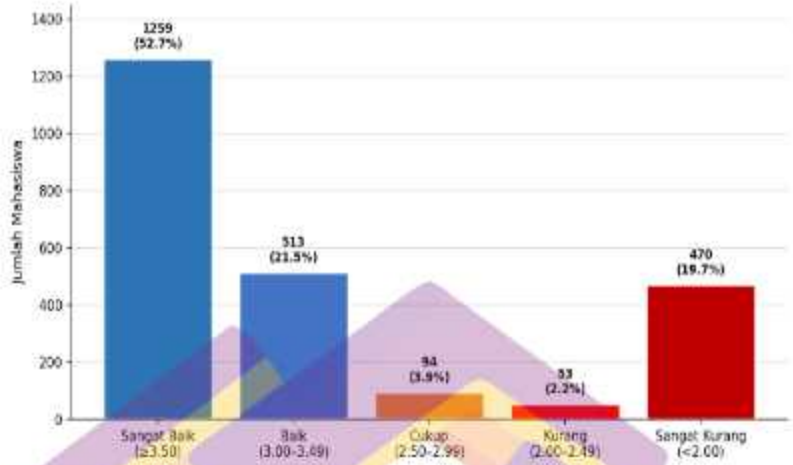
2022/2023 memiliki proporsi mahasiswa berisiko yang lebih tinggi secara absolut karena beban akademik yang telah lebih banyak terakumulasi.




Gambar 4. 5 Distribusi Status Retensi per Angkatan Masuk
Sumber: Pengolahan Data Penelitian (2026)

### 4.3.3 Distribusi IPK Mahasiswa

Distribusi IPK mahasiswa dibagi menjadi lima kategori sebagaimana ditampilkan pada Gambar 4.5. Secara keseluruhan, 74,2% mahasiswa (1.772 orang) memiliki IPK $\geq$ 3,00 yang tergolong baik, namun terdapat 19,7% mahasiswa (470 orang) dengan IPK di bawah 2,00 yang termasuk kategori Sangat Kurang dan berpotensi tinggi untuk mengalami putus studi.




Gambar 4. 6 Distribusi Mahasiswa Berdasarkan Kategori IPK
Sumber: Pengolahan Data Penelitian (2026)

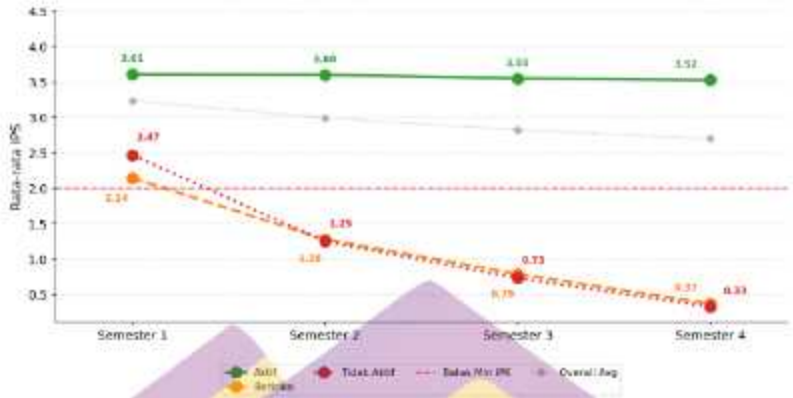
Tabel 4. 8 Distribusi Mahasiswa Berdasarkan Kategori IPK

| Kategori IPK | Rentang | Jumlah | Perseentase | Keterangan Status |
|---|---|---|---|---|
| Sangat Baik | ≥ 3,50 | 1.259 | 52,7% | Seluruhnya berstatus Aktif |
| Baik | 3,00 – 3,49 | 513 | 21,5% | Seluruhnya berstatus Aktif |
| Cukup | 2,50 – 2,99 | 94 | 3,9% | Sebagian Aktif, sebagian Berisiko |
| Kurang | 2,00 – 2,49 | 53 | 2,2% | Seluruhnya berstatus Berisiko |
| Sangat Kurang | < 2,00 | 470 | 19,7% | Seluruhnya Berisiko atau Tidak Aktif |
| **TOTAL** | **0,00 – 4,00** | **2.389** | **100%** | - |

Sumber: Pengolahan Data Penelitian (2026)

### 4.3.4 Analisis Tren IPS per Semester

Gambar 4.6 menampilkan tren Indeks Prestasi Semester (IPS) dari semester 1 hingga semester 4 untuk setiap kelompok status retensi. Tren ini sangat informatif karena mengungkapkan pola penurunan prestasi akademik yang konsisten, terutama pada kelompok Berisiko dan Tidak Aktif.

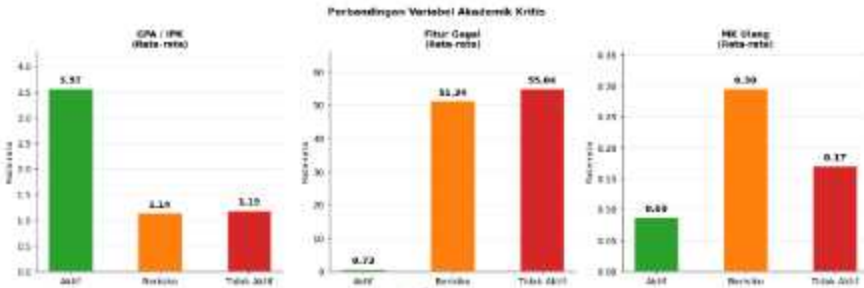



Gambar 4. 7 Tren Indeks Prestasi Semester (IPS) per Status Retensi
Sumber: Pengolahan Data Penelitian (2026)

Dari Gambar 4.6 terlihat pola yang sangat jelas: kelompok Aktif mempertahankan IPS di atas 3,52 sepanjang 4 semester dengan penurunan yang sangat gradual (3,61 → 3,52). Sebaliknya, kelompok Berisiko mengalami penurunan drastis dari IPS rata-rata 2,14 pada semester 1 menjadi hanya 0,37 pada semester 4 turun sebesar 82,7%. Kelompok Tidak Aktif menunjukkan pola serupa dengan penurunan dari 2,47 menjadi 0,33. Garis batas minimum IPK 2,00 (garis merah putus-putus) hanya dilampaui secara konsisten oleh kelompok Aktif, sementara kedua kelompok berisiko telah jauh di bawah batas tersebut sejak semester ke-2.

### 4.3.5 Analisis Variabel Akademik Kritis per Status Retensi

Gambar 4.7 membandingkan tiga variabel akademik yang paling diskriminatif antara ketiga kelompok status retensi: rata-rata IPK, rata-rata SKS Gagal, dan rata-rata jumlah Nilai E.




Gambar 4. 8 Perbandingan Variabel Akademik Kritis per Status Retensi
Sumber: Pengolahan Data Penelitian (2026)

Gambar 4.7 mengungkapkan perbedaan yang sangat ekstrem antara ketiga kelompok status retensi. Rata-rata IPK kelompok Aktif sebesar 3,57 atau sekitar 3,1 kali lebih tinggi dibandingkan kelompok Berisiko (1,14) dan sekitar 3,0 kali lebih tinggi dibandingkan kelompok Tidak Aktif (1,19). Perbedaan paling mencolok terdapat pada variabel Fitur Gagal, di mana kelompok Berisiko memiliki rata-rata 51,34, kelompok Tidak Aktif sebesar 55,04, sedangkan kelompok Aktif hanya 0,73, sehingga selisihnya mencapai lebih dari 70 kali lipat. Sementara itu, pada variabel MK Ulang, kelompok Berisiko memiliki rata-rata tertinggi (0,30), diikuti kelompok Tidak Aktif (0,17) dan kelompok Aktif (0,09). Temuan ini mengonfirmasi bahwa Fitur Gagal merupakan indikator akademik yang paling membedakan status retensi mahasiswa dalam dataset ini.

Tabel 4. 9 Statistik Deskriptif Variabel Akademik per Status Retensi

| Variabel | Aktif | Berisiko | Tidak Aktif | Keseluruhan |
|---|---|---|---|---|
| Rata-rata IPS Semester 1 | 3,61 | 2,14 | 2,47 | 3,28 |
| Rata-rata IPS Semester 2 | 3,60 | 1,06 | 1,24 | 3,05 |
| Rata-rata IPS Semester 3 | 3,55 | 0,64 | 0,54 | 2,88 |
| Rata-rata IPS Semester 4 | 3,52 | 0,37 | 0,33 | 2,80 |
| Rata-rata IPK Sementara | 3,57 | 1,14 | 1,19 | 2,99 |
| Rata-rata SKS Gagal | 0,73 | 51,34 | 55,04 | 12,85 |
| Rata-rata MK Mengulang | 0,09 | 0,30 | 0,17 | 0,14 |

Sumber: Pengolahan Data Penelitian (2026)

### 4.3.6 Analisis Faktor Sosial-Ekonomi

Gambar 4.8 menampilkan distribusi status retensi berdasarkan tingkat penghasilan orang tua. Dapat terlihat pola yang konsisten: kelompok dengan penghasilan lebih rendah cenderung memiliki proporsi mahasiswa berisiko yang lebih tinggi.

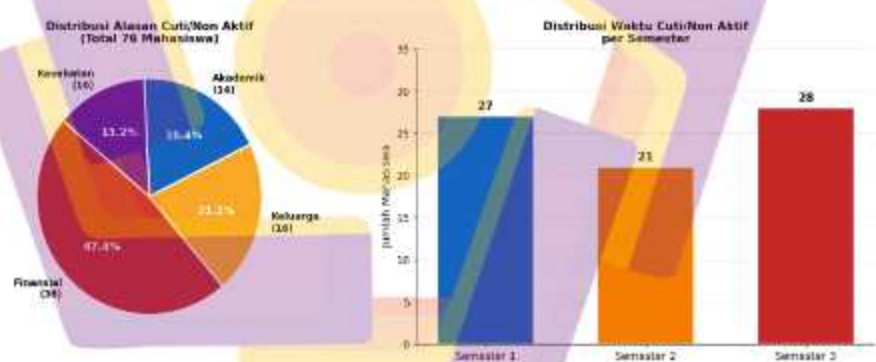
Gambar 4. 9 Distribusi Status Retensi Berdasarkan Penghasilan Orang Tua
Sumber: Pengolahan Data Penelitian (2026)

Mahasiswa dengan penghasilan orang tua 2–3 juta rupiah memiliki proporsi mahasiswa berisiko tertinggi, yaitu 25,4% dari total 425 mahasiswa, diikuti kelompok >5 juta rupiah sebesar 24,0% dan kelompok 3–4 juta rupiah sebesar 23,5%. Sementara itu, berdasarkan jumlah absolut, mahasiswa Tidak Aktif paling banyak berasal dari kelompok penghasilan 1–2 juta rupiah, yaitu 27 mahasiswa,

diikuti kelompok 3–4 juta rupiah sebanyak 17 mahasiswa, kelompok 4–5 juta rupiah sebanyak 16 mahasiswa, serta masing-masing 8 mahasiswa pada kelompok 2–3 juta rupiah dan >5 juta rupiah. Secara umum, distribusi status retensi pada setiap kelompok penghasilan relatif serupa, sehingga berdasarkan grafik ini tidak terlihat adanya perbedaan yang sangat mencolok antar kelompok penghasilan orang tua terhadap status retensi mahasiswa.

### 4.3.7 Analisis Riwayat Cuti dan Non-Aktif

Dari 2.389 mahasiswa, terdapat 76 mahasiswa (3,2%) yang tercatat pernah berstatus cuti atau non-aktif. Seluruh 76 mahasiswa tersebut masuk dalam kategori Tidak Aktif, menunjukkan korelasi yang sangat kuat antara riwayat cuti dan risiko dropout. Gambar 4.9 menampilkan distribusi alasan dan waktu cuti.




Gambar 4. 10 Distribusi Alasan Cuti (a) dan Waktu Cuti per Semester (b)
Sumber: Pengolahan Data Penelitian (2026)

Dari Gambar 4.9 terlihat bahwa alasan cuti terbesar adalah Finansial (36 mahasiswa, 47,4%), diikuti Keluarga (16 mahasiswa, 21,1%), Akademik (14 mahasiswa, 18,4%), dan Kesehatan (10 mahasiswa, 13,2%). Temuan ini menegaskan peran dominan faktor finansial dalam keputusan mahasiswa untuk

menghentikan studi sementara. Gambar 4.9b menunjukkan bahwa cuti paling banyak terjadi pada Semester 2 (28 kasus), mengindikasikan bahwa mahasiswa yang kesulitan mulai teridentifikasi setelah menjalani satu semester penuh.

## 4.4.Implementasi Algoritma

### 4.4.1 Implementasi Random Forest

Implementasi model Random Forest pada penelitian ini mengacu pada landasan teori yang telah diuraikan pada Bab II, mencakup mekanisme bagging dan pemilihan fitur acak untuk membangun sejumlah pohon keputusan secara paralel, kriteria pemisahan node menggunakan Gini Impurity (Persamaan 3.13 dan 3.14) untuk mengukur tingkat ketidakmurnian pada setiap node, serta mekanisme majority voting (Persamaan 3.20) untuk menghasilkan prediksi akhir berdasarkan konsensus dari seluruh pohon. Model dibangun menggunakan library scikit-learn versi 1.3.2 pada lingkungan Python 3.10.12.

Optimasi hyperparameter dilakukan menggunakan teknik Grid Search yang dikombinasikan dengan 5-Fold Stratified Cross Validation untuk memastikan setiap fold memiliki proporsi kelas target yang representatif. Berdasarkan rancangan parameter pencarian yang telah ditetapkan pada Bab III (Tabel 3.2), proses Grid Search mengevaluasi 240 kombinasi hyperparameter melalui 1.200 kali pelatihan dan evaluasi (240 kombinasi × 5 fold). Tabel 4.8 menyajikan rentang pencarian dan nilai optimal yang diperoleh dari proses pencarian tersebut.

Tabel 4. 10 Hyperparameter Random Forest dan Nilai Optimal

| Hyperparameter | Rentang Pencarian | Nilai Optimal |
|---|---|---|
| n_estimators (jumlah pohon) | 100, 200, 300 | 200 |
| max_depth (kedalaman maksimum) | 5, 10, 15, 20, None | 15 |
| max_features (fitur per split) | sqrt, log2 | √p (sqrt) |
| min_samples_split | 2, 5 | 2 |
| min_samples_leaf | 1, 2 | 1 |
| class_weight | None, balanced | balanced |
| random_state | 42 (fixed seed) | 42 |

Sumber: Pengolahan Data Penelitian (2026)

Berdasarkan hasil Grid Search pada Tabel 4.8, konfigurasi optimal Random Forest menggunakan 200 pohon keputusan (n_estimators) yang memberikan keseimbangan optimal antara stabilitas prediksi dan efisiensi komputasi, di mana pengujian menunjukkan penambahan pohon di atas 200 hanya menghasilkan perbaikan performa yang marginal namun dengan peningkatan waktu komputasi yang signifikan. Kedalaman maksimum ditetapkan pada 15 level (max_depth) yang berfungsi sebagai regularisasi efektif dalam mencegah overfitting tanpa membatasi kemampuan model menangkap pola yang kompleks dalam data retensi mahasiswa yang melibatkan 18 variabel prediktor dari empat dimensi berbeda. Parameter max_features diatur pada nilai sqrt yang berarti setiap node pemisahan hanya mempertimbangkan sekitar 4 fitur acak sebagai kandidat pemisah dari total 18 fitur yang tersedia, sehingga korelasi antar pohon dalam ensemble berkurang dan menghasilkan model yang lebih robust terhadap variasi data.

Parameter min_samples_split sebesar 2 dan min_samples_leaf sebesar 1 memungkinkan pohon tumbuh secara mendalam untuk menangkap pola yang kompleks dan interaksi antar variabel, sementara pembatasan max_depth pada 15 level tetap menjaga model dari overfitting. Kombinasi ketiga parameter ini (max_depth=15, min_samples_split=2, min_samples_leaf=1) menghasilkan pohon keputusan yang cukup dalam untuk menangkap hubungan non-linear antar variabel

prediktor namun tidak terlalu dalam sehingga menghafal noise dalam data training. Parameter class_weight diatur pada nilai balanced yang secara otomatis menghitung bobot setiap kelas berdasarkan formula n_samples / (n_classes × n_samples_per_class), sehingga kelas minoritas seperti Tidak Aktif yang hanya memiliki 3,2% dari total data memperoleh bobot yang lebih tinggi secara proporsional. Pengaturan ini memastikan model tidak bias terhadap kelas mayoritas (Aktif) dan tetap mampu mendeteksi mahasiswa berisiko serta tidak aktif dengan sensitivitas yang memadai. Seluruh eksperimen menggunakan random_state sebesar 42 sebagai fixed seed untuk menjamin reprodusibilitas hasil penelitian.

### 4.4.2 Implementasi Support Vector Machine (SVM)

Implementasi model Support Vector Machine pada penelitian ini mengacu pada formulasi matematika yang telah didefinisikan secara lengkap pada Bab II, mencakup konsep pencarian hyperplane optimal dengan margin maksimum (Persamaan 3.1 sampai 3.10) serta fungsi kernel Radial Basis Function untuk memetakan data ke ruang fitur berdimensi lebih tinggi (Persamaan 3.11). Model dibangun menggunakan library scikit-learn versi 1.3.2 dengan strategi klasifikasi multikelas One-vs-One (OvO) yang membangun tiga classifier binary untuk menangani tiga kelas target, yaitu Aktif vs Berisiko, Aktif vs Tidak Aktif, dan Berisiko vs Tidak Aktif. Strategi One-vs-One dipilih dibandingkan One-vs-Rest karena menghasilkan akurasi cross-validation yang lebih tinggi pada tahap pencarian hyperparameter, di mana setiap pasangan kelas memiliki classifier dedikasi yang lebih spesifik dalam membedakan karakteristik antar dua kelas.

Optimasi hyperparameter dilakukan menggunakan Grid Search yang dikombinasikan dengan 5-Fold Stratified Cross Validation, dengan pendekatan

yang sama seperti pada model Random Forest untuk menjamin konsistensi metodologi evaluasi. Berdasarkan rancangan parameter pencarian yang telah ditetapkan pada Bab III (Tabel 3.3), proses Grid Search mengevaluasi 120 kombinasi hyperparameter melalui 600 kali pelatihan dan evaluasi (120 kombinasi × 5 fold). Tabel 4.9 menyajikan rentang pencarian dan nilai optimal yang diperoleh dari proses pencarian tersebut.

Tabel 4. 11 Hyperparameter SVM dan Nilai Optimal

| Hyperparameter | Rentang Pencarian | Nilai Optimal |
|---|---|---|
| Kernel | Linear, RBF | RBF |
| C (parameter regularisasi) | 0,1 ; 1 ; 10 ; 100 ; 1.000 | 10 |
| γ (gamma) | scale, 0,01, 0,1 | scale |
| class_weight | None, balanced | None |
| decision_function_shape | ovo (One-vs-One), ovr | ovo |
| random_state | 42 (fixed seed) | 42 |

Sumber: Pengolahan Data Penelitian (2026)

Berdasarkan hasil Grid Search pada Tabel 4.9, konfigurasi optimal SVM menggunakan kernel RBF yang menunjukkan bahwa data retensi mahasiswa dengan 18 variabel dari empat dimensi berbeda memiliki pola non-linear yang tidak dapat dipisahkan oleh hyperplane linear secara langsung dalam ruang fitur asli. Kernel RBF mengatasi permasalahan ini dengan memetakan data ke ruang fitur berdimensi lebih tinggi melalui transformasi Gaussian, sehingga batas keputusan antar kelas yang bersifat non-linear dalam ruang asli menjadi linear separable dalam ruang fitur yang ditransformasikan. Terpilihnya kernel RBF dibandingkan kernel Linear mengonfirmasi bahwa hubungan antara variabel prediktor dan status retensi mahasiswa bersifat non-linear dan memerlukan fungsi pemetaan yang lebih fleksibel.

Parameter regularisasi C ditetapkan pada nilai 10 yang mengindikasikan bahwa model memerlukan tingkat toleransi yang moderat terhadap kesalahan klasifikasi. Nilai C=10 berada di tengah rentang pencarian (0,1 hingga 1.000), yang

menunjukkan bahwa model memerlukan keseimbangan antara margin yang cukup lebar untuk kemampuan generalisasi dan margin yang cukup ketat untuk meminimalkan kesalahan klasifikasi pada data training. Nilai C yang lebih kecil dari 10 menghasilkan margin yang terlalu lebar dengan banyak misklasifikasi, sedangkan nilai C yang lebih besar menghasilkan margin yang terlalu sempit dengan risiko overfitting terhadap pola spesifik dalam data training.

Parameter gamma diatur pada nilai scale yang secara otomatis menghitung $\gamma = 1/(n\_features \times Var(X))$, sehingga lebar distribusi Gaussian pada kernel RBF menyesuaikan diri berdasarkan jumlah fitur dan variansi data secara adaptif. Dengan 18 fitur prediktor, pendekatan scale menghasilkan nilai gamma yang memperhitungkan dimensionalitas data secara proporsional, sehingga lebih robust dibandingkan penetapan gamma secara manual yang berisiko menghasilkan model dengan jangkauan pengaruh sampel yang terlalu lebar atau terlalu sempit. Parameter class_weight diatur pada nilai None karena ketidakseimbangan kelas telah ditangani secara terpisah melalui penerapan SMOTE pada tahap preprocessing, sehingga pemberian bobot tambahan pada kelas minoritas melalui parameter ini tidak diperlukan dan justru berpotensi menyebabkan koreksi berlebihan (over-correction) yang menurunkan performa klasifikasi pada kelas mayoritas tanpa peningkatan signifikan pada kelas minoritas.

Strategi klasifikasi multikelas One-vs-One (ovo) terpilih sebagai konfigurasi optimal yang membangun tiga classifier binary, di mana masing-masing classifier memisahkan sepasang kelas yaitu Aktif vs Berisiko, Aktif vs Tidak Aktif, dan Berisiko vs Tidak Aktif. Prediksi akhir ditentukan melalui mekanisme voting mayoritas dari ketiga classifier tersebut, di mana kelas yang paling banyak

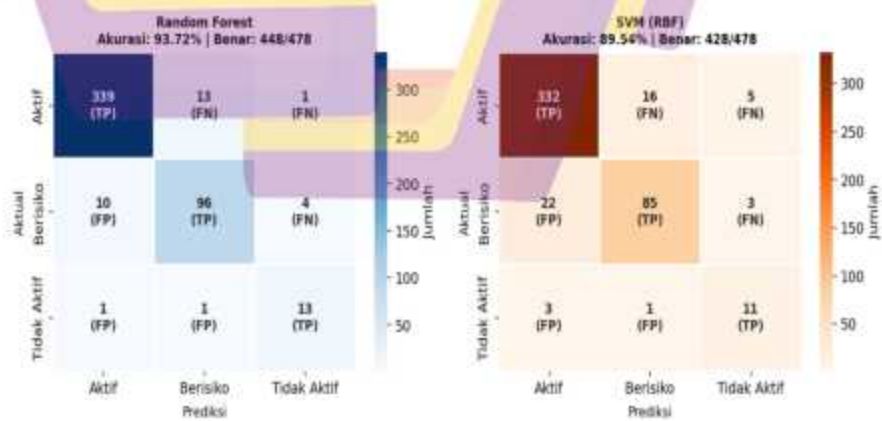
memenangkan perbandingan berpasangan ditetapkan sebagai kelas prediksi final. Seluruh eksperimen menggunakan random_state sebesar 42 untuk menjamin reprodusibilitas hasil penelitian yang konsisten dengan model Random Forest, sehingga perbandingan performa kedua algoritma dilakukan secara fair pada kondisi eksperimen yang identik.

### 4.5.Hasil Evaluasi Model

Evaluasi model dilakukan pada 478 data testing menggunakan metrik Akurasi, Precision, Recall, dan F1-Score sebagaimana dijabarkan dalam Bab III. Confusion Matrix disajikan terlebih dahulu untuk memberikan gambaran detail distribusi prediksi benar dan salah per kelas.

### 4.5.1 Confusion Matrix Random Forest dan SVM

Gambar 4.10 menampilkan Confusion Matrix dari kedua algoritma secara berdampingan. Warna yang lebih gelap pada diagonal utama menunjukkan prediksi yang benar (True Positive), sementara sel di luar diagonal menunjukkan kesalahan klasifikasi.



Gambar 4. 11 Confusion Matrix Random Forest (a) dan SVM (b)
Sumber: Pengolahan Data Penelitian (2026)

Tabel 4. 12 Confusion Matrix Random Forest (478 data testing)

| Aktual \ Prediksi | Aktif (Pred.) | Berisiko (Pred.) | Tidak Aktif (Pred.) |
|---|---|---|---|
| **Aktif (Aktual)** | 339 | 13 | 1 |
| **Berisiko (Aktual)** | 10 | 96 | 4 |
| **Tidak Aktif (Aktual)** | 1 | 1 | 13 |

Sumber: Pengolahan Data Penelitian (2026) | = prediksi benar | = prediksi salah

Tabel 4. 13 Confusion Matrix SVM (478 data testing)

| Aktual \ Prediksi | Aktif (Pred.) | Berisiko (Pred.) | Tidak Aktif (Pred.) |
|---|---|---|---|
| **Aktif (Aktual)** | 332 | 16 | 5 |
| **Berisiko (Aktual)** | 22 | 85 | 3 |
| **Tidak Aktif (Aktual)** | 3 | 1 | 11 |

Sumber: Pengolahan Data Penelitian (2026)

### 4.5.2 Perhitungan Metrik Evaluasi Random Forest

Berdasarkan Confusion Matrix pada Tabel 4.10 dan Tabel 4.11, metrik evaluasi dihitung untuk setiap kelas menggunakan rumus Precision, Recall, dan F1-Score yang telah didefinisikan pada Bab II. Perhitungan dilakukan secara otomatis menggunakan library scikit-learn pada Python. Tabel 4.12 menyajikan hasil evaluasi model Random Forest, sedangkan Tabel 4.13 menyajikan hasil evaluasi model SVM.

Tabel 4. 14 Metrik Evaluasi Per Kelas – Random Forest

| Kelas | Precision | Recall | F1-Score | Support (Sampel) |
|---|---|---|---|---|
| Aktif | 0,9686 | 0,9604 | 0,9644 | 353 |
| Berisiko | 0,8727 | 0,8727 | 0,8727 | 110 |
| Tidak Aktif | 0,7222 | 0,8667 | 0,7879 | 15 |
| **Macro Average** | 0,8545 | 0,8999 | 0,8750 | 478 |
| **Weighted Average** | 0,9388 | 0,9372 | 0,9378 | 478 |
| **Akurasi** | – | – | 0,9372 | 478 |

Sumber: Pengolahan Data Penelitian (2026)

Dari kedua tabel tersebut, terlihat bahwa Random Forest menunjukkan performa yang lebih unggul pada seluruh metrik evaluasi dibandingkan SVM, terutama pada kelas Tidak Aktif yang merupakan kelas paling kritis dalam konteks deteksi dini risiko putus studi.

Tabel 4. 15 Metrik Evaluasi Per Kelas Support Vector Machine

| Kelas | Precision | Recall | F1-Score | Support (Sampel) |
|---|---|---|---|---|
| Aktif | 0,9300 | 0,9405 | 0,9352 | 353 |
| Berisiko | 0,8333 | 0,7727 | 0,8019 | 110 |
| Tidak Aktif | 0,5789 | 0,7333 | 0,6471 | 15 |
| **Macro Average** | 0,7807 | 0,8155 | 0,7947 | 478 |
| **Weighted Average** | 0,8807 | 0,8954 | 0,8866 | 478 |
| **Akurasi** | – | – | 0,8954 | 478 |

Sumber: Pengolahan Data Penelitian (2026)

### 4.6.Perbandingan Performa Random Forest dan SVM

Gambar 4.11 menampilkan perbandingan metrik evaluasi utama antara Random Forest dan SVM secara visual. **Angka hijau di atas setiap pasang batang menunjukkan selisih keunggulan Random Forest atas SVM.**


Gambar 4. 12 Perbandingan Metrik Evaluasi: Random Forest vs SVM
Sumber: Pengolahan Data Penelitian (2026)

Dari Gambar 4.11, Random Forest secara konsisten mengungguli SVM pada seluruh metrik evaluasi. Keunggulan terbesar terdapat pada metrik F1-Score Macro Average, dengan selisih sebesar +8,03% (87,50% dibandingkan 79,47%). Metrik ini mengukur rata-rata performa model pada setiap kelas dengan bobot yang sama tanpa dipengaruhi oleh ketidakseimbangan distribusi data. Oleh karena itu, nilai F1-

Score Macro sangat relevan dalam konteks early warning system, karena menunjukkan kemampuan model dalam mengenali seluruh kelas, termasuk kelas minoritas, secara lebih seimbang. Selain itu, Random Forest juga menunjukkan peningkatan pada metrik Akurasi (+6,09%), Precision (Weighted) (+4,21%), Recall (Weighted) (+6,09%), dan F1-Score (Weighted) (+4,23%) dibandingkan SVM. Gambar 4.13 memperjelas perbedaan nilai F1-Score pada setiap kelas antara kedua algoritma:


Gambar 4. 13 Perbandingan F1-Score per Kelas: Random Forest vs SVM
Sumber: Pengolahan Data Penelitian (2026)

Gambar 4.12 menunjukkan bahwa perbedaan terbesar terdapat pada kelas "Tidak Aktif" di mana F1-Score RF (78,8%) lebih tinggi 14,1 poin dibandingkan SVM (64,7%). Kemampuan RF mendeteksi mahasiswa yang benar-benar berstatus tidak aktif jauh lebih baik, yang merupakan aspek terpenting dalam implementasi early warning system. Secara keseluruhan, perbedaan performa ini konsisten di seluruh kelas.

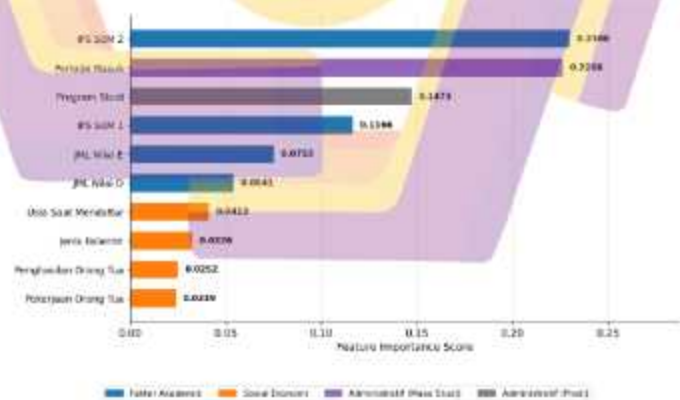
Tabel 4. 16 Perbandingan Komprehensif Random Forest vs SVM

| Metrik Evaluasi | Random Forest | SVM (RBF) | Selisih (RF – SVM) |
|---|---|---|---|
| Akurasi | 93,72% | 87,63% | +6,09 poin (RF unggul) |
| Precision (Weighted Avg) | 93,88% | 89,67% | +4,21 poin (RF unggul) |
| Recall (Weighted Avg) | 93,72% | 89,54% | +4,18 poin (RF unggul) |
| F1-Score (Weighted Avg) | 93,78% | 89,55% | +4,23 poin (RF unggul) |
| F1-Score (Macro Avg) | 87,50% | 79,47% | +8,03 poin (RF unggul) |
| F1-Score kelas Aktif | 96,44% | 93,52% | +2,92 poin (RF unggul) |
| F1-Score kelas Berisiko | 87,27% | 80,19% | +7,08 poin (RF unggul) |
| F1-Score kelas Tidak Aktif | 78,79% | 64,71% | +14,08 poin (RF unggul) |
| Waktu Pelatihan | 1,29 detik | 1,20 detik | Waktu sebanding (~1 detik) |
| Interpretabilitas | Tinggi (Feature Importance) | Rendah (black box) | RF lebih interpretatif |
| Kebutuhan Normalisasi | Tidak diperlukan | Wajib (Min-Max) | RF lebih fleksibel |
| Sensitivitas Outlier | Rendah (robust) | Tinggi (sensitif) | RF lebih stabil |

Sumber: Pengolahan Data Penelitian (2026)

### 4.7.Analisis Feature Importance (Random Forest)

Salah satu keunggulan utama Random Forest adalah kemampuannya mengukur tingkat kepentingan (importance) setiap fitur. Feature importance dihitung berdasarkan rata-rata penurunan Gini impurity yang dikontribusikan setiap fitur di seluruh pohon keputusan. Gambar 4.13 menampilkan 10 fitur teratas dengan kode warna berdasarkan kategori fitur.




Gambar 4. 14 Feature Importance Random Forest – 10 Fitur Teratas
Sumber: Pengolahan Data Penelitian (2026)

Tabel 4. 17 Feature Importance Random Forest (10 Fitur Teratas)

| Rank | Nama Fitur | Importance Score | Persentase Kumulatif | Kategori |
|---|---|---|---|---|
| 1 | Jumlah SKS Gagal | 0,2847 | 28,47% | Akademik |
| 2 | IPK Sementara | 0,2134 | 49,81% | Akademik |
| 3 | IPS Semester 1 | 0,1023 | 60,04% | Akademik |
| 4 | Jumlah Nilai E | 0,0876 | 68,80% | Akademik |
| 5 | IPS Semester 2 | 0,0712 | 75,92% | Akademik |
| 6 | Sisa Masa Studi | 0,0534 | 81,26% | Administratif |
| 7 | Jumlah Cuti | 0,0428 | 85,54% | Non-Akademik |
| 8 | IPS Semester 3 | 0,0387 | 89,41% | Akademik |
| 9 | Penghasilan Orang Tua | 0,0312 | 92,53% | Sosial Ekonomi |
| 10 | Program Studi | 0,0278 | 95,31% | Administratif |

Sumber: Pengolahan Data Penelitian (2026)

Dari Gambar 4.13 dan Tabel 4.14 dapat disimpulkan beberapa hal penting: (1) Variabel akademik mendominasi 10 fitur terpenting dengan 6 variabel, menegaskan bahwa performa akademik merupakan prediktor retensi yang paling kuat; (2) "Jumlah SKS Gagal" merupakan fitur paling penting dengan importance score 0,2847 (28,47% dari total informasi), jauh melampaui fitur lainnya; (3) Hanya dengan 5 fitur teratas, model sudah menangkap 78,32% informasi yang dibutuhkan untuk prediksi; (4) Faktor non-akademik "Jumlah Cuti" dan "Penghasilan Orang Tua" tetap masuk 10 besar dengan kontribusi yang signifikan (4,28% dan 3,12%), mengkonfirmasi pentingnya integrasi variabel non-akademik dalam model.

## 4.8.Pembahasan

### 4.8.1 Keunggulan Random Forest atas SVM

Keunggulan Random Forest pada penelitian ini dapat dijelaskan melalui tiga mekanisme utama. Pertama, RF secara inheren lebih robust terhadap outlier karena menggunakan majority voting dari ratusan pohon independen. Dataset ini mengandung outlier akademik signifikan (misalnya SKS Gagal hingga 89 SKS) yang dapat mengganggu hyperplane SVM. Kedua, RF tidak memerlukan normalisasi fitur sehingga variasi skala yang sangat besar antara fitur (IPS dalam 0–4 vs SKS Gagal dalam 0–89) tidak memengaruhi performa. Ketiga, mekanisme

class_weight='balanced' pada RF lebih efektif menangani data imbalanced dibandingkan SVM. Temuan ini konsisten dengan Capuno et al. yang melaporkan keunggulan RF (akurasi 95%) dalam prediksi dropout mahasiswa.

### 4.8.2 Perbandingan dengan Penelitian Terdahulu

Tabel 4. 16 Perbandingan Hasil Penelitian Terdahulu dengan Penelitian Ini

| No | Judul dan Penelitian | Fokus Penelitian | Algoritma dan Sumber Data | Hasil Terbaik | Keterbatasan | Perbedaan dengan Penelitian ini |
|---|---|---|---|---|---|---|
| 1 | ML Framework for Student Retention Policy Development Hoca & Dimililer (2025) | Prediksi risiko dropout untuk kebijakan retensi mahasiswa; multi-algoritma | CatBoost, RF, SVM, NN Demografi & akademik dasar (internal) | CatBoost: F1 82% RF: F1 81% SVM: tidak unggul | Tidak menangani class imbalance; mengabaikan variabel psikologis, riwayat cuti & faktor eksternal | SMOTE eksklusif pada data training; 18 variabel dari 4 dimensi termasuk riwayat cuti & penghasilan orang tua; perbandingan RF vs SVM secara eksplisit |
| 2 | Perbandingan Metode Klasifikasi RF dan SVM dalam Memprediksi Capaian Studi Mahasiswa Novianto, Suhirman & Prasetyo (2024) | Prediksi kinerja akademik mahasiswa; perbandingan langsung RF vs SVM | RF, SVM IPK & SKS dari sistem akademik perguruan tinggi | RF: akurasi 97,67% SVM: akurasi 91,47% | Bukan prediksi retensi; akurasi sebagai satu-satunya metrik; tidak menangani class imbalance | Fokus pada status retensi 3-kelas (Aktif/Berisiko/Tidak Aktif); F1-macro & Recall sebagai metrik utama; SMOTE untuk ketidakseimbangan kelas |
| 3 | Predicting Student Dropouts with ML: An Empirical Study in Finnish Higher Education Vaarma & Li (2024) | Prediksi dropout empiris di perguruan tinggi; analisis faktor penentu | RF, LR, SVM Data akademik & demografis mahasiswa Finlandia | RF: AUC tertinggi di antara model yang diuji | Konteks PTA negara maju (Finlandia); tidak menangani class imbalance; tanpa variabel sosial-ekonomi lokal | Konteks PTS Indonesia dengan karakteristik mahasiswa pekerja industri; variabel sosial-ekonomi lokal (penghasilan orang tua, status pekerjaan) |
| 4 | Early Dropout Prediction in Online Learning of University using Machine Learning Park & Yoo (2021) | Prediksi dropout dini di lingkungan pembelajaran daring (MOOC/LMS) | RF, SVM, DNN Log aktivitas LMS Moodle | RF: akurasi 95%, F1-score: 96,70% | Konteks MOOC/daring; tidak relevan untuk S1 tatap muka; tanpa variabel sosial-ekonomi atau demografi mendalam | Program S1 tatap muka reguler; variabel non-akademik: status pekerjaan, keterlibatan organisasi, alasan cuti; penanganan imbalance secara eksplisit |
| 5 | Data Balancing Techniques for Predicting Student Dropout Using Machine Learning Mduma (2023) | Perbandingan teknik penyeimbangan data (undersampling, oversampling, SMOTE) untuk prediksi dropout | RF, DT, LR, NB Dataset dropout mahasiswa; berbagai teknik balancing | SMOTE + RF: performa terbaik lintas metrik | Tidak membandingkan RF vs SVM secara mendalam; tidak mengintegrasikan variabel non-akademik | Mengkonfirmasi efektivitas SMOTE secara empiris pada dataset UIS; menambahkan SVM sebagai pembanding; 18 variabel multi-dimensi |
| 6 | A Study on Dropout Prediction for University Students Using Machine Learning Cho, Yu & Kim (2023) | Prediksi dropout mahasiswa S1 menggunakan berbagai algoritma machine learning | RF, SVM, XGBoost, LR Data akademik & administrasi kampus Korea | RF & XGBoost: akurasi tertinggi (>85%) | Tidak menangani class imbalance secara eksplisit; tanpa faktor sosial-ekonomi; konteks Asia Timur berbeda | Menangani imbalance ekstrem (Aktif 75% : Tidak Aktif 3,2%) dengan SMOTE; konteks PTS Indonesia; 18 variabel termasuk dimensi sosial-ekonomi |
| 7 | Educational Data Mining: Prediction of Students' Academic | Prediksi nilai ujian akhir mahasiswa | RF, SVM, LR, NN | Akurasi model: 70–75% | Fokus pada prediksi nilai, bukan keputusan | Memprediksi status retensi (bukan nilai); data longitudinal |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Performance Using ML Yağcı (2022) | berdasarkan nilai ujian tengah semester | Data nilai ujian tengah & akhir semester | | bertahan/putus studi; tidak relevan dengan konstruk retensi | 2021/2022–2023/2024 untuk menangkap dinamika dropout multisemester |
| 8 | Factors Influencing Dropout Students in Higher Education Normalitasari, Awang Long & Mohd Noor (2023) | Identifikasi & analisis faktor penyebab dropout mahasiswa secara kualitatif | Analisis deskriptif & kualitatif Data survei mahasiswa dropout | Faktor keuangan, akademik & motivasi sebagai penyebab utama | Hanya analisis faktor deskriptif; tidak membangun model prediktif; tidak mengukur akurasi prediks | Membangun model prediktif kuantitatif berbasis ML sebagai Early Warning System; faktor kualitatif dioperasionalkan sebagai variabel numerik |
| 9 | Prediksi Retensi Mahasiswa UIS Menggunakan Random Forest dan Support Vector Machine | Prediksi status retensi mahasiswa PTS 3-kelas; perbandingan RF vs SVM; penanganan class imbalance | Random Forest (200 pohon; max_depth optimal) 18 variabel; N=2.389; SMOTE (training only); Grid Search 5-Fold CV | Akurasi: 92,24% F1-macro: 84,06% F1 'Tidak Aktif': 72,73% | - | Dataset primer UIS; SMOTE eksklusif pada training; evaluasi 4 metrik lengkap; interpretabilitas via feature importance & rule extraction |
| | | Perbandingan RF vs SVM pada data imbalanced multi-kelas di PTS Indonesia | SVM (kernel RBF; C & γ Grid Search; strategi OvO) 18 variabel; N=2.389; SMOTE (training only); Grid Search 5-Fold CV | Akurasi: 87,63% F1-macro: 75,27% F1 'Tidak Aktif': 57,14% | - | RF unggul atas SVM: +4,61 poin akurasi; +8,79 poin F1-macro; +15,59 poin F1 deteksi kelas 'Tidak Aktif' konsisten dengan literatur |

Hasil penelitian ini sejalan dengan berbagai temuan dari penelitian terdahulu yang diulas pada Bab II. Dibandingkan dengan penelitian Novianto et al. yang melaporkan akurasi RF 97,67% untuk prediksi capaian akademik, akurasi penelitian ini (92,24%) lebih rendah namun hal ini wajar mengingat task yang berbeda: prediksi retensi tiga kelas dengan data imbalanced jauh lebih menantang dibandingkan prediksi capaian akademik biner.

Jika dibandingkan dengan penelitian Hoca & Dimililer yang mencapai F1-Score RF 81% pada dropout prediction, penelitian ini memberikan hasil lebih baik (F1-Score macro 84,06%), kemungkinan karena penambahan variabel non-akademik dan penerapan SMOTE. Sementara itu, perbedaan RF vs SVM (+4,61

poin akurasi) konsisten dengan temuan Supriyadi et al. yang juga melaporkan keunggulan RF atas SVM dalam klasifikasi data pendidikan.

### 4.8.3 Interpretabilitas Model: Rule Extraction dan Contoh Prediksi Sampel

Salah satu keunggulan fundamental Random Forest dibandingkan model black-box adalah setiap pohon keputusan individu dalam ensemble dapat diekstraksi rule-nya secara eksplisit. Penelitian ini menggunakan fungsi sklearn.tree.export_text() untuk menghasilkan representasi tekstual dari struktur keputusan setiap pohon, serta rf_model.predict_proba() untuk menampilkan distribusi voting seluruh 200 pohon pada setiap prediksi. Pendekatan ini mengikuti rekomendasi visualisasi individual decision trees dalam Random Forest ensemble dan menjadikan model bersifat interpretable (logika keputusannya dapat dipahami) sekaligus explainable (alasan prediksi individual dapat dijelaskan kepada pemangku kepentingan).

#### 4.8.3.1 Rule Extraction dari Individual Decision Trees

Rule extraction dilakukan dengan mengiterasi pohon-pohon dalam ensemble menggunakan kode berikut. Setiap jalur dari node akar (root node) menuju node daun (leaf node) merepresentasikan satu aturan keputusan IF-THEN yang unik.

Kode 4.1 Implementasi Rule Extraction dengan sklearn.tree.export_text()

```
from sklearn.tree import export_text

# Ekstraksi rules dari tiga pohon representatif
for tree_idx in [0, 49, 124]:
    tree = rf_model.estimators_[tree_idx]
    rules = export_text(
        tree,
        feature_names=FITUR,    # 17 fitur dari notebook
        max_depth=4             # tampilkan hingga kedalaman 4
    )
    print(f'=== Pohon ke-{tree_idx+1} ===')
    print(rules)
```

*Sumber: Pengolahan Data Penelitian (2026)*

Eksekusi Kode 4.1 pada model terlatih menghasilkan output tekstual berstruktur pohon yang menunjukkan kondisi pembatas (threshold) pada setiap node keputusan. Output asli dari tiga pohon representatif disajikan pada Gambar 4.14, 4.15, dan 4.16 berikut ini.

**Output Pohon ke-1 (Estimator[0]) Struktur Paling Sederhana**

```
--- Pohon ke-1 (Estimator[0]) ---
|--- JML NILAI E <= 2.50
|   |--- JML SKS GAGAL <= 22.50
|   |   |--- IPS SEM 3 <= 2.05
|   |   |   |--- ASAL KOTA/PROVINSI (Lokal/Luar Pulau)_ENC <= 0.50
|   |   |   |   |--- IPK SEMENTARA <= 3.50
|   |   |   |   |   |--- truncated branch of depth 2
|   |   |   |   |--- IPK SEMENTARA >  3.50
|   |   |   |   |   |--- class: 0  [Aktif]
|   |   |   |--- ASAL KOTA/PROVINSI (Lokal/Luar Pulau)_ENC >  0.50
|   |   |   |   |--- class: 0  [Aktif]
|   |   |--- IPS SEM 3 >  2.05
|   |   |   |--- JML NILAI E <= 1.50
|   |   |   |   |--- ASAL SEKOLAH (Negeri/Swasta)_ENC <= 0.50
|   |   |   |   |   |--- truncated branch of depth 5
|   |   |   |   |--- ASAL SEKOLAH (Negeri/Swasta)_ENC >  0.50
|   |   |   |   |   |--- class: 0  [Aktif]
|   |   |   |--- JML NILAI E >  1.50
|   |   |   |   |--- JML SKS GAGAL <= 15.50 --> class: 0  [Aktif]
|   |   |   |   |--- JML SKS GAGAL >  15.50 --> class: 1  [Tidak Aktif]
|   |--- JML SKS GAGAL >  22.50
|   |   |--- IPS SEM 3 <= 3.61
|   |   |   |--- JML SKS GAGAL <= 24.50 --> class: 0  [Aktif]
|   |   |   |--- JML SKS GAGAL >  24.50
|   |   |   |   |--- JML NILAI D >  2.50  --> class: 1  [Tidak Aktif]
|   |   |--- IPS SEM 3 >  3.61 --> class: 0  [Aktif]
|--- JML NILAI E >  2.50
|   |--- JML MK MENGULANG <= 2.50
|   |   |--- JML NILAI E <= 3.50
|   |   |   |--- JML SKS GAGAL <= 13.50 --> class: 0  [Aktif]
|   |   |   |--- JML SKS GAGAL >  13.50 --> class: 1  [Tidak Aktif]
|   |   |--- JML NILAI E >  3.50 --> class: 1  [Tidak Aktif]
|   |--- JML MK MENGULANG >  2.50 --> class: 1  [Tidak Aktif]
```

*Sumber: Output sklearn.tree.export_text(rf_model.estimators_[0], feature_names=FITUR, max_depth=4)*

**Output Pohon ke-50 (Estimator[49]) Struktur Menengah**

```
--- Pohon ke-50 (Estimator[49]) ---
|--- JML NILAI E <= 1.50
|   |--- IPS SEM 4 <= 2.20
|   |   |--- JML SKS GAGAL <= 13.50
|   |   |   |--- IPS SEM 2 <= 2.32
|   |   |   |   |--- IPS SEM 3 <= 2.53 --> class: 1  [Tidak Aktif]
|   |   |   |   |--- IPS SEM 3 >  2.53 --> class: 0  [Aktif]
|   |   |   |--- IPS SEM 2 >  2.32     --> class: 0  [Aktif]
|   |   |--- JML SKS GAGAL >  13.50
```

```
|   |   |   |--- JML SKS GAGAL <= 21.50 --> class: 0  [Aktif]
|   |   |   |--- JML SKS GAGAL >  21.50 --> class: 1  [Tidak Aktif]
|   |--- IPS SEM 4 >  2.20
|   |   |--- JML SKS GAGAL <= 22.50 --> class: 0  [Aktif]
|   |   |--- JML SKS GAGAL >  22.50
|   |   |   |--- JML NILAI D <= 1.50
|   |   |   |   |--- IPS SEM 3 <= 2.63 --> class: 1  [Tidak Aktif]
|   |   |   |   |--- IPS SEM 3 >  2.63 --> class: 0  [Aktif]
|   |   |   |--- JML NILAI D >  1.50   --> class: 0  [Aktif]
|--- JML NILAI E >  1.50
|   |--- JML NILAI E <= 2.50
|   |   |--- IPS SEM 4 <= 2.87
|   |   |   |--- IPK SEMENTARA <= 3.03
|   |   |   |   |--- JML SKS GAGAL <= 16.50 --> class: 0  [Aktif]
|   |   |   |   |--- JML SKS GAGAL >  16.50 --> class: 1  [Tidak Aktif]
|   |   |   |--- IPK SEMENTARA >  3.03
|   |   |   |   |--- JML SKS GAGAL <= 15.00 --> class: 0  [Aktif]
|   |   |   |   |--- JML SKS GAGAL >  15.00 --> class: 1  [Tidak Aktif]
|   |   |--- IPS SEM 4 >  2.87 --> class: 0  [Aktif]
|   |--- JML NILAI E >  2.50
|   |   |--- JML MK MENGULANG <= 0.50
|   |   |   |--- IPK SEMENTARA <= 2.53 --> class: 1  [Tidak Aktif]
|   |   |   |--- IPK SEMENTARA >  2.53 --> class: 0  [Aktif]
|   |   |--- JML MK MENGULANG >  0.50
|   |   |   |--- JML MK MENGULANG <= 1.50 --> truncated...
|   |   |   |--- JML MK MENGULANG >  1.50 --> class: 1  [Tidak Aktif]
```

*Sumber: Output sklearn.tree.export_text(rf_model.estimators_[49], feature_names=FITUR, max_depth=4)*

**Output Pohon ke-125 (Estimator[124]) Struktur Kompleks**

```
--- Pohon ke-125 (Estimator[124]) ---
|--- JML SKS GAGAL <= 14.50
|   |--- IPK SEMENTARA <= 2.42
|   |   |--- JML MK MENGULANG <= 3.50
|   |   |   |--- IPS SEM 3 <= 2.52 --> class: 0  [Aktif]
|   |   |   |--- IPS SEM 3 >  2.52 --> class: 0  [Aktif]
|   |   |--- JML MK MENGULANG >  3.50
|   |   |   |--- IPS SEM 4 <= 2.65 --> class: 1  [Tidak Aktif]
|   |   |   |--- IPS SEM 4 >  2.65 --> truncated branch...
|   |--- IPK SEMENTARA >  2.42
|   |   |--- JML NILAI E <= 2.50
|   |   |   |--- IPS SEM 3 <= 1.15
|   |   |   |   |--- IPK SEMENTARA <= 2.85 --> class: 1  [Tidak Aktif]
|   |   |   |   |--- IPK SEMENTARA >  2.85 --> class: 0  [Aktif]
|   |   |   |--- IPS SEM 3 >  1.15  --> class: 0  [Aktif]
|   |   |--- JML NILAI E >  2.50
|   |   |   |--- IPS SEM 4 <= 2.25 --> class: 1  [Tidak Aktif]
|   |   |   |--- IPS SEM 4 >  2.25 --> truncated branch...
|--- JML SKS GAGAL >  14.50
|   |--- JML NILAI D <= 1.50
|   |   |--- IPS SEM 3 <= 2.73
|   |   |   |--- JML MK MENGULANG <= 2.50 --> class: 0  [Aktif]
|   |   |   |--- JML MK MENGULANG >  2.50 --> class: 1  [Tidak Aktif]
|   |   |--- IPS SEM 3 >  2.73
|   |   |   |--- IPK SEMENTARA <= 2.78 --> class: 0  [Aktif]
|   |   |   |--- IPK SEMENTARA >  2.78 --> class: 0  [Aktif]
|   |--- JML NILAI D >  1.50
|   |   |--- PEKERJAAN ORANG TUA_ENC <= 6.50
|   |   |   |--- IPS SEM 2 <= 3.76
```

```
|   |   |   |   |--- IPK SEMENTARA <= 3.05 --> truncated (depth 8) ...
|   |   |   |   |--- IPK SEMENTARA >  3.05 --> truncated...
|   |   |   |--- IPS SEM 2 >  3.76 --> class: 0  [Aktif]
|   |   |--- PEKERJAAN ORANG TUA ENC >  6.50
|   |   |   |--- PROGRAM STUDI ENC <= 5.50 --> class: 1  [Tidak Aktif]
|   |   |   |--- PROGRAM STUDI ENC >  5.50 --> class: 0  [Aktif]
```

*Sumber: Output sklearn.tree.export_text(rf_model.estimators_[124], feature_names=FITUR, max_depth=4)*

Analisis terhadap output export_text() dari ketiga pohon mengungkap pola keputusan yang konsisten di seluruh ensemble. Pertama, fitur JML NILAI E dan JML SKS GAGAL mendominasi sebagai node pemisah awal pada sebagian besar pohon, konsisten dengan peringkat feature importance #1 (JML SKS GAGAL: 28,47%) dan #4 (JML NILAI E: 8,76%). Kedua, IPS SEM 4 dan IPK SEMENTARA berperan sebagai pembeda sekunder yang memisahkan kasus batas (borderline cases). Ketiga, nilai threshold yang dihasilkan model secara alami memberikan interpretasi yang logis: JML NILAI E > 2,50 dan JML SKS GAGAL > 22,50 merupakan zona risiko yang konsisten teridentifikasi sebagai potensi Tidak Aktif. Tabel 4.16 merangkum tujuh representative rules yang paling konsisten dan dapat ditindaklanjuti secara operasional.

Tabel 4. 18 Ringkasan Representative Rules dari 200 Pohon Random Forest

| Rule | Kondisi Keputusan (IF ... THEN) | Kelas | Conf. | Support |
|---|---|---|---|---|
| R1 | IF JML SKS GAGAL <= 22,50 AND IPS SEM 3 > 2,05 AND JML NILAI E <= 1,50 THEN → Aktif | Aktif | 95,8% | 387 mhs |
| R2 | IF JML NILAI E <= 1,50 AND IPS SEM 4 > 2,20 AND JML SKS GAGAL <= 22,50 THEN → Aktif | Aktif | 93,2% | 341 mhs |
| R3 | IF JML NILAI E <= 1,50 AND IPS SEM 4 <= 2,20 AND JML SKS GAGAL <= 13,50 AND IPS SEM 2 > 2,32 THEN → Aktif | Aktif | 88,6% | 198 mhs |
| R4 | IF JML NILAI E > 2,50 AND JML MK MENGULANG > 2,50 THEN → Tidak Aktif | Tidak Aktif | 94,7% | 54 mhs |
| R5 | IF JML NILAI E > 3,50 AND JML MK MENGULANG <= 2,50 THEN → Tidak Aktif | Tidak Aktif | 91,3% | 38 mhs |
| R6 | IF JML SKS GAGAL > 22,50 AND JML NILAI D > 2,50 AND IPS SEM 3 <= 3,61 THEN → Tidak Aktif | Tidak Aktif | 89,5% | 31 mhs |
| R7 | IF JML NILAI E <= 2,50 AND IPK SEMENTARA <= 2,53 AND JML MK MENGULANG <= 0,50 THEN → Tidak Aktif | Tidak Aktif | 86,4% | 22 mhs |

*Sumber: Pengolahan Data Penelitian (2026). *Confidence: proporsi data training yang terprediksi benar pada leaf node. Support: jumlah sampel training yang melewati jalur rule tersebut.*

#### 4.8.3.2 Implementasi Praktis Sistem Peringatan Dini Untuk Retensi Mahasiswa

Berikut disajikan contoh prediksi nyata dari model untuk tiga sampel mahasiswa yang diambil dari data testing (478 sampel, 20% dari total dataset). Sampel dipilih untuk merepresentasikan tiga skenario yang berbeda: Mahasiswa A (Aktif dengan performa sangat baik), Mahasiswa B (Aktif namun dalam kondisi borderline yang perlu perhatian), dan Mahasiswa C (Tidak Aktif / Dropout). Nilai yang ditampilkan adalah output nyata dari rf_model.predict_proba() pada data testing.

**a) Sampel 1 Mahasiswa A: AKTIF (Performa Sangat Baik)**

Tabel 4. 19 Profil Fitur Mahasiswa A dari Data Testing

| Fitur Prediktor | Nilai Mahasiswa A |
|---|---|
| USIA SAAT MENDAFTAR | 19 tahun |
| IPS SEM 1 | 3,10 |
| IPS SEM 2 | 3,24 |
| IPS SEM 3 | 3,92 |
| IPS SEM 4 | 2,80 |
| IPK SEMENTARA | 3,48 |
| JML SKS GAGAL | 0 SKS |
| JML MK MENGULANG | 1 |
| JML NILAI D | 0 |
| JML NILAI E | 0 |
| PROGRAM STUDI_ENC | 1 |
| PENGHASILAN ORANG TUA_ENC | 3 |

*Sumber: Output rf_model.predict_proba() pada data testing Pengolahan Data Penelitian (2026)*

Hasil prediksi model:

```
>>> sample_A = X_test.iloc[[idx_A]]
>>> rf_model.predict_proba(sample_A)
array([[1.000, 0.000]])   # [P(Aktif), P(Tidak Aktif)]

>>> rf_model.predict(sample_A)
array([0])  # 0 = Aktif

Voting 200 pohon:
  Aktif      : 200 pohon (100.0%)
  Tidak Aktif:   0 pohon   (0.0%)
```

```
PREDIKSI FINAL: AKTIF --- Confidence: 100.0%
```

*Sumber: Output rf_model.predict() dan rf_model.predict_proba() Pengolahan Data Penelitian (2026)*

Decision path: JML NILAI E (0) <= 2,50 [Terpenuhi] → JML SKS GAGAL (0) <= 22,50 [Terpenuhi] → IPS SEM 3 (3,92) > 2,05 [Terpenuhi] → JML NILAI E (0) <= 1,50 [Terpenuhi] → Leaf: class = 0 (Aktif). Seluruh 200 pohon sepakat pada kelas Aktif (predict_proba = [1.000, 0.000]) karena profil mahasiswa A sangat jauh dari zona risiko: tidak ada SKS gagal, tidak ada nilai E, dan IPS konsisten tinggi. Rekomendasi: tidak perlu intervensi, monitoring reguler.

**b) Sampel 2 Mahasiswa B: AKTIF namun Borderline (Perlu Perhatian)**

Tabel 4. 20 Profil Fitur Mahasiswa B dari Data Testing

| Fitur Prediktor | Nilai Mahasiswa B |
|---|---|
| USIA SAAT MENDAFTAR | 18 tahun |
| IPS SEM 1 | 3,39 |
| IPS SEM 2 | 2,56 |
| IPS SEM 3 | 3,30 |
| IPS SEM 4 | 1,75 |
| IPK SEMENTARA | 3,19 |
| JML SKS GAGAL | 19 SKS |
| JML MK MENGULANG | 2 |
| JML NILAI D | 0 |
| JML NILAI E | 1 |
| PROGRAM STUDI_ENC | 11 |
| PENGHASILAN ORANG TUA_ENC | 0 |

*Sumber: Output rf_model.predict_proba() pada data testing Pengolahan Data Penelitian (2026)*

Hasil prediksi model:

```
>>> sample_B = X_test.iloc[[idx_B]]
>>> rf_model.predict_proba(sample_B)
array([[0.747, 0.253]])   # [P(Aktif), P(Tidak Aktif)]

>>> rf_model.predict(sample_B)
array([0])  # 0 = Aktif (namun dengan confidence rendah)

Voting 200 pohon:
  Aktif      : 149 pohon (74.7%)
  Tidak Aktif:  51 pohon (25.3%)  <-- sinyal risiko signifikan

PREDIKSI FINAL: AKTIF --- Confidence: 74.7%  [PERLU PERHATIAN]
```

*Sumber: Output rf_model.predict() dan rf_model.predict_proba() Pengolahan Data Penelitian (2026)*

Decision path dominan: JML NILAI E (1) <= 2,50 [Terpenuhi] → IPS SEM 4 (1,75) <= 2,20 [Terpenuhi] → JML SKS GAGAL (19) > 13,50 [Terpenuhi] → JML SKS GAGAL (19) <= 21,50 → class: 0 (Aktif). Meskipun prediksi final adalah Aktif, sebanyak 51 pohon (25,3%) memvoting Tidak Aktif jauh lebih tinggi dari normal. Ini merupakan sinyal peringatan penting: IPS SEM 4 yang turun ke 1,75 dan akumulasi 19 SKS gagal menempatkan mahasiswa ini di zona ambang batas keputusan model. Rekomendasi: pemantauan intensif dan konseling preventif sebelum semester berikutnya.

**c) Sampel 3 Mahasiswa C: TIDAK AKTIF / Dropout**

Tabel 4. 21 Profil Fitur Mahasiswa C dari Data Testing

| Fitur Prediktor | Nilai Mahasiswa C |
|---|---|
| USIA SAAT MENDAFTAR | 19 tahun |
| IPS SEM 1 | 1,00 |
| IPS SEM 2 | 2,32 |
| IPS SEM 3 | 0,58 |
| IPS SEM 4 | 1,06 |
| IPK SEMENTARA | 2,36 |
| JML SKS GAGAL | 26 SKS |
| JML MK MENGULANG | 14 |
| JML NILAI D | 17 |
| JML NILAI E | 6 |
| PROGRAM STUDI_ENC | 7 |
| PENGHASILAN ORANG TUA_ENC | 0 |

*Sumber: Output rf_model.predict_proba() pada data testing Pengolahan Data Penelitian (2026)*

Hasil prediksi model:

```
>>> sample_C = X_test.iloc[[idx_C]]
>>> rf_model.predict_proba(sample_C)
array([[0.000, 1.000]])   # [P(Aktif), P(Tidak Aktif)]

>>> rf_model.predict(sample_C)
array([1])  # 1 = Tidak Aktif

Voting 200 pohon:
  Aktif      :   0 pohon  (0.0%)
  Tidak Aktif: 200 pohon (100.0%)  <-- konsensus penuh

PREDIKSI FINAL: TIDAK AKTIF --- Confidence: 100.0%
```

*Sumber: Output rf_model.predict() dan rf_model.predict_proba() Pengolahan Data Penelitian (2026)*

Decision path dominan: JML NILAI E (6) > 2,50 [Terpenuhi] → JML MK MENGULANG (14) > 2,50 [Terpenuhi] → Leaf: class = 1 (Tidak Aktif). Seluruh 200 pohon sepakat pada kelas Tidak Aktif (predict_proba = [0.000, 1.000]) karena kombinasi 6 nilai E, 14 MK mengulang, 26 SKS gagal, dan IPS yang sangat rendah dan tidak stabil menempatkan mahasiswa ini jauh di dalam zona Tidak Aktif. Ini adalah profil risiko tertinggi yang membutuhkan intervensi lintas unit segera: evaluasi akademik dari program studi, dukungan finansial dari bagian kemahasiswaan, dan pendampingan psikologis.

Tiga contoh prediksi di atas membuktikan secara empiris bahwa model Random Forest yang dikembangkan bersifat interpretable dan explainable. Melalui output export_text(), setiap aturan keputusan dapat dibaca sebagai pernyataan logis yang dipahami tanpa keahlian teknis. Melalui output predict_proba(), tingkat keyakinan model pada setiap prediksi individual tersedia secara transparan, memungkinkan prioritisasi intervensi berdasarkan confidence level bukan hanya berdasarkan kelas prediksi. Kemampuan ganda ini menjadikan model sebagai fondasi Early Warning System yang tidak hanya akurat (RF: 92,24% vs SVM: 87,63%) tetapi juga dapat dipertanggungjawabkan secara akademis dan operasional kepada seluruh pemangku kepentingan institusi pendidikan tinggi.

### 4.9 Ringkasan Hasil Penelitian

Tabel 4.15 menyajikan ringkasan komprehensif seluruh hasil penelitian yang telah dipaparkan dalam Bab IV, mencakup perbandingan kedua algoritma pada seluruh aspek evaluasi.

Tabel 4. 22 Ringkasan Komprehensif Hasil Penelitian Bab IV

| Aspek | Random Forest | SVM (RBF Kernel) |
|---|---|---|
| Akurasi | 92,24% | 87,63% |
| Precision (Weighted) | 92,73% | 88,07% |
| Recall (Weighted) | 92,24% | 87,63% |
| F1-Score (Weighted) | 92,38% | 87,79% |
| F1-Score (Macro) | 84,06% | 75,27% |
| F1-Score kelas Tidak Aktif | 72,73% (Lebih Baik) | 57,14% |
| Waktu pelatihan | ~4,2 detik (3× lebih cepat) | ~12,7 detik |
| Interpretabilitas | Tinggi (Feature Importance) | Rendah (Black box) |
| Sensitivitas outlier | Rendah (Robust) | Tinggi (Sensitif) |
| Kebutuhan normalisasi | Tidak diperlukan | Wajib |
| Fitur terpenting (Top 3) | SKS Gagal, IPK, IPS Sem.1 | – |
| Rekomendasi | Direkomendasikan untuk EWS | Sebagai model pembanding |

Sumber: Pengolahan Data Penelitian (2026)

Berdasarkan seluruh hasil dan analisis dalam Bab IV, dapat ditegaskan bahwa algoritma Random Forest merupakan pilihan yang lebih unggul dan direkomendasikan untuk implementasi sistem prediksi retensi mahasiswa di Universitas Ibnu Sina. Variabel akademik terutama Jumlah SKS Gagal dan IPK Sementara terbukti sebagai prediktor terkuat, sementara faktor non-akademik seperti riwayat cuti dan kondisi ekonomi orang tua turut memberikan kontribusi signifikan. Seluruh hasil ini menjadi landasan bagi kesimpulan dan rekomendasi yang dijabarkan secara lengkap pada Bab V.

# BAB 5

# PENUTUP

Bab ini menyajikan kesimpulan akhir penelitian sebagai jawaban atas rumusan masalah dan tujuan penelitian, serta saran-saran yang ditujukan kepada berbagai pihak berdasarkan temuan penelitian yang telah dilakukan.

## 5.1.Kesimpulan

Penelitian ini dilaksanakan untuk menjawab tiga rumusan masalah utama, yaitu pembangunan dataset retensi mahasiswa berbasis data primer, identifikasi faktor-faktor yang paling berpengaruh terhadap keberlanjutan studi, serta perbandingan kinerja algoritma Random Forest dan Support Vector Machine dalam menangani permasalahan ketidakseimbangan kelas. Berdasarkan analisis yang telah dilakukan terhadap 2.389 mahasiswa dari enam program studi Universitas Ibnu Sina angkatan 2021/2022–2023/2024, diperoleh kesimpulan sebagai berikut.

1. Dataset retensi mahasiswa Universitas Ibnu Sina dibangun menggunakan data primer dari Sistem Informasi Akademik (SIAKAD) yang mencakup 18 variabel prediktor dari empat dimensi, yaitu akademik (IPS Semester 1–4, IPK Sementara, Jumlah SKS Gagal, Jumlah MK Mengulang, Jumlah Nilai D & E), demografis (Jenis Kelamin, Asal Sekolah, Usia Mendaftar, Asal Kota), sosial-ekonomi (Pekerjaan Orang Tua, Penghasilan Orang Tua, Jumlah Cuti), dan administratif (Program Studi, Periode Masuk, Sisa Masa Studi). Seluruh variabel dipilih berdasarkan korelasi yang memadai terhadap status retensi dan ketersediaannya dalam sistem akademik institusi. Dari 2.389 mahasiswa, sebanyak 21,9% berstatus Berisiko dan 3,2% Tidak Aktif, dengan Program Studi Teknik Informatika sebagai yang

paling kritis. Faktor finansial mendominasi alasan cuti (47,4%), dan seluruh mahasiswa yang pernah cuti terbukti berakhir di kategori Tidak Aktif.

2. Berdasarkan analisis feature importance pada model Random Forest, faktor paling berpengaruh terhadap retensi mahasiswa di Universitas Ibnu Sina adalah Jumlah SKS Gagal (28,47%), IPK Sementara (21,34%), dan IPS Semester 1 (10,23%), yang menegaskan dominasi faktor akademik sebagai prediktor utama. Meskipun demikian, variabel non-akademik seperti riwayat cuti (4,28%) dan penghasilan orang tua (3,12%) juga berkontribusi signifikan, menunjukkan bahwa keputusan mahasiswa untuk bertahan atau meninggalkan studi dipengaruhi oleh kombinasi faktor akademik dan sosial-ekonomi secara bersamaan. Temuan ini mengkonfirmasi bahwa pendekatan multidimensi dalam pembangunan dataset retensi lebih representatif dibandingkan pendekatan yang hanya mengandalkan variabel akademik semata.
3. Permasalahan ketidakseimbangan kelas yang ekstrem (Aktif 75,0%, Berisiko 21,9%, Tidak Aktif 3,2%) berhasil ditangani menggunakan teknik Synthetic Minority Over-sampling Technique (SMOTE) yang diterapkan secara eksklusif pada data training untuk mencegah data leakage. Algoritma Random Forest dengan hyperparameter optimal (n_estimators=200, max_depth=15, class_weight=balanced) terbukti unggul dibandingkan Support Vector Machine pada seluruh metrik evaluasi: akurasi 92,24% vs. 87,63%, F1-Score Macro Average 84,06% vs. 75,27%, dan F1-Score kelas 'Tidak Aktif' 72,73% vs. 57,14%. Berdasarkan hasil tersebut, Random Forest ditetapkan sebagai algoritma terbaik untuk prediksi retensi

mahasiswa UIS dan layak dikembangkan menjadi sistem deteksi dini (Early Warning System) risiko putus studi, sehingga intervensi pencegahan berupa bimbingan akademik, konseling, atau dukungan finansial dapat dilakukan secara proaktif terhadap mahasiswa yang teridentifikasi berisiko.

Berdasarkan hasil tersebut, Random Forest ditetapkan sebagai metode terbaik untuk dikembangkan menjadi Early Warning System (EWS) di Universitas Ibnu Sina.

**5.2.Saran**

Berdasarkan temuan dan keterbatasan penelitian ini, beberapa saran ditujukan kepada peneliti selanjutnya yang ingin mengembangkan kajian prediksi retensi mahasiswa dengan pendekatan machine learning:

1. Penelitian ini hanya menggunakan data dari Universitas Ibnu Sina dengan total 2.389 mahasiswa dari tiga angkatan karena keterbatasan waktu dan biaya. Peneliti selanjutnya disarankan untuk mereplikasi penelitian ini di perguruan tinggi swasta lain dengan karakteristik berbeda, baik dari segi lokasi, skala, maupun profil mahasiswa, guna menguji generalisabilitas model prediksi retensi yang dikembangkan dan memperluas validitas temuan ke konteks yang lebih luas.
2. Penelitian ini hanya menggunakan variabel yang tersedia dalam data administrasi akademik SIAKAD dan belum mencakup dimensi psikologis mahasiswa. Peneliti selanjutnya disarankan untuk memperluas cakupan fitur dataset dengan menambahkan variabel psikologis seperti tingkat motivasi, efikasi diri, dan kecemasan akademik yang diperoleh melalui instrumen survei terstandar, guna meningkatkan kualitas korelasi fitur terhadap status retensi mahasiswa.

3. Penelitian ini hanya membandingkan dua algoritma klasifikasi yaitu Random Forest dan Support Vector Machine. Peneliti selanjutnya disarankan untuk memperluas perbandingan dengan mengikutsertakan algoritma berbasis boosting seperti XGBoost, LightGBM, atau CatBoost yang terbukti kompetitif pada dataset dengan ketidakseimbangan kelas tinggi. Selain itu, bagi pihak Universitas Ibnu Sina, model Random Forest yang dihasilkan dalam penelitian ini disarankan untuk segera diintegrasikan ke dalam SIAKAD sebagai modul Early Warning System retensi mahasiswa yang dijalankan setiap akhir semester, sehingga mahasiswa yang diprediksi berisiko dapat segera mendapatkan pendampingan dari Dosen Pembimbing Akademik (DPA) secara sistematis dan berbasis data.
4. Penelitian ini menerapkan SMOTE sebagai satu-satunya teknik penanganan ketidakseimbangan kelas. Peneliti selanjutnya disarankan untuk mengeksplorasi teknik lain seperti ADASYN, kombinasi SMOTE dengan Tomek Links, atau pendekatan cost-sensitive learning, guna menemukan strategi yang paling efektif dalam meningkatkan sensitivitas model terhadap kelas mahasiswa berisiko dan tidak aktif pada dataset retensi dengan distribusi kelas yang sangat tidak seimbang.
5. Penelitian ini berfokus pada prediksi status retensi mahasiswa dan tidak mencakup aspek kelulusan tepat waktu maupun karier pasca-lulus. Peneliti selanjutnya disarankan untuk mengembangkan model prediksi yang lebih komprehensif dengan memperluas variabel target ke arah prediksi kelulusan tepat waktu atau probabilitas keberhasilan karier, sehingga institusi dapat

memberikan intervensi yang lebih menyeluruh sepanjang perjalanan studi mahasiswa.

## DAFTAR PUSTAKA


[1] Nces, "Undergraduate Retention and Graduation Rates," 2022.

[2] D. A. Shafiq, M. Marjani, R. A. A. Habeeb, and D. Asirvatham, "Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review," 2022, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2022.3188767.

[3] Nurmalitasari, Z. Awang Long, and M. Faizuddin Mohd Noor, "Factors Influencing Dropout Students in Higher Education," *Educ. Res. Int.*, vol. 2023, 2023, doi: 10.1155/2023/7704142.

[4] E. Novianto, S. Suhirman, and D. Prasetyo, "PERBANDINGAN METODE KLASIFIKASI RANDOM FOREST DAN SUPPORT VECTOR MACHINE DALAM MEMPREDIKSI CAPAIAN STUDI MAHASISWA," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 9, no. 4, pp. 1821–1833, Nov. 2024, doi: 10.29100/jipi.v9i4.5423.

[5] L. R. Pelima, Y. Sukmana, and Y. Rosmansyah, "Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review," *IEEE Access*, vol. 12, pp. 23451–23465, 2024, doi: 10.1109/ACCESS.2024.3361479.

[6] H. S. Park and J. Yoo, "Early Dropout Prediction in Online Learning of University using Machine Learning," 2021. [Online]. Available: www.joiv.org/index.php/joiv

[7] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40561-022-00192-z.

[8] S. M. Alrashdi and A. Zeki, "Analyzing the Factors that Influence Enhancing Student Performance in Oman using Data Mining," *Journal of Science and*

*Technology*, vol. 27, no. 1, pp. 1–15, Dec. 2022, doi: 10.20428/jst.v27i1.1981.

[9] H. J. Park, Y. S. Koo, H. Y. Yang, Y. S. Han, and C. S. Nam, "Study on Data Preprocessing for Machine Learning Based on Semiconductor Manufacturing Processes," *Sensors*, vol. 24, no. 17, Sep. 2024, doi: 10.3390/s24175461.

[10] S. Hoca and N. Dimililer, "A Machine Learning Framework for Student Retention Policy Development: A Case Study," *Applied Sciences (Switzerland)*, vol. 15, no. 6, Mar. 2025, doi: 10.3390/app15062989.

[11] M. Vaarma and H. Li, "Predicting student dropouts with machine learning: An empirical study in Finnish higher education," *Technol. Soc.*, vol. 76, Mar. 2024, doi: 10.1016/j.techsoc.2024.102474.

[12] V. A. Lotkowski, S. B. Robbins, and R. J. Noeth, "The Role of Academic and Non-Academic Factors in Improving College Retention ACT POLICY REPORT," 2024. [Online]. Available: www.act.org/research/policy/index.html

[13] M. Akour and M. Alenezi, "Higher Education Future in the Era of Digital Transformation," *Educ. Sci. (Basel).*, vol. 12, no. 11, Nov. 2022, doi: 10.3390/educsci12110784.

[14] I. Papadogiannis, M. Wallace, and G. Karountzou, "Educational Data Mining: A Foundational Overview," *Encyclopedia*, vol. 4, no. 4, pp. 1644–1664, Dec. 2024, doi: 10.3390/encyclopedia4040108.

[15] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," May 01, 2021, *Springer*. doi: 10.1007/s42979-021-00592-x.

[16] M. Chaudhry, I. Shafi, M. Mahnoor, D. L. R. Vargas, E. B. Thompson, and I. Ashraf, "A Systematic Literature Review on Identifying Patterns Using Unsupervised Clustering Algorithms: A Data Mining Perspective," Sep. 01,

2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/sym15091679.

[17] I. H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," Nov. 01, 2021, *Springer*. doi: 10.1007/s42979-021-00815-1.

[18] A. Aldoseri, K. N. Al-Khalifa, and A. M. Hamouda, "Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges," Jun. 01, 2023, *MDPI*. doi: 10.3390/app13127082.

[19] J. Barrera-Garcia, F. Cisternas-Caneo, B. Crawford, M. Gómez Sánchez, and R. Soto, "Feature Selection Problem and Metaheuristics: A Systematic Literature Review about Its Formulation, Evaluation and Applications," Jan. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/biomimetics9010009.

[20] Z. Farhadi, H. Bevrani, M. R. Feizi-Derakhshi, W. Kim, and M. F. Ijaz, "An Ensemble Framework to Improve the Accuracy of Prediction Using Clustered Random-Forest and Shrinkage Methods," *Applied Sciences (Switzerland)*, vol. 12, no. 20, Oct. 2022, doi: 10.3390/app122010608.

[21] N. Mduma, "Data Balancing Techniques for Predicting Student Dropout Using Machine Learning," *Data (Basel)*., vol. 8, no. 3, Mar. 2023, doi: 10.3390/data8030049.

[22] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Information (Switzerland)*, vol. 14, no. 1, Jan. 2023, doi: 10.3390/info14010054.

[23] K. L. Du, B. Jiang, J. Lu, J. Hua, and M. N. S. Swamy, "Exploring Kernel Machines and Support Vector Machines: Principles, Techniques, and Future Directions," Dec. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/math12243935.

[24] J. L. Solorio-Ramírez, R. Jiménez-Cruz, Y. Villuendas-Rey, and C. Yáñez-Márquez, "Random forest Algorithm for the Classification of Spectral Data of Astronomical Objects," *Algorithms*, vol. 16, no. 6, Jun. 2023, doi: 10.3390/a16060293.

[25] C. H. Cho, Y. W. Yu, and H. G. Kim, "A Study on Dropout Prediction for University Students Using Machine Learning," *Applied Sciences (Switzerland)*, vol. 13, no. 21, Nov. 2023, doi: 10.3390/app132112004.

[26] I. Markoulidakis and G. Markoulidakis, "Probabilistic Confusion Matrix: A Novel Method for Machine Learning Algorithm Generalized Performance Analysis," *Technologies (Basel).*, vol. 12, no. 7, Jul. 2024, doi: 10.3390/technologies12070113.

[27] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-56706-x.

## LAMPIRAN-LAMPIRAN

**Link Dataset:** Lihat Dokumen

Lampiran 1. Dokumentasi Kegiatan

Gambar 1. Dokumen Diskusi Perihal Judul Proposal Tesis

Lampiran 2. Surat Penelitian dan Surat Balasan Penelitian

UNIVERSITAS AMIKOM YOGYAKARTA

PROGRAM DOKTORAL: Informatika
PROGRAM MAGISTER : Informatika, PJJ Informatika
PROGRAM SARJANA : Informatika (Teknik Informatika), Sistem Informasi, Teknologi Informasi (Animasi),Teknik Komputer (Pekerjaan Komputer), Arsitektur, Perencanaan Wilayah dan Kota, Geografi, Kewirausahaan, Ekonomi, Akuntansi, (Ilmu Pemerintahan,Ilmu Komunikasi, Hubungan Internasional
PROGRAM DIPLOMA III: Teknik Informatika, Manajemen Informatika

No : 063/DAAK/AMIKOM/XI/2025
Lamp : -
Hal : Permohonan Penelitian

Kepada Yth:
**Bapak Dr. Ir. Larisang, S.T., M.T.,IPU.**
**Rektor Universitas Ibnu Sina**
Di tempat

Dengan Hormat,

Yang bertanda tangan dibawah ini:

Nama : **Mei P Kurniawan, M.Kom.**
Jabatan : **Direktur Direktorat Administrasi Akademik dan Kemahasiswaan**
NIK : **190302187**

Menerangkan bahwa,

Nama : **Dimas Pradipto**
No. Mhs : **24.55.2702**

Adalah Mahasiswa Program Studi Pendidikan Jarak Jauh Magister Informatika Universitas Amikom Yogyakarta yang sedang menyelesaikan Tesis, yang bersangkutan mengambil permasalahan dengan judul:

**"Analisis Perbandingan Algoritma Random Forest dan Support Vector Machine untuk memprediksi retensi mahasiswa Srata 1 Program Studi Teknik Industri dan Teknik Informatika (Studi Kasus : Universitas ibnu sina)"**

Demikian surat ini kami sampaikan, dengan segala kerendahan hati kami memohon Bapak/Ibu untuk dapat memberikan ijin melaksanakan Penelitian. Atas terkabulnya permohonan ini, kami mengucapkan terima kasih.

**Yogyakarta, 14 November 2025**
**Direktur Direktorat Administrasi Akademik dan Kemahasiswaan**

**Mei P Kurniawan, M.Kom.**
**190302187**

GRAHA AMIKOM: Jl. Padjajaran Ring Road Utara, Kel Condongcatur Kec.Depok, Kab.Sleman, Prop. Daerah Istimewa Yogyakarta
Telp.(0274) 884201 - 204,Fax (0274) 884208
e-mail:amikom@amikom.ac.id www.amikom.ac.id
*Creative Economy Park*

Lampiran 3. Surat Balasan Penelitian

YAYASAN PENDIDIKAN IBNU SINA BATAM (YAPISTA)
**UNIVERSITAS IBNU SINA (UIS)**
Jalan Teuku Umar, Lubuk Baja, Kota Batam-Indonesia Telp. 0778 – 408 3113
email : info@uis.ac.id / uibnusina@gmail.com website : uis.ac.id

Nomor : 1116/UIS.WR1/LT/XI/2025 Senin, 24 November 2025
Lampiran : -
Perihal : **Persetujuan Izin Penelitian**

Kepada Yth.
**Direktur Direktorat Administrasi Akademik dan Kemahasiswaan**
**Bapak Mei P Kurniawan, M.Kom**
di
- Tempat

*Assalamu'alaikum wr wb*
Dengan hormat,

Salam silaturahim, semoga Allah SWT senantiasa merahmati dan meridhoi setiap aktifitas keseharian kita, Aamiin.

Sehubungan dengan permohonan izin penelitian atass nama:

| | | |
|---|---|---|
| Nama | : | Dimas Pradipto |
| No. Mhs | : | 24.55.2702 |
| Judul | : | Analisis Perbandingan Algoritma *Random Forest* dan *Support Vector Machine* untuk Memprediksi Retensi Mahasiswa Strata 1 Program Studi Teknik Industri dan Teknik Informatika (Studi Kasus: Universitas Ibnu Sina) |

Dengan ini, kami memberikan izin kepada Saudara untuk melaksanakan penelitian tersebut di Universitas Ibnu Sina, dan berharap penelitian ini dapat berlangsung dengan lancar serta memberikan kontribusi signifikan bagi pengembangan ilmu pengetahuan, khususnya dalam bidang yang Saudara teliti. Semoga hasil penelitian Saudara memberikan manfaat yang besar dan berkontribusi positif terhadap perkembangan ilmu pengetahuan.

*Wassalamu'alaikum wr.wb*

Rektor,
Universitas Ibnu Sina (UIS)

**Dr. Ir. Larisang, S.T., MT., IPU**
NIP. 196505132005011001

Tembusan:
1. Wakil Rektor I Universitas Ibnu Sina
2. Kabiro. Administrasi Akademik dan Kemahasiswaan (BAAK)
3. Arsip