TESIS

ANALISIS KOMPARASI ALGORITMA SUPPORT VECTOR MACHINE DAN RANDOM FOREST UNTUK MEMPREDIKSI KELULUSAN SANTRI DALAM MELANJUTKAN STUDI KE TIMUR TENGAH

(Studi kasus : Pondok Pesantren Imam Bukhari)

Disusun oleh:

Nama	**: Mahmud Zunus Amirudin**
NIM	**: 24.51.1640**
Konsentrasi	**: Business Intelligence**

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2025

TESIS

ANALISIS KOMPARASI ALGORITMA SUPPORT VECTOR MACHINE DAN RANDOM FOREST UNTUK MEMPREDIKSI KELULUSAN SANTRI DALAM MELANJUTKAN STUDI KE TIMUR TENGAH

(Studi kasus : Pondok Pesantren Imam Bukhari)

COMPARISON ANALYSIS OF SUPPORT VECTOR MACHINE AND RANDOM FOREST FOR PREDICTING THE GRADUATION OF STUDENTS IN PURSUING FURTHER STUDIES IN THE MIDDLE EAST

(Case Study: Imam Bukhori Islamics Boarding School)

Diajukan untuk memenuhi salah satu syarat mencapai derajat Pascasarjana
Program Studi S2 Informatika

Disusun oleh:

Nama	**: Mahmud Zunus Amirudin**
NIM	**: 24.51.1640**
Konsentrasi	**: Business Intelligence**

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2025

HALAMAN PERSETUJUAN

ANALISIS KOMPARASI ALGORITMA SUPPORT VECTOR MACHINE DAN RANDOM FOREST UNTUK MEMPREDIKSI KELULUSAN SANTRI DALAM MELANJUTKAN STUDI KE TIMUR TENGAH

(Studi kasus: Pondok Pesantren Imam Bukhari)

COMPARISON ANALYSIS OF SUPPORT VECTOR MACHINE AND RANDOM FOREST FOR PREDICTING THE GRADUATION OF STUDENTS IN PURSUING FURTHER STUDIES IN THE MIDDLE EAST

(Case Study: Imam Bukhori Islamics Boarding School)

yang disusun dan diajukan oleh

Mahmud Zunus Amirudin
24.51.1640

telah disetujui oleh Dosen Pembimbing Tesis
pada tanggal 10 November 2025

Dosen Pembimbing,

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

HALAMAN PENGESAHAN

ANALISIS KOMPARASI ALGORITMA SUPPORT VECTOR MACHINE DAN RANDOM FOREST UNTUK MEMPREDIKSI KELULUSAN SANTRI DALAM MELANJUTKAN STUDI KE TIMUR TENGAH

(Studi kasus: Pondok Pesantren Imam Bukhari)

COMPARISON ANALYSIS OF SUPPORT VECTOR MACHINE AND RANDOM FOREST FOR PREDICTING THE GRADUATION OF STUDENTS IN PURSUING FURTHER STUDIES IN THE MIDDLE EAST

(Case Study: Imam Bukhori Islamics Boarding School)

yang disusun dan diajukan oleh

Mahmud Zunus Amirudin
24.51.1640

Telah dipertahankan di depan Dewan Penguji
pada tanggal 10 November 2025

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
I Made Artha Agastya, S.T., M.Eng., PhD **NIK. 190302352**	
Hanif Al Fatta, S.Kom., M.Kom., Ph.D. **NIK. 190302096**	
Prof. Dr. Kusrini, M.Kom. **NIK. 190302106**	

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer
Tanggal 10 November 2025

DEKAN FAKULTAS ILMU KOMPUTER

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

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Mahmud Zunus Amirudin
NIM : 24.51.1640

Menyatakan bahwa Tesis dengan judul berikut:

ANALISIS KOMPARASI ALGORITMA SUPPORT VECTOR MACHINE DAN RANDOM FOREST UNTUK MEMPREDIKSI KELULUSAN SANTRI DALAM MELANJUTKAN STUDI KE TIMUR TENGAH

(Studi kasus: Pondok Pesantren Imam Bukhari)

Dosen Pembimbing: Prof. Dr. Kusrini, M.Kom.

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, 10 November 2025

Yang Menyatakan,

Mahmud Zunus Amirudin

HALAMAN PERSEMBAHAN

Dengan rasa syukur yang mendalam ke hadirat اللهُ سُبْحَانَهُ وَتَعَالَى, karya sederhana ini saya persembahkan kepada:

1. Istri dan anak-anak tercinta atas doa, dukungan, dan pengorbanan yang tiada henti.
2. Seluruh keluarga besar Pondok Pesantren Imam Bukhari serta civitas akademika yang telah memberikan bimbingan, motivasi, dan bantuan dalam proses penyelesaian tesis ini.

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat الله سبحانه وتعالى, atas limpahan rahmat, taufik, dan hidayah-Nya sehingga penulis dapat menyelesaikan tesis yang berjudul:

"Analisis Komparasi Algoritma Support Vector Machine dan Random Forest untuk Memprediksi Kelulusan Santri dalam Melanjutkan Studi ke Timur Tengah (Studi Kasus: Pondok Pesantren Imam Bukhari)"

Tesis ini disusun sebagai salah satu syarat untuk memperoleh gelar magister, serta sebagai bentuk kontribusi ilmiah dalam pengembangan sistem pendukung keputusan berbasis machine learning di lembaga pendidikan Islam.

Ucapan terima kasih yang sebesar-besarnya penulis sampaikan kepada:

1. **Ibu Prof. Dr. Kusrini, M.Kom.**, selaku Dosen Pembimbing yang telah memberikan arahan, bimbingan, serta motivasi selama proses penelitian dan penulisan tesis ini.
2. **Bapak Hanif Al Fatta, S.Kom., M.Kom., Ph.D. dan Bapak I Made Artha Agastya, S.T., M.Eng., PhD**, yang telah memberikan masukan berharga untuk penyempurnaan karya ini.
3. **Seluruh dosen dan staf civitas akademika** di lingkungan program studi yang telah membantu dalam berbagai proses akademik dan administratif.
4. **Keluarga besar Pondok Pesantren Imam Bukhari**, atas dukungan, kerja sama, serta izin penggunaan data penelitian.
5. **Orang tua, istri, dan anak-anak tercinta**, atas doa, kesabaran, dan dukungan yang tiada henti.

Penulis menyadari bahwa tesis ini masih jauh dari sempurna. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan demi perbaikan di masa mendatang.

Akhirnya, semoga karya ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan, khususnya dalam penerapan algoritma machine learning untuk pendidikan Islam.

Yogyakarta, 12 November 2025
Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

DAFTAR LAMBANG DAN SINGKATAN

$\phi(x)$	Fungsi transformasi fitur
C	Parameter regulasi
Gamma	Parameter kernel
GPL	General Public License
KNN	K-Nearest Neighbor
RF	Random Forest
SMOTE	Sythetic Minority Over-Sampling Technique
SVM	Support Vector Machines

DAFTAR ISTILAH

Accuracy	Persentase prediksi yang benar dari seluruh data yang diuji
Classification	Proses pengelompokan data ke dalam kategori tertentu berdasarkan pola dari data pelatihan
Cross Validation	Metode evaluasi model dengan membagi data menjadi beberapa bagian untuk menguji keandalan dan kestabilan hasil
F1-Score	Ukuran kinerja model yang menggabungkan *precision* dan *recall* secara seimbang
Feature Importance	Nilai yang menunjukkan tingkat pengaruh atau kontribusi masing-masing fitur terhadap hasil prediksi model
Feature Selection	Proses pemilihan fitur yang paling relevan terhadap variabel target untuk meningkatkan efisiensi dan akurasi model
Machine Learning	Kecerdasan buatan (*Artificial Intelligence*) yang memungkinkan komputer mempelajari pola dari data dan membuat prediksi tanpa pemrograman eksplisit
Overfitting	Kondisi ketika model terlalu menyesuaikan diri dengan data pelatihan sehingga performanya menurun pada data baru
Precision	Rasio antara jumlah prediksi benar untuk kelas tertentu dibandingkan dengan total prediksi pada kelas tersebut
Recall	Rasio antara jumlah data yang benar-benar termasuk dalam kelas tertentu yang berhasil dideteksi oleh model

INTISARI

Penelitian ini bertujuan untuk menganalisis dan membandingkan performa dua algoritma *machine learning*, yaitu *Support Vector Machine* (SVM) dan *Random Forest* (RF), dalam memprediksi kelulusan santri Pondok Pesantren Imam Bukhari yang melanjutkan studi ke Timur Tengah. Data yang digunakan meliputi nilai akademik, kemampuan bahasa Arab, serta jumlah hafalan Al-Qur'an. Metode penelitian mencakup tahap pra-pemrosesan data, pembagian data latih dan uji, pelatihan model, evaluasi menggunakan metrik akurasi, presisi, recall, dan F1-score, serta analisis *feature importance* untuk mengidentifikasi faktor dominan yang memengaruhi hasil prediksi. Hasil penelitian menunjukkan bahwa algoritma *Random Forest* memiliki akurasi tertinggi sebesar 71,14% dibandingkan *Support Vector Machine* sebesar 64,98%. Meskipun demikian, SVM menunjukkan sensitivitas lebih baik dalam mengenali santri yang berpotensi tidak lulus, sementara RF lebih stabil dalam mengklasifikasikan santri lulus. Faktor yang paling berpengaruh terhadap kelulusan santri adalah hafalan Al-Qur'an, diikuti oleh mata pelajaran keagamaan seperti Hadits, Musthalah, dan Faraidh. Kesimpulan dari penelitian ini menunjukkan bahwa *Random Forest* lebih sesuai digunakan untuk meningkatkan akurasi keseluruhan, sedangkan SVM lebih efektif dalam mendeteksi risiko ketidaklulusan. Penelitian ini diharapkan dapat menjadi dasar pengembangan sistem pendukung keputusan berbasis data bagi lembaga pendidikan Islam dalam proses seleksi dan pembinaan santri.

Kata kunci: Machine learning, Support Vector Machine, Random Forest, prediksi kelulusan, pesantren, Timur Tengah.

ABSTRACT

This study aims to analyze and compare the performance of two machine learning algorithms, namely Support Vector Machine (SVM) and Random Forest (RF), in predicting the graduation of students (santri) from Pondok Pesantren Imam Bukhari who intend to continue their studies in the Middle East. The dataset includes academic grades, Arabic language proficiency, and the number of memorized Qur'anic verses. The research method consists of data preprocessing, data splitting into training and testing sets, model training, evaluation using accuracy, precision, recall, and F1-score metrics, as well as feature importance analysis to identify the dominant factors influencing prediction outcomes. The results show that the Random Forest algorithm achieved the highest accuracy of 71.14%, compared to 64.98% for the Support Vector Machine. However, SVM demonstrated better sensitivity in identifying students at risk of not passing, while RF showed more stability in classifying successful students. The most influential factor in predicting graduation was Qur'an memorization, followed by Islamic subjects such as Hadith, Musthalah, and Faraidh. In conclusion, Random Forest is more suitable for maximizing overall prediction accuracy, whereas SVM is more effective for detecting potential non-graduation cases. This study is expected to serve as a foundation for developing data-driven decision support systems in Islamic educational institutions to assist in the selection and guidance of students.

Keywords: *Machine learning, Support Vector Machine, Random Forest, graduation prediction, Islamic boarding school, Middle East.*

BAB 1

PENDAHULUAN

1.1. Latar Belakang

Perkembangan teknologi yang pesat telah menghasilkan ledakan data yang sangat besar di berbagai sektor, salah satunya adalah pada sektor pendidikan. Adanya sistem informasi yang digunakan oleh sekolah dan pondok pesantren dapat menghasilkan banyak data yang dapat dipergunakan untuk keperluan yang lebih besar. Salah satu yang dapat dimanfaatkan adalah penggunaan dataset nilai akademik santri untuk diolah menggunakan machine learning agar dapat dipergunakan untuk memprediksi kelulusan santri untuk dapat diterima di universitas di Timur Tengah.

Ada beberapa algoritma machine learning yang biasa digunakan dalam memprediksi sebuah kasus, diantaranya adalah support vector machine dan random forest. Beberapa penelitian menunjukkan bahwa akurasi yang dimiliki metode klasifikasi SVM dan Random Forest cenderung lebih baik saat dibandingkan dengan beberapa metode klasifikasi lain. Selain itu SVM mampu melakukan klasifikasi pada data non linear serta dengan Random Forest tidak akan terjadi overfit seiring dengan penambahan jumlah pohon.

Penelitian ini menggunakan data santri di Pondok Pesantren Imam bukhari dengan didasari bahwa santri selama belajar di pondok pesantren pasti memiliki tujuan untuk melanjutkan ke jenjang yang lebih tinggi, terutama di universitas – universitas yang berada di luar negeri, seperti Saudi Arabia, Mesir, Turki dan

negara- negara lainnya. Hal ini dikarenakan dalam pembelajaran Agama Islam berasal dari negara Saudi Arabia dan banyak ulama yang mengajarkan Islam secara murni berada di negara – negara tersebut. Disamping itu belajar di negara-negara tersebut juga mendapatkan beasiswa penuh sampai dengan lulus kuliahnya.

Sistem pembelajaran di negara - negara Timur Tengah memiliki karakteristik tersendiri dibandingkan sistem pembelajaran di negara lainnya, salah satunya adalah sistem pembelajaran berbasis Al Qur'an dan Hadits Nabi Muhammad, serta dalam kegiatan pembelajarannya menggunakan bahasa Arab sebagai bahasa pengantarnya. Di samping itu pada pembelajaran di negara Timur Tengah kebanyakan mahasiswanya mengambil bidang studi Agama Islam dan turunannya, oleh karena itu rata-rata mahasiswa yang diterima di universitas di timur tengah berasal dari pondok pesantren, karena adanya kesamaan sistem belajar dan materi yang diajarkan di pondok pesantren hampir sama dengan materi yang diajarkan di universitas di Timur Tengah.

Dalam pelaksanaan ujian masuk perguruan tinggi di Timur Tengah, universitas-universitas tersebut memiliki beberapa persamaan dalam persyaratannya, antara lain mewajibkan adanya ijazah negara, wawancara calon mahasiswa dengan penguji secara online, dan juga melampirkan nilai dari sekolah asal selama periode tertentu.

Pada ujian masuk perguruan tinggi di negara Saudi Arabia, ada beberapa hal yang menjadi acuan yang wajib untuk dilampirkan atau dimasukkan sebagai syarat wajib dalam pendaftaran bea siswanya. Salah satunya adalah nilai - nilai semua mata pelajaran yang ada di ijazah (nilai semester terakhir), terjemahan seluruh

dokumen dari penerjemah yang sudah bersertifikat, data calon mahasiswa dan juga dokumen surat berkelakuan baik (yang juga sudah diterjemahkan oleh penerjemah yang sudah bersertifikat). Adapun syarat yang tidak wajib tetapi memiliki pengaruh yang besar adalah surat rekomendasi dari ulama yang dikenal oleh universitas di Saudi Arabia, biasanya diberikan oleh *mudhir* (kepala) pondok pesantren tempat calon mahasiswa itu belajar.

Oleh karena itu maka peneliti menggunakan nilai akademik dari santri sebagai variabel untuk membuat prediksi kemungkinan santri dapat diterima di universitas di Timur Tengah. Data akademik ini terdiri dari beberapa mata pelajaran yang diajarkan pada jenjang *Tsanawiyah* (setingkat SMA) yang terdiri dari beberapa mata pelajaran, antara lain : Hafalan Al Qur'an, Tajwid, Tauhid, Hadits, Fiqih, Tafsir, Ulumul Qur'an, Musthalah, Usul Fiqih, Faraidh, Al-Adab Wa As-Suluk, Tarikh Islam, Tadris, Qawa'idul Lughah, Muthala'ah, Balaghah, Adab-Lughah, Ta'bir, Bahasa Indonesia, Bahasa Inggris dan Matematika.

Selain data nilai akademik tersebut diatas, jumlah hafalan Al Qur'an dan kemampuan bahasa Arab juga digunakan sebagai variabel yang digunakan untuk memprediksi kelulusan santri dalam melanjutkan studi ke universitas di Timur Tengah. Dataset yang digunakan pada penelitian ini adalah data dari santri di pondok pesantren Imam Bukhari Karanganyar, hal ini dengan pertimbangan beberapa hal, antara lain:

1. Sistem pembelajaran di pondok pesantren ini sudah menggunakan kurikulum yang diadaptasi dari kurikulum di Saudi Arabia yang dikombinasikan dengan kurikulum Indonesia.

2. Pondok pesantren ini juga memiliki pengajar-pengajar ilmu agama yang sebagian besar lulusan dari universitas-universitas di Timur Tengah.
3. Bahasa yang digunakan selama pembelajaran di kelas pada pondok pesantren ini sudah menggunakan bahasa Arab dalam penyampaian materi yang diajarkan.
4. *Masyaikh* (Pengajar Ilmu Agama) dari Universitas di Saudi Arabia, khususnya dari Universitas Islam Madinah sering berkunjung dan memberikan tambahan ilmu kepada santri - santri di pondok pesantren ini dengan menggunakan bahasa Arab langsung.
5. Adanya pendampingan khusus kepada santri yang hendak mendaftar ke universitas di Timur tengah ketika masa pendaftaran dan di support langsung oleh seluruh civitas akademik di pondok pesantren, termasuk dalam hal pembuatan Paspor, SKCK dan juga pembuatan rekomendasi dari ustadz (pengajar). Bahkan dalam pendaftarannya disediakan laboratorium Komputer khusus untuk mempersiapkan berkas pendaftaran, melakukan proses pendaftaran dan juga pada waktu wawancara calon mahasiswa dengan pihak penguji universitas.
6. Pondok pesantren ini telah memiliki banyak alumni yang belajar di Timur Tengah, sehingga akan lebih mudah dalam mendapatkan data.

Data yang digunakan oleh peneliti antara lain data santri, data nilai akademik seluruh mata pelajaran dan juga data diterima atau tidak santri di universitas di Timur Tengah. Data nilai akademik santri tersebut diolah menggunakan 2 metode *machine learning classification*, yaitu *Support Vector Machine* (SVM) dan *Random*

Forest (RF). Dan dari kedua metode ini dibandingkan agar dapat diketahui metode yang paling akurat dalam membuat prediksi dengan studi kasus ini.

Penelitian ini memiliki beberapa manfaat bagi pondok pesantren, antara lain dengan adanya pemanfaatan metode ini diharapkan membuat pondok pesantren bisa memprediksi peserta didiknya dalam melanjutkan pendidikan di universitas di Timur Tengah. Selain itu dengan adanya penelitian ini juga diharapkan membuat pondok pesantren bisa mengetahui jenis - jenis mata pelajaran yang menjadi dasar penting untuk dapat diterima di universitas di Timur Tengah. Dengan adanya nilai akurasi yang dihasilkan, maka dapat diketahui metode mana yang paling cocok digunakan pada prediksi ini, sehingga pondok pesantren dapat lebih mengembangkan lagi penelitian ini sehingga dapat diimplementasikan kepada santri-santri yang sedang belajar di pondok pesantren.

Pondok pesantren yang memiliki variabel yang sama diharapkan akan juga dapat menggunakan penelitian ini untuk dapat diimplementasikan di pondok pesantrennya dengan minimal memiliki variabel nilai bahasa arab, jumlah hafalan Al Qur'an dan juga nilai akademik untuk beberapa mata pelajaran yang sejenis dengan yang ada di penelitian ini.

1.2. Rumusan Masalah

Berdasar latar belakang, maka rumusan masalah dari penelitian ini adalah:

1. Bagaimana strategi penanganan *dataset* yang *imbalanced* agar model prediksi memiliki performa optimal dalam memprediksi kelulusan santri ke Timur Tengah?

2. Bagaimana pengaruh pemilihan fitur (*feature selection*) dan pengaturan hyperparameter pada algoritma *Support Vector Machine* (SVM) dan Random Forest (RF) terhadap akurasi prediksi kelulusan santri?
3. Bagaimana perbandingan performa model prediksi menggunakan algoritma SVM dan RF berdasarkan metrik akurasi, *precision*, *recall*, dan *F1-score* setelah dilakukan optimasi parameter, sehingga dapat diketahui kondisi kapan masing-masing algoritma lebih tepat digunakan?

1.3. Batasan Masalah

Penelitian ini memiliki beberapa batasan untuk memastikan fokus dan arah yang jelas:

a. Menggunakan dataset yang berasal dari nilai akademik umum, nilai akademik agama, kemampuan Bahasa Arab, dan jumlah hafalan Al-Qur'an dari santri Pondok Pesantren Imam Bukhari.
b. Status diterima atau tidak diterima di universitas Timur Tengah juga menjadi bagian dataset.
c. Faktor non-akademik seperti keaktifan organisasi, surat rekomendasi lembaga, atau data demografis lain tidak dimasukkan dalam analisis ini.
d. Dataset yang digunakan berasal dari Pondok Pesantren Imam Bukhari, sehingga hasil penelitian tidak dapat digeneralisasikan ke pesantren lain atau universitas di luar Timur Tengah. Namun, hasil penelitian ini diupayakan dapat diadaptasi ke pesantren lain dengan variabel yang sama.

e. Dataset memiliki distribusi kelas yang tidak seimbang (imbalanced data), sehingga penelitian ini menggunakan teknik penyeimbangan data seperti SMOTE (Synthetic Minority Over-sampling Technique) atau metode lain untuk meningkatkan performa model. Tidak semua metode balancing data akan diuji, hanya metode yang dipilih berdasarkan studi literatur.

f. Dataset dibagi menjadi 80% data latih dan 20% data uji untuk memastikan evaluasi model yang adil. Teknik validasi cross-validation (k-fold atau stratified) tidak digunakan dalam penelitian ini untuk menghindari kompleksitas tambahan.

g. Fokus penelitian adalah membandingkan algoritma *support vector machine* dan *random forest* untuk klasifikasi. Algoritma lain seperti Naïve Bayes, KNN, atau Neural Network tidak digunakan dalam penelitian ini.

h. Untuk SVM, penelitian ini melakukan tuning parameter C, gamma, dan kernel (Linear, RBF, Polynomial). Sedangkan Untuk RF, tuning dilakukan pada jumlah pohon (n_estimators), kedalaman maksimum (max_depth), dan jumlah fitur (max_features). Grid Search atau Random Search digunakan sebagai teknik tuning, tetapi tidak semua kemungkinan kombinasi akan diuji.

i. Penilaian kinerja model dilakukan berdasarkan akurasi, precision, recall, F1-score, dan ROC-AUC Score. Evaluasi tidak mencakup analisis biaya implementasi atau dampak praktis dari penggunaan model ini di lingkungan pendidikan.

j. Dataset yang digunakan terbatas pada data historis dari tahun tertentu, yang mencakup santri yang telah diterima atau tidak diterima di universitas Timur Tengah. Ukuran dataset dapat mempengaruhi hasil analisis dan generalisasi

model, sehingga hasil penelitian hanya berlaku pada cakupan dataset yang digunakan.

k. Dalam dataset, terdapat kemungkinan bahwa beberapa santri tidak diterima di universitas Timur Tengah bukan karena ketidakmampuan akademik, tetapi karena memang tidak mendaftar atau memilih tidak melanjutkan studi ke luar negeri. Hal ini dapat menyebabkan bias dalam model prediksi. Namun, penelitian ini tidak dapat sepenuhnya memisahkan faktor ini karena keterbatasan data mengenai alasan santri tidak melanjutkan studi.

1.4. Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Menganalisis model prediksi kelulusan santri untuk melanjutkan studi ke Timur Tengah dengan menggunakan algoritma *Support Vector Machine* dan *Random Forest*.
2. Mengidentifikasi faktor-faktor yang berpengaruh terhadap kelulusan santri berdasarkan data nilai akademik, kemampuan Bahasa Arab, dan jumlah hafalan Al-Qur'an.
3. Membandingkan performa kedua algoritma berdasarkan hasil evaluasi kuantitatif untuk menentukan metode yang lebih efektif.

1.5. Manfaat Penelitian

Penelitian ini diharapkan memberikan manfaat sebagai berikut:

a. Bagi pengembangan ilmu pengetahuan dan teknologi, penelitian ini dapat memperluas penerapan algoritma *machine learning*, khususnya *Support Vector Machine* dan *Random Forest*, dalam bidang pendidikan Islam yang masih jarang diteliti.

b. Bagi lembaga pendidikan Islam, hasil penelitian ini diharapkan dapat menjadi dasar dalam merancang proses seleksi dan pembinaan santri berbasis data, sehingga pengambilan keputusan dapat dilakukan secara lebih objektif dan efisien.

c. Bagi Pondok Pesantren Imam Bukhari, penelitian ini dapat memberikan gambaran mengenai faktor-faktor akademik dan non-akademik yang berpotensi memengaruhi kelulusan santri untuk melanjutkan studi ke Timur Tengah, sehingga dapat dimanfaatkan dalam perencanaan program pembinaan yang lebih terarah.

d. Bagi pengembang sistem pendukung keputusan (*Decision Support System*), penelitian ini dapat menjadi acuan awal dalam penerapan model prediksi berbasis algoritma klasifikasi untuk membantu proses penilaian dan seleksi santri secara otomatis.

e. Bagi peneliti selanjutnya, penelitian ini dapat dijadikan referensi untuk mengembangkan model prediksi yang lebih kompleks, menggunakan data yang lebih luas, atau menerapkan metode *ensemble learning* lain untuk meningkatkan ketepatan prediksi pada bidang pendidikan Islam.

BAB 2

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Beberapa penelitian mengenai prediksi kelulusan siswa atau mahasiswa menggunakan algoritma machine learning telah banyak dilakukan. Studi-studi tersebut memberikan wawasan mengenai berbagai pendekatan dan algoritma yang digunakan, seperti Support Vector Machine (SVM), Random Forest, dan metode lainnya. Beberapa penelitian itu diantaranya Implementasi Data Mining Untuk Memprediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan *Random Forest* (Nurfadilla et al., 2022) menerapkan metode Random Forest untuk memprediksi kelulusan mahasiswa tepat waktu. Penelitian ini menyoroti kekuatan Random Forest dalam menangani data yang kompleks dan menghasilkan model prediksi yang andal. Penelitian ini menunjukkan bahwa Random Forest dapat menjadi alternatif untuk algoritma lainnya.

Kemudian Penerapan Algoritma *Support Vector Machine* (SVM) untuk Prediksi Tingkat Kelulusan Siswa SMA (Wulandari et al., 2024), pada penelitian ini algoritma SVM digunakan untuk memprediksi tingkat kelulusan siswa SMA. Studi ini berfokus pada keakuratan prediksi yang dicapai dengan menggunakan SVM dalam konteks pendidikan menengah. Hasilnya menunjukkan bahwa SVM dapat digunakan untuk memetakan kemungkinan kelulusan berdasarkan data akademik siswa, dimana nilai akurasi sebesar 98.81% untuk siswa kelas XII, 96.49% untuk siswa kelas XI dan 98.25% untuk siswa kelas X.

Kemudian pada penelitian Perbandingan Metode Klasifikasi *Random Forest* dan *Support Vector Machine* Dalam Memprediksi Capaian Studi Mahasiswa (Novianto et al., 2024), dimana pada penelitian ini membandingkan kinerja algoritma Random Forest dan SVM dalam memprediksi capaian studi mahasiswa. Penelitian ini menyimpulkan bahwa kedua algoritma memiliki kelebihan masing-masing, tergantung pada kompleksitas data yang digunakan, walaupun setelah dilakukan seleksi fitur, didapatkan hasil akurasi yang sama, yaitu 97,6%. Penelitian ini juga menekankan pentingnya seleksi fitur untuk meningkatkan akurasi prediksi. Kemudian penelitian dengan judul Pola Prediksi Kelulusan Siswa Madrasah Aliyah Swasta dengan *Support Vector Machine* dan *Random Forest* (Darmawan et al., 2023) mengeksplorasi pola prediksi kelulusan siswa Madrasah Aliyah menggunakan algoritma SVM dan Random Forest. Hasil penelitian ini menunjukkan bahwa kombinasi kedua algoritma dapat meningkatkan efisiensi prediksi dibandingkan hanya menggunakan satu algoritma. Dan dari hasil pengujian dapat diketahui bahwa algoritma RF sedikit lebih baik daripada algoritma SVM pada hasil pengukuran parameter *Accuracy* (99,49% vs 98,98%), *Precision*(99,74% vs 99,23%), *Recall*(99,74% vs 99,74%), *F-Measure* (99,74% vs 99,48%), dan *Classification error*(0,005 vs 0,010), akan tetapi untuk parameter waktu konsumsi SVM(0,04 detik) lebih baik daripada RF(0,26 detik).

Kemudian penelitian dengan judul Prediksi Kelulusan Siswa dengan metode *Support Vector Machine* (SVM) di SMK Adiluhur (Lukman dan Herlinda, 2024), menerapkan metode SVM untuk memprediksi kelulusan siswa di SMK Adiluhur. Penelitian ini memanfaatkan dataset dari sekolah untuk mengembangkan model

prediktif yang memfasilitasi analisis faktor-faktor yang memengaruhi kelulusan siswa. Hasil penelitian menunjukkan bahwa metode SVM menghasilkan *Precision* (97%), *recall* (82%) dan *accuracy* (95.06%) yang berarti bahwa data ini tergolong data *excellent*, sehingga dapat digunakan untuk memprediksi kelulusan siswa.

Kemudian pada Prediksi Kelulusan Siswa Sekolah Menengah Pertama Menggunakan *Machine Learning* (Naibaho dan Zahra, 2023), penelitian ini menggunakan pendekatan *machine learning* untuk memprediksi kelulusan siswa sekolah menengah pertama. Algoritma klasifikasi yang digunakan adalah *decision tree, random forest,* dan *extreme gradient boosting* dengan *gridsearchCV* dan *k-fold=5*. Studi ini membahas implementasi algoritma *machine learning* dalam pendidikan dasar dan menengah serta hasil prediksi yang dapat digunakan untuk mendukung pengambilan keputusan sekolah. Pada penelitian ini algoritma *random forest* mengungguli metode lainnya dengan nilai 99,5%.

Kemudian Prediksi Kelulusan Tepat Waktu Siswa SMK Teknik Komputer Menggunakan Algoritma *Random Forest* (Fatunnisa dan Marcos, 2024) mengembangkan model prediksi menggunakan Random Forest untuk siswa SMK Teknik Komputer. Penelitian ini menyoroti akurasi algoritma dalam memprediksi kelulusan tepat waktu berdasarkan data akademik dan non-akademik. Pada penelitian ini, penggunaan algoritma *random forest* dengan dataset kelulusan siswa. Pendistribusian pemilihan data latih dan uji menggunakan metode *stratified random sampling* untuk memastikan representasi yang seimbang dari setiap kelas yang dihasilkan. Model *Random Forest* berhasil diperoleh melalui pelatihan dan evaluasi model menggunakan data uji, menunjukkan akurasi 1,0 atau setara dengan 100%.

Kemudian Seleksi Fitur dan Perbandingan Algoritma Klasifikasi untuk Prediksi Kelulusan Mahasiswa (Zeniarja et al., 2022) mengevaluasi seleksi fitur dan perbandingan berbagai algoritma klasifikasi untuk memprediksi kelulusan mahasiswa, antara lain algoritma *Naïve Bayes, Random Forest, Decision Tree, K-Nearest Neighbor (K-NN)* dan *Support Vector Machine* (SVM). Penelitian ini menunjukkan bahwa *random forest* memiliki nilai akurasi paling tinggi mencapai 77.35% yang dapat dikatakan sebagai *fair classification*.

Kemudian Penerapan Data Mining untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu Menggunakan *Random Forest* (Oon Wira Yuda et al., 2022) mengkaji penerapan data mining dengan metode Random Forest untuk mengklasifikasikan kelulusan mahasiswa tepat waktu. Penelitian ini menunjukkan bahwa Random Forest mampu menghasilkan model klasifikasi yang akurat dengan memanfaatkan data historis akademik mahasiswa. Dengan menggunakan algoritma random forest dan variable yang digunakan, diperoleh tingkat akurasi sebesar 98%. Studi ini juga memberikan rekomendasi untuk memperluas penggunaan metode ini di berbagai institusi pendidikan lainnya.

Dari berbagai penelitian di atas, dapat disimpulkan bahwa algoritma SVM dan Random Forest adalah dua metode yang sering digunakan dalam memprediksi kelulusan siswa atau mahasiswa. Kedua algoritma ini memiliki keunggulan dalam menangani data kompleks dan memberikan hasil prediksi yang akurat. Penelitian lanjutan dapat mengintegrasikan metode ini untuk menghasilkan model yang lebih andal dan efisien.

2.2. Keaslian Penelitian

Tabel 2.1 Matriks literatur review dan posisi penelitian

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Benchmarking State-of-the-Art Classification Algorithms for Predictive Modeling	Stefan Lessmann et al., European Journal of Operational Research, Vol 247, 2015	Membandingkan 41 algoritma klasifikasi termasuk SVM dan RF untuk modeling prediksi dalam berbagai dataset.	RF dan SVM adalah dua algoritma yang secara konsisten menunjukkan performa tinggi dalam prediksi berbasis klasifikasi.	Perlu kajian lebih lanjut pada validasi cross-dataset untuk meningkatkan kemampuan generalisasi algoritma.	Studi ini membahas benchmarking algoritma secara luas, sementara penelitian yang dilakukan lebih berfokus pada studi kasus spesifik santri untuk studi ke Timur Tengah.
2.	Analysis of Machine Learning Strategies for Prediction of Passing Undergraduate Admission Test	Md. Abul Ala Walid et al., Int. Journal of Information Management Data Insights, 2022	Menganalisis strategi pembelajaran mesin seperti SVM dan RF untuk prediksi ujian masuk universitas.	Model SVM berbasis SMOTE menunjukkan kinerja terbaik pada dataset dengan distribusi kelas tidak seimbang.	Perlu evaluasi lebih lanjut pada variasi algoritma lain dan implementasi real-time.	Penelitian ini relevan dalam penggunaan SVM dan RF, tetapi pada penelitian yang dilakukan memiliki perbedaan, antara lain karena melibatkan nilai hafalan Al-Qur'an dan Bahasa Arab untuk prediksi santri dari pondok pesantren.

3	Implementasi Data Mining Untuk Memprediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Random Forest	Nurfadilla Z dan Faisal, 2022	Menggunakan Random Forest untuk prediksi kelulusan tepat waktu	Random Forest menunjukkan kekuatan dalam menangani data kompleks dengan nilai akurasi mencapai 90,74%	Tidak dibahas tentang seleksi fitur	Penelitian ini akan membandingkan SVM dan RF dan evaluasi dilakukan menggunakan metrik akurasi, precision, recall, dan F1-score untuk mengetahui metode yang lebih efektif
4	Seleksi Fitur dan Perbandingan Algoritma Klasifikasi untuk Prediksi Kelulusan Mahasiswa.	Zeniarja, Salam, dan Ma'ruf, 2022	Mengevaluasi seleksi fitur dan perbandingan algoritma, antara lain *Naïve Bayes, Random Forest, Decision Tree, K-Nearest Neighbor (*K-NN*) dan Support Vector Machine (*SVM*)* untuk memprediksi kelulusan mahasiswa.	Seleksi fitur meningkatkan performa model	Tidak dijelaskan aplikasi langsung di dunia nyata	Penelitian sebelumnya membandingkan beberapa algoritma secara umum untuk prediksi kelulusan mahasiswa. Namun, penelitian yang dilakukan lebih fokus pada algoritma SVM dan Random Forest, dengan penerapan langsung pada konteks pondok pesantren Imam Bukhari. Penelitian yang dilakukan juga menggunakan variabel spesifik seperti hafalan

						Al-Qur'an dan Bahasa Arab, yang relevan untuk studi di Timur Tengah.
5	Pola Prediksi Kelulusan Siswa Madrasah Aliyah Swasta dengan SVM dan Random Forest	Darmawan et al., 2023	Mengintegrasikan SVM dan Random Forest untuk meningkatkan efisiensi prediksi	Kombinasi metode menghasilkan hasil lebih baik, dimana SVM menghasilkan akurasi 98,98% dan RF 99,49%.	Belum diuji dengan menggunakan dataset besar, dan hanya berdasarkan dataset sedikit.	Penelitian sebelumnya menggunakan SVM dan RF pada siswa Madrasah Aliyah untuk prediksi kelulusan. Namun, pada penelitian yang dilakukan lebih spesifik dengan fokus pada santri pondok pesantren, variabel tambahan seperti hafalan Al-Qur'an dan Bahasa Arab, serta tujuan studi di universitas Timur Tengah. Selain itu, dataset yang digunakan juga lebih besar dan relevan untuk konteks pendidikan berbasis agama.

6	Penerapan Algoritma Support Vector Machine (SVM) untuk Prediksi Tingkat Kelulusan Siswa SMA	Wulandari, Aviani, dan Saputra, 2024	Penerapan SVM dalam memprediksi kelulusan siswa SMA	SVM memberikan hasil akurat dalam konteks Siswa SMA	Dataset yang digunakan kurang luas karena hanya fokus pada siswa SMA	Penelitian sebelumnya hanya menggunakan algoritma SVM pada siswa SMA dengan fokus pada nilai akademik standar. Sementara itu, penelitian yang dilakukan tidak hanya menggunakan SVM tetapi juga membandingkannya dengan RF, serta mencakup variabel unik seperti hafalan Al-Qur'an dan nilai Bahasa Arab. Selain itu, penelitian yang dilakukan pada siswa pondok pesantren dengan kompleksitas pelajaran yang lebih tinggi dan tujuan studi ke luar negeri.
7	Perbandingan Metode Klasifikasi Random Forest dan Support Vector	Novianto, Suhirman, dan Prasetyo, 2024	Membandingkan kinerja Random Forest dan SVM	Kedua algoritma memiliki kelebihan sesuai data.Dan hasil	Keterbatasan dalam variasi data yang diuji. Dan juga fitur yang digunakan	Akan dilakukan perbandingan algoritma SVM dan RF dengan metode

	Machine Dalam Memprediksi Capaian Studi Mahasiswa			akurasinya sama, yaitu 97,67%.	hanya Forward Selection (FS)	evaluasi yang digunakan adalah metrik akurasi, precision, recall, dan F1-score
8	Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review	Lidya R. Pelima et al., IEEE Access, Vol 12, 2024	Mengulas penggunaan ML, termasuk SVM dan RF, untuk memprediksi kelulusan mahasiswa berdasarkan data akademik.	SVM dan RF adalah algoritma paling umum digunakan dalam prediksi kelulusan dengan akurasi hingga 90%.	Perlu eksplorasi lebih dalam mengenai variabel lain di luar akademik untuk meningkatkan keakuratan prediksi.	Penelitian sebelumnya Fokus pada prediksi kelulusan atau penerimaan di institusi pendidikan umum (sekolah, perguruan tinggi) berdasarkan data akademik standar seperti nilai rata-rata, IPK, jumlah SKS, atau hasil ujian masuk. sedangkan penelitian yang dilakukan lebih spesifik ke pondok pesantren dengan fokus pada Bahasa Arab dan hafalan Al-Qur'an, memberikan dimensi baru pada aplikasi ML di pendidikan berbasis agama.

2.3. Data *Mining*

Data mining merupakan proses untuk menemukan hubungan dalam data yang tidak diketahui oleh pengguna dan menyajikannya dengan cara yang dapat dipahami, sehingga hubungan tersebut dapat menjadi dasar pengambilan keputusan (Jr. McLeod.R dan Schell, 2007). Data mining juga sebuah proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data dengan cara melakukan penggalian pola-pola dari data, dan tujuan untuk bisa melakukan manipulasi data menjadi informasi yang berharga dengan cara mengenali pola yang penting atau menarik dari data yang terdapat dalam basis data (Suyadi, 2017).

Salah satu metode data mining adalah dengan klasifikasi. Metode klasifikasi memiliki beberapa algoritma dan setiap algoritma klasifikasi pada data mining pastinya memiliki kelebihan dan kekurangan, sehingga penelitian yang dilakukan juga digunakan untuk menganalisis performansi dari algoritma tersebut. Penerapan kinerja algoritma data mining dapat dilakukan berdasarkan kriteria antara lain melalui keakuratan, kesempurnaan, konsistensi, ketepatan, dapat dipercaya dan interpretabilitas (Sari Dewi, 2016).

2.4. Klasifikasi

Klasifikasi merupakan kegiatan yang mengelompokkan suatu benda yang memiliki beberapa ciri yang sama dan memisahkannya dengan benda yang tidak sama. Pada dasarnya di dunia perpustakaan dikenal ada 2 (dua) jenis kegiatan klasifikasi yaitu Klasifikasi Fundamental (*Fundamental Classification*) dan Klasifikasi Artifisial (*Artificial Classification*). Klasifikasi Fundamental

(*Fundamental Classification*) adalah klasifikasi bahan pustaka yang diambil berdasarkan subyek/isi buku, karena pemakai perpustakaan lebih banyak mencari informasi tentang subyek tertentu. Klasifikasi Artifisial (*Artificial Classification*), yaitu klasifikasi bahan pustaka yang diambil berdasarkan ciri-ciri yang ada pada bahan pustaka (Suyadi, 2017). Dalam klasifikasi terdapat dua pekerjaan utama yang harus dilakukan, yaitu pembangunan model sebagai prototipe yang hasilnya disimpan sebagai memori dan penggunaan model untuk melakukan pengenalan/klasifikasi/prediksi pada suatu objek, agar data lain diketahui berasal dari kelas mana objek datanya didapatkan (Dwi L. Arisandy, Victor W., dan Yeni D. Rahayu, 2016).

2.5. Random Forest

Random Forest merupakan sebuah metode klasifikasi pengembangan dari *Decision Tree*. Sesuai dengan namanya, metode ini menciptakan sebuah hutan (*forest*) dengan sejumlah pohon (*tree*). Analogi yang bisa digunakan dalam penyebutan *Random Forest* karena semakin banyak pohon (*tree*) pada hutan (*forest*), maka akan semakin kuat hutan terlihat. Jika diimplementasikan pada sebuah kasus, apabila semakin banyak pohon (*tree*), maka akan semakin besar pula akurasi yang didapatkan (Reinardus A. Haristu, 2019).

Banyaknya pohon yang akan dibentuk sangat berpengaruh terhadap tingkat akurasi hasil klasifikasi. Semakin banyak pohon, semakin akurat hasil klasifikasinya. Sebaliknya, jika jumlah pohon terlalu banyak, maka akan menyebabkan error dalam analisis dataset. Pohon keputusan digunakan untuk

memetakan nilai informasi yang digunakan untuk tingkat ketidakmurnian dengan nilai entropi. Untuk menghitung nilai *entropy* digunakan Persamaan 2.1.

$$Entropy = -\sum_{i} p(c|Y)\log_2 p(c|Y) \quad (2.1)$$

Dimana Y adalah himpunan kasus dan p(c|Y)p(c|Y)p(c|Y) merupakan proporsi nilai Y terhadap kelas c.

$$InformationGain(Y,a) = Entropy(Y) - \sum_{v \in Values(a)} \frac{|v|}{|a|} Entropy(Y_v) \quad (2.2)$$

Dimana:

Values(a) adalah semua nilai yang mungkin dalam himpunan kasus aaa.

Yv adalah subkelas dari Y dengan kelas v yang berhubungan dengan kelas a.

|v| adalah semua nilai yang sesuai dengan v.

Untuk penerapan algoritma *Random Forest*, diperlukan algoritma untuk membangun *tree*. Salah satu algoritma yang digunakan adalah algoritma *Classification and Regression Tree* (CART). CART membagi pohon keputusan (*decision binary tree*) dan dapat diterapkan pada variabel numerik dan kategori sekaligus dibuktikan bahwa CART cocok diterapkan untuk data variabel yang banyak dan kompleks (Fransiska A. Kurniawan dan Angelina P. Kurniati, 2011).

2.6. Support Vector Machine

SVM merupakan salah satu dari metode yang dikembangkan untuk mengatasi permasalahan yang tidak bisa diselesaikan dengan metode statistika klasik, terutama pada kasus klasifikasi dan prediksi. SVM salah satu teknik yang relatif baru dibandingkan dengan teknik lain, tetapi memiliki performansi yang lebih baik di berbagai bidang aplikasi seperti *bioinformatics*, pengenalan tulisan tangan, klasifikasi teks, dan lain sebagainya (Bendi V. Ramana, Prof. M. Surendra P. Babu, Prof. N. B. Venkateswarlu, 2011).

Dalam SVM, untuk memisahkan data terhadap kelasnya, SVM membangun sebuah *hyperplane* (bidang pemisah). *Hyperplane* yang baik bukan hanya memisahkan data, tetapi memiliki batasan (*margin*) yang paling besar (Munawaroh dan Raudatul, 2016). SVM dikembangkan dengan prinsip *linear classifier*. Namun, dalam kasus nyata sering dijumpai data yang tidak linear sehingga dikembangkan SVM untuk kasus non-linear dengan memasukkan konsep kernel. Dengan begitu, ada jaminan bahwa klasifikasi menggunakan SVM akan menghasilkan pemetaan yang sangat akurat (Muhammad I. Fachrudin, 2015).

Proses pembelajaran SVM adalah untuk menentukan *support vector*, cukup mengetahui fungsi kernel yang dipakai tanpa perlu mengetahui wujud dari fungsi non-linear. Persamaan SVM ditunjukkan pada Persamaan 2.3.

$$f(x) = w^T \phi(x) + b \tag{2.3}$$

Dimana:

b = bias

$\mathbf{x}$ = variabel input

$\mathbf{w}$ = parameter bobot

$\boldsymbol{\phi}(\mathbf{x})$ = fungsi transformasi fitur

Melihat dari konsep metode *Support Vector Machine* ini, muncul pemikiran apakah metode ini dapat digunakan untuk memprediksi kelulusan santri dalam mendaftar di universitas di Timur Tengah.

2.7. Machine Learning Dalam Dunia Pendidikan

Machine learning adalah teknologi yang memungkinkan komputer untuk belajar dari data dan membuat prediksi atau keputusan tanpa diprogram secara eksplisit. Dalam dunia pendidikan, *machine learning* digunakan untuk berbagai tujuan, seperti memprediksi kelulusan, mengidentifikasi siswa yang memerlukan intervensi, dan mengoptimalkan proses pembelajaran. Algoritma seperti SVM dan Random Forest telah terbukti efektif dalam membantu institusi pendidikan memahami pola data akademik dan meningkatkan pengambilan keputusan.

2.8. Python

Python merupakan bahasa pemrograman dinamis yang mendukung pemrograman berbasis objek. Python didistribusikan dengan beberapa lisensi yang berbeda dari beberapa versi. Prinsipnya, Python diperoleh dan dipergunakan secara bebas (*freeware*), bahkan untuk kepentingan komersial. Lisensi Python tidak bertentangan, secara definisi *Open Source* maupun *General Public License* (GPL). Lengkap dengan *source code*, *debugger*, dan *profiler*, antarmuka yang terkandung

di dalamnya untuk antarmuka aplikasi, sistem file, sistem GUI (antar muka pengguna grafis), dan basis datanya (Therzian, R. Perkasa, Helmy, W., dan Pauladie S., 2014).

Hal yang membedakan Python dengan bahasa lain adalah dalam hal aturan penulisan kode program. Bahasa Python mendukung hampir di semua sistem operasi besar, baik itu sistem operasi Linux, hampir semua distribusi sudah menyertakan Python di dalamnya. Dengan kode yang simpel dan mudah diimplementasikan, seorang programmer dapat lebih mengutamakan pengembangan aplikasi yang dibuat (R.H. Sianipar dan Hamzan.W., 2015).

BAB 3

METODE PENELITIAN

3.1. Jenis, Sifat dan Pendekatan Penelitian

Pada penelitian ini menggunakan jenis penelitian eksperimen (*Experiment*) dengan menggunakan *dataset* dari sistem informasi pondok pesantren Imam Bukhari untuk mengevaluasi algoritma *Support Vector Machine* dan *Random Forest* dalam memprediksi kelulusan santri dalam mendaftar di universitas di Timur Tengah. Sifat penelitian ini adalah penelitian evaluasi yang bertujuan untuk mengevaluasi atau menilai efektivitas kinerja dari algoritma *Support Vector Machine* dan *Random Forest* dalam memprediksi kelulusan santri dalam mendaftar di universitas di Timur Tengah.

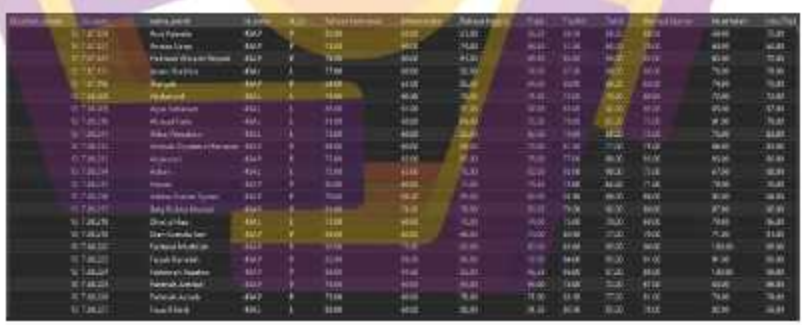
Tujuan dari penelitian ini adalah untuk membantu pondok pesantren dalam pengembangan sistem pendukung keputusan dalam proses seleksi santri yang berpotensi melanjutkan studi ke universitas di Timur Tengah menggunakan algoritma yang paling tepat dan akurat. Penelitian ini menggunakan metode kuantitatif dengan pendekatan *machine learning*, metode ini adalah salah satu jenis metode yang spesifikasinya sistematis, terencana dan terstruktur dengan jelas dari awal hingga pembuatan desain penelitiannya (Sugiyono, 2013).

3.2. Metode Pengumpulan Data

Dataset yang digunakan diperoleh melalui aplikasi sistem informasi akademik yang ada di pondok pesantren Imam Bukhari yang meliputi data nilai

akademik santri pada semester VI (semester akhir) di jenjang *Tsanawiyyah* (setingkat SMA) yang terdiri dari 21 mata pelajaran antara lain : Hafalan Al Qur'an, Tajwid, Tauhid, Hadits, Fiqih, Tafsir, Ulumul Qur'an, Musthalah, Usul Fiqih, Faraidh, Al-Adab Wa As-Suluk, Tarikh Islam, Tadris, Qawa'idul Lughah, Muthala'ah, Balaghah, Adab-Lughah, Ta'bir, Bahasa Indonesia, Bahasa Inggris dan Matematika. Disamping itu juga ditambahkan jumlah hafalan Al Qur'an dan nilai mata pelajaran Bahasa Arab dipergunakan sebagai dasar kemampuan berbahasa arab. Untuk hasil *outcome*, juga diperlukan status penerimaan (diterima/ditolak) di universitas di Timur Tengah.

Dataset ini merupakan data nilai akademik santri mulai tahun ajaran 2016/2017 sampai dengan tahun ajaran 2023/2024, yang diperoleh langsung dari aplikasi sistem informasi akademik yang sudah ada, adapun tampilan dari data yang diperoleh adalah seperti di tunjukkan pada Gambar 3.1.



Gambar 3.1 Tampilan data nilai akademik santri.

Untuk mendukung proses klasifikasi, data yang digunakan dalam penelitian ini terdiri dari sejumlah variabel *input* (fitur) dan satu variabel target sebagai label. Variabel-variabel tersebut merepresentasikan aspek akademik dan kemampuan

santri, seperti nilai pelajaran agama, kemampuan Bahasa Arab, serta jumlah hafalan Al-Qur'an. Berikut Tabel 3.1 menyajikan ringkasan variabel yang digunakan dalam penelitian.

Tabel 3.1 Variabel Dataset

No	Nama Variabel	Tipe Data	Deskripsi
1	Nilai Bahasa Arab	Numerik	Nilai akademik pada mata pelajaran Bahasa Arab
2	Hafalan Al-Qur'an	Numerik	Jumlah hafalan santri dalam satuan lembar (hasil konversi dari teks)
3	Nilai Aqidah	Numerik	Nilai akademik pada mata pelajaran Aqidah
4	Nilai Fiqih	Numerik	Nilai akademik pada mata pelajaran Fiqih
5	Nilai Hadits	Numerik	Nilai akademik pada mata pelajaran Hadits
6	Nilai Tafsir	Numerik	Nilai akademik pada mata pelajaran Tafsir
7	Nilai Nahwu	Numerik	Nilai akademik pada mata pelajaran Nahwu
8	Nilai Shorof	Numerik	Nilai akademik pada mata pelajaran Shorof
9	Nilai Bahasa Inggris	Numerik	Nilai akademik mata pelajaran Bahasa Inggris (opsional jika tersedia)
10	Nilai Matematika	Numerik	Nilai akademik mata pelajaran Matematika (opsional jika tersedia)

3.3. Metode Analisis Data

Dalam penelitian ini, peneliti menggunakan pendekatan *machine learning* dengan metode *Support Vector Machine* dan *Random Forest* untuk menganalisis data dan menghasilkan model prediksi diterimanya santri di universitas Timur Tengah. Adapun alat dan *software* yang digunakan untuk penelitian ini adalah menggunakan *platform google colabs*, dengan bahasa python, dan *library pandas*,

scikit-learn dan *NumPy*. Adapun metode analisis data yang digunakan dalam penelitian ini yaitu:

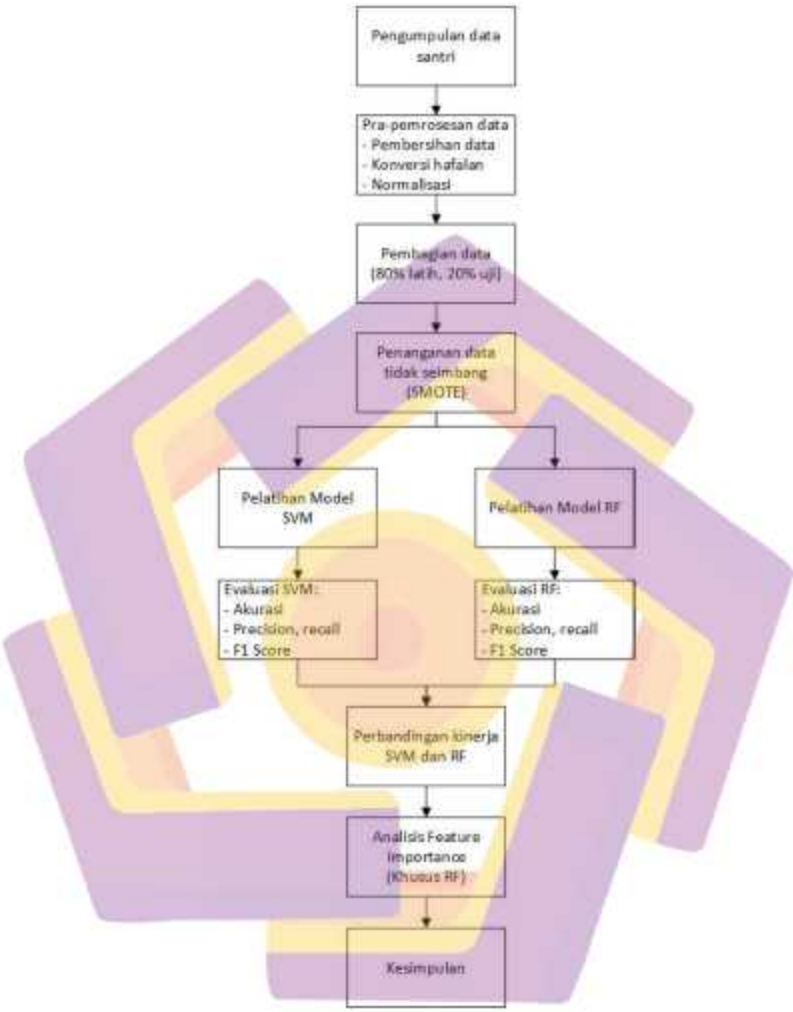
1. Analisis deskriptif untuk memberikan gambaran seberapa akurat model dalam memprediksi keberhasilan santri untuk diterima di universitas di Timur Tengah, dan faktor - faktor apa saja yang mempengaruhi peluang diterimanya santri.
2. Analisis untuk mengetahui variabel yang paling signifikan dengan menggunakan bobot dari model linear SVM sebagai indikasi pentingnya fitur pada algoritma SVM. Sedangkan pada *Random Forest* akan dievaluasi dengan *feature importance*, sehingga dapat diketahui seberapa signifikan setiap variabel terhadap hasil prediksi.

Adapun penerapan penelitian ini di pondok pesantren adalah untuk mengetahui variabel mana yang paling signifikan dalam mempengaruhi peluang kelulusan santri di universitas yang dituju. sehingga lembaga pendidikan dapat menyusun kriteria penerimaan dan pembinaan santri berdasarkan faktor paling signifikan.

Hasil prediksi yang dihasilkan oleh model mungkin menunjukkan *false negative* yang sebenarnya tidak relevan. Beberapa santri dengan prediksi "tidak lulus" mungkin memiliki nilai yang cukup baik tetapi tidak melanjutkan ke luar negeri karena alasan pribadi. Oleh karena itu, perlu dilakukan evaluasi tambahan untuk melihat apakah faktor ketidaklulusan lebih dipengaruhi oleh akademik atau faktor eksternal lainnya.

3.4. Alur Penelitian

Tahap ini menjelaskan secara sistematis tahapan-tahapan yang dilakukan dalam penelitian, mulai dari jenis dan pendekatan penelitian hingga evaluasi kinerja algoritma. Penelitian ini menggunakan pendekatan kuantitatif berbasis *machine learning* dengan tujuan membandingkan performa algoritma *Support Vector Machine* (SVM) dan *Random Forest* (RF) dalam memprediksi kelulusan santri di Pondok Pesantren Imam Bukhari. Alur penelitian ditunjukkan pada Gambar 3.1.



Gambar 3.2 Alur Penelitian

3.4.1 Pra-Pemrosesan Data

Pra-pemrosesan data merupakan tahap penting dalam penelitian berbasis *machine learning* untuk memastikan kualitas data yang digunakan dapat diproses

secara optimal oleh algoritma. Pada tahap ini, dilakukan beberapa proses utama seperti pembersihan data, transformasi format data, serta normalisasi nilai numerik. Langkah pertama adalah penghapusan kolom-kolom non-relevan, seperti ID santri, nama, dan jenis kelamin, yang tidak berkontribusi dalam proses klasifikasi. Kemudian dilakukan konversi pada kolom jumlah hafalan yang awalnya berbentuk teks seperti "2 Juz 10 Lembar" menjadi angka dalam format lembar, dengan asumsi satu juz setara dengan 20 lembar.

Selanjutnya, seluruh fitur yang mengandung nilai akademik diubah ke dalam format numerik agar dapat diolah oleh algoritma pembelajaran mesin. Setelah itu, data yang mengandung nilai kosong (*missing values*) dihapus agar tidak mengganggu proses pelatihan model. Setelah seluruh data numerik bersih dan lengkap, dilakukan proses normalisasi menggunakan metode *standard scaling* agar seluruh fitur berada dalam skala yang seragam.

Salah satu tahapan penting dalam *preprocessing* adalah normalisasi data, khususnya ketika algoritma yang digunakan sensitif terhadap skala fitur, seperti *Support Vector Machine* (SVM). Pada penelitian ini digunakan metode *Min-Max Scaling*, yaitu teknik normalisasi yang mengubah rentang data ke skala [0, 1] (Ambarwari, Adrian, & Herdiyeni, 2021). Normalisasi dilakukan terhadap seluruh fitur numerik seperti nilai akademik dan jumlah hafalan agar setiap fitur memiliki kontribusi yang seimbang dalam proses pelatihan model. Persamaan normalisasi data dengan metode *Min-Max Scaling* ditunjukkan pada Persamaan 3.1.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.1}$$

Keterangan:

Xnorm: Nilai variabel ternormalisasi

Xmax: Nilai variabel maksimal

Xmin: Nilai variabel minimal

X: Nilai variabel awal

Tahapan pra-pemrosesan ini diselesaikan sebelum proses pembagian data dan pelatihan model dilakukan. Proses ini dilaksanakan menggunakan bahasa pemrograman Python dengan bantuan pustaka *Pandas* dan *Scikit-Learn*.

3.4.2 Pembagian Data Latih dan Uji

Setelah melalui tahap pra-pemrosesan, dataset dibagi menjadi dua bagian, yaitu data latih dan data uji. Pembagian ini bertujuan untuk memisahkan data yang digunakan dalam proses pelatihan model dengan data yang digunakan untuk menguji performa model secara objektif. Dalam penelitian ini, data dibagi dengan rasio 80% sebagai data latih dan 20% sebagai data uji, menggunakan metode *stratified split*. Metode ini digunakan untuk memastikan bahwa distribusi kelas pada data uji tetap proporsional dan mewakili kondisi awal dari *dataset*, terutama dalam situasi *class imbalance*.

Data latih digunakan untuk membangun dan melatih model klasifikasi menggunakan algoritma *Support Vector Machine* (SVM) dan *Random Forest* (RF), termasuk proses *tuning* parameter. Sementara itu, data uji digunakan sebagai data yang tidak terlihat oleh model sebelumnya, sehingga hasil evaluasi dapat memperlihatkan performa model yang sesungguhnya. Proses pembagian data ini

dilakukan menggunakan fungsi *train_test_split* dari pustaka *Scikit-Learn*, dengan parameter *stratify* diaktifkan berdasarkan label kelas, serta pengaturan *random state* agar hasilnya konsisten.

3.4.3 Penanganan Ketidakseimbangan Data

Salah satu permasalahan umum dalam klasifikasi adalah ketidakseimbangan jumlah data antar kelas (*class imbalance*), jumlah data pada kelas mayoritas jauh lebih banyak dibandingkan kelas minoritas. Dalam penelitian ini, kelas mayoritas terdiri dari santri yang tidak melanjutkan studi ke Timur Tengah, sedangkan kelas minoritas adalah santri yang melanjutkan studi. Ketidakseimbangan ini dapat menyebabkan model pembelajaran mesin bias terhadap kelas mayoritas dan mengabaikan kelas minoritas, sehingga menurunkan kemampuan model dalam mengenali santri yang sebenarnya berpotensi diterima.

Untuk mengatasi hal tersebut, digunakan metode *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE bekerja dengan membuat sampel sintetis dari kelas minoritas berdasarkan jarak antar tetangga terdekat dalam ruang fitur. Teknik ini dilakukan hanya pada data latih, sehingga tidak memengaruhi distribusi data uji yang tetap mencerminkan kondisi asli. Proses penerapan SMOTE dilakukan dengan menggunakan pustaka *imbalanced-learn* pada Python. Setelah SMOTE diterapkan, distribusi antar kelas menjadi seimbang sehingga dapt membantu model dalam belajar secara adil terhadap kedua kelas.

3.4.4 Perancangan dan Pelatihan Model

Setelah data dibersihkan, dibagi, dan diseimbangkan, langkah selanjutnya adalah membangun dan melatih model klasifikasi. Penelitian ini menggunakan dua algoritma yaitu *Support Vector Machine* (SVM) dan *Random Forest* (RF). Masing-masing algoritma dirancang, dilatih, dan diuji dengan konfigurasi parameter tertentu menggunakan data latih yang telah diproses sebelumnya.

1. *Support Vector Machine* (SVM)

Support Vector Machine merupakan algoritma klasifikasi yang bekerja dengan mencari *hyperplane* optimal untuk memisahkan kelas dalam ruang fitur berdimensi tinggi. Dalam penelitian ini, SVM digunakan dengan kernel *Radial Basis Function* (RBF) karena kemampuannya dalam menangani data non-linear. Beberapa parameter penting pada SVM yang digunakan adalah:

a. C: Parameter regulasi yang mengontrol *trade-off* antara margin maksimal dan kesalahan klasifikasi.
b. Gamma: Parameter kernel yang menentukan seberapa jauh pengaruh satu titik data terhadap lainnya.

Model SVM dibangun menggunakan pustaka *Scikit-Learn*, dengan proses pelatihan dilakukan pada data latih hasil SMOTE. Setelah pelatihan selesai, model diuji menggunakan data uji untuk mengukur akurasi, *precision, recall*, dan *F1-score*.

2. *Random Forest*

Random Forest adalah algoritma berbasis *ensemble learning* yang membangun banyak pohon keputusan (*decision trees*) secara acak, kemudian

menggabungkan hasil prediksi dari seluruh pohon untuk menghasilkan keputusan akhir. Metode ini dikenal tangguh terhadap *overfitting* dan sangat efektif dalam klasifikasi data yang kompleks. Parameter penting yang digunakan dalam Random Forest antara lain:

a. *n_estimators*: Jumlah pohon yang dibentuk dalam satu hutan.
b. *max_depth*: Kedalaman maksimum setiap pohon (opsional, bisa dibiarkan default).
c. *random_state*: Untuk memastikan hasil eksperimen dapat direproduksi.

Model RF juga dibangun menggunakan pustaka *Scikit-Learn* dan dilatih menggunakan data latih hasil SMOTE. Evaluasi dilakukan dengan metrik yang sama seperti SVM, dan hasilnya dibandingkan untuk melihat performa relatif masing-masing algoritma.

Sebelum model dilatih, dilakukan tahap seleksi fitur (*feature selection*) untuk memastikan bahwa hanya fitur yang relevan dan informatif yang digunakan dalam proses klasifikasi. Seleksi fitur ini dilakukan secara konseptual dan eksploratif berdasarkan dua pendekatan:

a. Analisis korelasi antar variabel, dengan memeriksa hubungan antara nilai mata pelajaran dan variabel target untuk menghindari fitur yang memiliki korelasi sangat tinggi (multikolinearitas). Fitur dengan korelasi kuat antar sesamanya dipertimbangkan untuk dieliminasi agar tidak menimbulkan bias pada model.
b. Pertimbangan relevansi akademik, yaitu pemilihan fitur yang secara logis berkaitan langsung dengan keberhasilan santri dalam melanjutkan studi ke

Timur Tengah, seperti nilai pelajaran agama, kemampuan bahasa Arab, dan jumlah hafalan Al-Qur'an.

Melalui proses ini, model difokuskan hanya pada fitur-fitur yang berpengaruh secara signifikan terhadap prediksi kelulusan santri, sehingga hasil analisis *feature importance* pada tahap berikutnya menjadi lebih representatif dan stabil.

3.4.5 Evaluasi Model

Evaluasi model merupakan tahap penting untuk menilai kinerja algoritma dalam menyelesaikan tugas klasifikasi. Dalam penelitian ini, model dievaluasi menggunakan data uji yang telah dipisahkan sebelumnya dan tidak mengalami proses *oversampling*, sehingga mencerminkan performa model dalam kondisi data yang sebenarnya. Terdapat empat metrik utama yang digunakan dalam evaluasi, yaitu:

1. Akurasi (*Accuracy*)

   Akurasi menunjukkan proporsi prediksi yang benar terhadap keseluruhan data uji. Metrik ini mengukur sejauh mana model mampu mengklasifikasikan data dengan tepat secara umum.

2. *Precision*

   Precision mengukur ketepatan prediksi positif. Dalam konteks ini, precision menunjukkan seberapa banyak santri yang diprediksi "lulus ke Timur Tengah" yang benar-benar lulus.

3. *Recall* (*Sensitivity*)

Recall menunjukkan kemampuan model dalam menangkap seluruh kasus positif. Hal ini menandakan bahwa *recall* mengukur seberapa banyak santri yang benar-benar lulus ke Timur Tengah berhasil dikenali oleh model.

4. *F1-score*

   F1-score merupakan rata-rata harmonis antara *precision* dan *recall*. Metrik ini sangat berguna ketika terjadi ketidakseimbangan kelas, karena mempertimbangkan baik kesalahan positif maupun negatif.

Selain metrik di atas, digunakan juga *confusion matrix* untuk melihat rincian prediksi benar dan salah pada masing-masing kelas. *Confusion matrix* divisualisasikan dalam bentuk diagram untuk mempermudah interpretasi. Seluruh proses evaluasi dilakukan menggunakan pustaka *Scikit-Learn* dalam bahasa pemrograman Python, dengan model yang telah dilatih sebelumnya menggunakan data latih hasil SMOTE. Hasil evaluasi dari algoritma SVM dan Random Forest kemudian dibandingkan untuk menentukan model dengan performa terbaik.

3.4.6 Perbandingan Kinerja Model

Setelah proses pelatihan dilakukan untuk masing-masing algoritma, yakni *Support Vector Machine* (SVM) dan *Random Forest* (RF), langkah selanjutnya adalah melakukan evaluasi dan perbandingan performa kedua model tersebut. Evaluasi dilakukan berdasarkan beberapa metrik utama seperti akurasi, *precision*, *recall*, dan *F1-score*, dengan fokus khusus pada kemampuan model dalam mengenali santri yang lulus maupun tidak lulus. Hasil perbandingan ini akan

menjadi dasar dalam menentukan model yang paling tepat digunakan dalam sistem pendukung keputusan seleksi santri.

3.4.7 Analisis *Feature Importance*

Tahap ini menjelaskan perbandingan performa dua algoritma *machine learning*, yaitu *Support Vector Machine* (SVM) dan *Random Forest* (RF), dalam melakukan klasifikasi kelulusan santri. Setelah kedua model dilatih menggunakan data yang telah diproses dan diseimbangkan dengan metode SMOTE, evaluasi kinerja dilakukan berdasarkan metrik akurasi, *precision*, *recall*, dan *F1-score*. Pemilihan metrik ini bertujuan untuk memperoleh gambaran menyeluruh terkait kemampuan masing-masing model dalam mengenali santri yang berpotensi lulus maupun tidak lulus.

Perbandingan ini menjadi dasar untuk menentukan algoritma yang lebih unggul dan sesuai digunakan sebagai bagian dari sistem pendukung keputusan dalam proses seleksi calon mahasiswa untuk studi ke Timur Tengah. Selain evaluasi numerik, visualisasi *confusion matrix* juga digunakan untuk memperkuat hasil klasifikasi masing-masing model.

BAB 4

HASIL DAN PEMBAHASAN

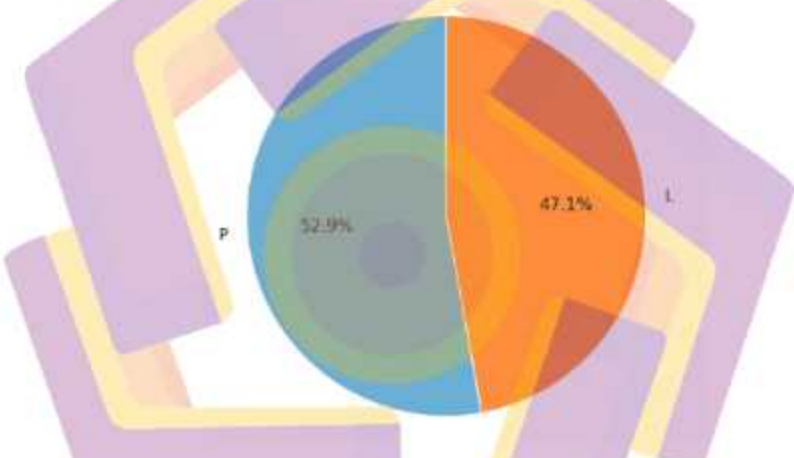
4.1 Hasil

4.1.1 Pengumpulan Data

Data penelitian ini diperoleh dari basis data akademik Pondok Pesantren Imam Bukhari yang dikelola oleh bagian Tata Usaha dan Bidang Akademik. Data tersebut dikumpulkan melalui proses ekspor dari sistem informasi akademik internal pesantren ke dalam format Microsoft Excel (.xlsx). Proses pengambilan dilakukan dengan izin resmi dari pimpinan pondok, serta melibatkan staf akademik untuk memastikan kelengkapan dan validitas data. Data penelitian ini mencakup nilai akademik, kemampuan bahasa Arab, dan jumlah hafalan Al-Qur'an para santri yang telah menyelesaikan masa belajarnya. *Dataset* berjumlah 2.333 baris dan memuat 27 kolom yang terdiri dari variabel identitas (ID tahun ajaran, ID santri, nama, ID kelas, jenis kelamin), nilai mata pelajaran umum (Bahasa Indonesia, Matematika, Bahasa Inggris), nilai mata pelajaran agama (Fiqih, Tauhid, Tafsir, Ulumul Qur'an, Musthalah, Usul Fiqih, Faraidh, Al-Adab Wa As-Suluk, Tarikh Islam, Tadris, Qawa'idul Lughah, Muthala'ah, Balaghah, Adab-Lughah, Ta'bir, Hadits, Tajwid), dan variabel hafalan Al-Qur'an.

4.1.2 Analisis Deskriptif

Analisis deskriptif dilakukan untuk memahami karakteristik umum dari data yang digunakan dalam penelitian ini. *Dataset* berisi informasi mengenai santri tingkat akhir Pondok Pesantren Imam Bukhari yang menjadi calon untuk melanjutkan studi ke perguruan tinggi di Timur Tengah. Total terdapat 2.333 data santri, yang terdiri atas 1.233 santri perempuan (52,9%) dan 1.100 santri laki-laki (47,1%). Distribusi santri berdasarkan jenis kelamin ditunjukkan pada Gambar 4.1.

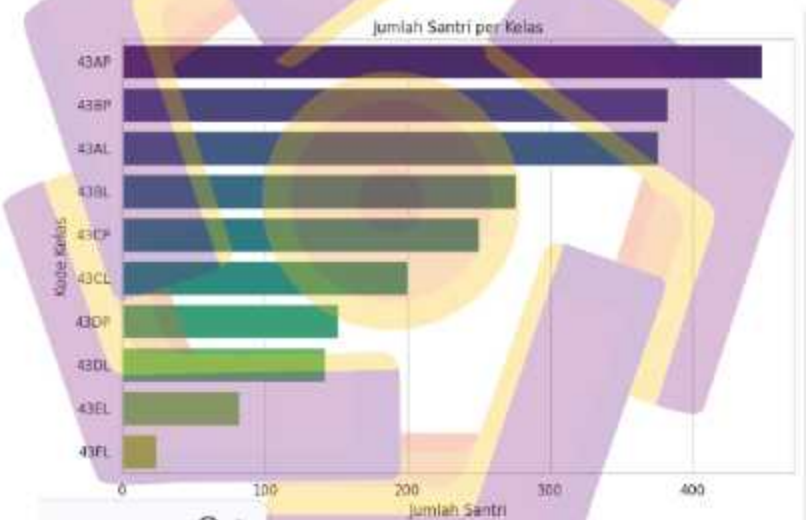


Gambar 4.1 Distribusi Santri Berdasarkan Jenis Kelamin

Santri yang menjadi objek dalam penelitian ini berasal dari berbagai kelas akhir tingkat atas di Pondok Pesantren Imam Bukhari. Terdapat 10 kelas berbeda, yang secara umum terdiri dari kombinasi huruf identifikasi SMA (4), penanda kelas (3A sampai F), serta jenis kelamin (L untuk laki-laki, P untuk perempuan). Kode seperti 43AP dapat diartikan sebagai kelas 3A perempuan dari SMA, dan seterusnya. Distribusi jumlah santri per kelas disajikan pada Tabel 4.1 dan Gambar 4.2.

Tabel 4.1 Tabel Distribusi Santri Per Kelas

No	Kode Kelas	Jumlah Santri
1	43AP	449
2	43BP	383
3	43AL	376
4	43BL	276
5	43CP	250
6	43CL	200
7	43DP	151
8	43DL	142
9	43EL	82
10	43FL	24
	Total	2.333



Gambar 4.2 Grafik Distribusi Santri Berdasarkan Kelas

Berdasarkan distribusi ini, dapat disimpulkan bahwa:

a. Kelas dengan jumlah santri terbanyak adalah kelas 43AP sebanyak 449 santri, diikuti oleh kelas 43BP dan 43AL.

b. Kelas dengan jumlah santri paling sedikit adalah 43FL, yaitu hanya 24 santri. Ini menunjukkan bahwa kelas 43FL merupakan kelas khusus atau

memiliki jumlah siswa terbatas karena alasan tertentu (misalnya penjurusan atau program khusus).

c. Secara umum, kelas perempuan (kode P) memiliki jumlah santri yang lebih banyak dibandingkan kelas laki-laki (kode L). Misalnya, gabungan kelas 43AP, 43BP, 43CP, dan 43DP mencakup total 1.233 santri perempuan, sedangkan total santri laki-laki dari kelas 43AL, 43BL, 43CL, 43DL, 43EL, dan 43FL berjumlah 1.100 santri.

Distribusi ini dapat berpengaruh terhadap pola pembelajaran, pendekatan pengajaran, serta kecenderungan hasil kelulusan yang akan dipelajari lebih lanjut dalam model prediksi. Ketimpangan jumlah siswa per kelas juga dapat menjadi indikasi awal apakah terdapat bias struktural dalam penugasan kelas atau seleksi internal sebelum kelulusan.

4.1.3 Pra-Pemrosesan Data

Pra-pemrosesan data merupakan tahapan penting dalam siklus *machine learning* yang bertujuan untuk memastikan bahwa data yang digunakan dalam pelatihan model berada dalam kondisi yang bersih, terstruktur, dan siap diproses. Tahapan ini mencakup beberapa proses berikut:

1. Pemeriksaan dan Pembersihan Data

Langkah pertama yang dilakukan adalah memeriksa kelengkapan data pada setiap kolom, khususnya untuk atribut yang bersifat numerik seperti nilai akademik, jumlah hafalan, serta atribut target berupa status kelulusan. Pemeriksaan ini juga digunakan untuk mendeteksi adanya kolom yang tidak relevan dengan analisis atau

memiliki format data yang tidak sesuai (misalnya angka yang tersimpan sebagai teks). Ditemukan beberapa entri dengan nilai kosong (*missing values*) pada sebagian variabel. Data tersebut dibersihkan dengan pendekatan:

a. Baris dengan nilai kosong pada variabel kunci seperti status kelulusan (label) atau total hafalan dihapus. Proses ini menggunakan fungsi isnull().sum() dari pustaka Pandas sehingga dapat diketahui kolom mana saja yang memiliki data hilang dan seberapa banyak jumlahnya. Pada variabel numerik seperti nilai mata pelajaran, ditemukan format angka yang menggunakan tanda koma (,) sebagai pemisah desimal. Format ini harus diubah menjadi titik (.) agar dapat dikenali sebagai tipe *float* oleh Python. Selain itu, kolom seperti *Jumlah Hafalan* yang awalnya menggunakan format teks deskriptif (misalnya "5 Juz 3 Lembar") diubah menjadi nilai numerik agar dapat diproses lebih lanjut.
b. Nilai non-numerik atau *outlier* yang ekstrem dicek secara manual untuk memastikan kesesuaian. Nilai yang tidak sesuai rentang logis (misalnya nilai ujian di bawah 0 atau di atas 100) ditandai dan dikoreksi atau dihapus sesuai kebijakan pengolahan data.

Setelah tahap ini, data yang tersisa memiliki format yang konsisten dan dapat diolah. Hasil pemeriksaan dan pembersihan data ditunjukkan pada Gambar 4.3.

```
Jumlah data awal: 2333

Jumlah missing values per kolom:
id_tahun_ajaran          0
id_santri                0
nama_santri              0
id_kelas                 0
lk_pr                    0
Bahasa Indonesia         3
Matematika               2
Bahasa Inggris           3
Fiqih                    3
Tauhid                 449
Tafsir                   3
Ulumul Qur'an            3
Musthalah                3
Usul Fiqih               3
Faraidh                  2
Al-Adab Wa As-Suluk    429
Tarikh Islam             2
Tadris                   2
Qawa'idul Lughah         3
Muthala'ah             122
Balaghah               122
Adab-Lughah            121
Ta'bir                 122
Hadits                   3
Tajwid                 366
Hafalan Qur'an           3
Jumlah Hafalan           0
kelulusan                0
dtype: int64
```

Gambar 4.3 Hasil Pemeriksaan dan Pembersihan Data

Hasil pemeriksaan menunjukkan bahwa tidak ditemukan data duplikat di dalam *dataset*, yang berarti setiap entri data mewakili individu santri yang unik. Namun, ditemukan sejumlah nilai kosong (*missing values*) pada beberapa kolom, terutama pada mata pelajaran tertentu. Beberapa mata pelajaran memiliki jumlah nilai kosong yang relatif kecil, seperti Bahasa Indonesia, Matematika, dan Bahasa Inggris, masing-masing hanya kehilangan 2–3 data. Namun, terdapat juga beberapa kolom dengan jumlah nilai kosong yang cukup signifikan. Misalnya kolom Tauhid memiliki 449 data kosong, kolom Al-Adab Wa As-Suluk memiliki 429 data kosong, dan kolom Tajwid memiliki 366 data kosong. Sementara mata pelajaran

seperti Muthala'ah, Balaghah, dan Ta'bir masing-masing memiliki lebih dari 120 nilai kosong.

Banyaknya *missing values* pada kolom-kolom tersebut menandakan bahwa tidak semua santri mengambil mata pelajaran yang sama, hal ini disebabkan oleh perbedaan kurikulum atau kelas. Oleh karena itu, untuk menjaga kualitas *dataset* yang akan digunakan dalam pelatihan model *machine learning*, hanya data yang memiliki nilai lengkap pada seluruh fitur akademik yang relevan dan nilai hafalan yang dipertahankan. Hal ini bertujuan agar proses pelatihan model tidak terganggu oleh data yang tidak lengkap. Setelah proses pembersihan, jumlah total data yang siap digunakan adalah 1.168 baris, yaitu data yang memenuhi syarat lengkap dan valid dari segi akademik maupun hafalan. Data ini kemudian menjadi dasar untuk seluruh tahapan berikutnya seperti transformasi, normalisasi, dan pelatihan model prediksi.

2. *Encoding* Variabel Kategorikal

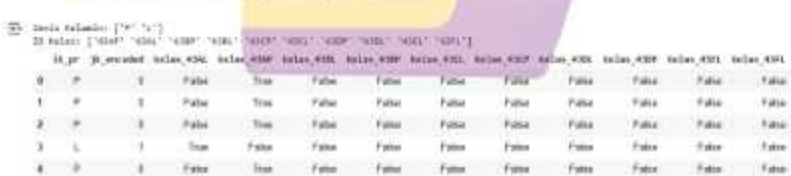
Proses encoding variabel kategorikal dilakukan untuk mengubah data non-numerik menjadi representasi numerik yang dapat diproses oleh algoritma machine learning. Tahap ini dilakukan karena sebagian besar algoritma, termasuk Support Vector Machine (SVM) dan Random Forest (RF), hanya dapat mengolah data numerik. Variabel kategorikal seperti jenis kelamin (jk) dan kelas (id_kelas) tidak dapat langsung digunakan dalam algoritma *machine learning* karena model hanya menerima input numerik. Oleh karena itu, dilakukan proses *encoding*.

a. Variabel jk dikonversi menjadi jk_encoded menggunakan metode *label encoding*, di mana nilai 'P' (Perempuan) dikodekan sebagai 0 dan 'L' (Laki-

laki) sebagai 1. Teknik ini dipilih karena sederhana, efisien, dan tidak menambah dimensi data

b. Variabel id_kelas dikodekan menggunakan *One-Hot Encoding* agar setiap kelas diubah menjadi kolom biner terpisah. Misalnya, jika seorang santri berada di kelas 43AP, maka kolom kelas_43AP akan bernilai 1, sementara kolom kelas lainnya 0. *One-Hot Encoding* dipilih karena variabel kelas tidak memiliki urutan tertentu (*non-ordinal*), sehingga metode ini mencegah algoritma menganggap ada hubungan numerik antar kelas.

Setelah kedua variabel dikodekan, dilakukan seleksi fitur untuk menentukan variabel mana yang digunakan pada tahap pelatihan model. Hanya jk_encoded yang digunakan karena berdasarkan pertimbangan awal, variabel kelas (id_kelas) tidak menunjukkan hubungan signifikan dengan variabel target dan berpotensi menimbulkan masalah multikolinearitas atau *overfitting*. Namun, dalam tahap pemodelan, hanya variabel jk_encoded yang digunakan karena variabel kelas (id_kelas) dianggap tidak memiliki pengaruh signifikan terhadap hasil kelulusan, atau bahkan dapat menimbulkan multikolinearitas dan *overfitting*. Hasil *Encoding* Variabel Kategorikal ditampilkan pada Gambar 4.4.



Gambar 4.4 Hasil *Encoding* Variabel Kategorikal

Teknik ini mengubah setiap kategori kelas menjadi kolom *biner* terpisah. Misalnya, kolom kelas_43AP bernilai *True* jika santri berada di kelas 43AP dan

False jika tidak. Dengan demikian, dari 10 kategori kelas, dihasilkan 10 kolom tambahan: kelas_43AL, kelas_43AP, kelas_43BL, kelas_43BP, kelas_43CL, kelas_43CP, kelas_43DL, kelas_43DP, kelas_43EL, dan kelas_43FL.

3. Normalisasi Data

Sebagian besar model pembelajaran mesin, khususnya *Support Vector Machine* (SVM) sensitif terhadap skala data. Oleh karena itu perlu dilakukan normalisasi terhadap seluruh variabel numerik menggunakan *Min-Max Scaling* untuk membawa nilai semua fitur ke rentang [0, 1]. Fitur yang dinormalisasi antara lain:

a. Nilai Bahasa Arab
b. Nilai Fiqih, Hadits, Tafsir, Aqidah
c. Nilai Nahwu dan Shorof
d. Nilai Bahasa Inggris
e. Jumlah Hafalan Al-Qur'an
f. Kelulusan

Sebelum melakukan normalisasi, dilakukan konversi nilai dari format *string* ke numerik. Ini mencakup penggantian pemisah desimal koma (,) menjadi titik (.), serta konversi tipe data menjadi *float*. Setelah memastikan tidak ada nilai teks atau kosong yang tersisa, maka proses normalisasi dengan *MinMaxScaler* berhasil dilakukan tanpa *error*. Hasil normalisasi data ditunjukkan pada Gambar 4.5.

	Bahasa Indonesia	Matematika	Bahasa Inggris	Fiqih	Tauhid	Tafsir	Ulumul Qur'an	Musthalah	Ushul Fiqih	Furuudh	...	Tarikh Islam	Tadris	Qawa'idul Lughah	Mutholo'ah	K
825	0.000000	0.00	0.000000	0.0000	0.00000	0.00	0.0000	0.000	0.0	0.000	...	0.000	0.000	0.000	0.0000	
826	0.000000	0.00	0.000000	0.0000	0.00000	0.00	0.0000	0.000	0.0	0.000	...	0.000	0.000	0.000	0.0000	
827	0.000000	0.00	0.000000	0.0000	0.00000	0.00	0.0000	0.000	0.0	0.000	...	0.000	0.000	0.000	0.0000	
828	0.000000	0.00	0.000000	0.0000	0.00000	0.00	0.0000	0.000	0.0	0.000	...	0.000	0.000	0.000	0.0000	
829	0.000000	0.00	0.000000	0.0000	0.00000	0.00	0.0000	0.000	0.0	0.000	...	0.000	0.000	0.000	0.0000	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
2326	0.893939	0.79	0.760726	0.6250	0.00000	0.68	0.6250	0.785	0.0	0.855	...	0.765	0.805	0.880	0.6625	
2327	0.906566	0.67	0.821584	0.9525	0.00000	0.52	0.9475	0.905	0.0	0.820	...	0.825	0.930	0.965	0.9200	
2328	0.830808	0.66	0.765790	0.7405	0.00000	0.71	0.7575	0.825	0.0	0.840	...	0.625	0.765	0.830	0.7325	
2329	0.631818	0.67	0.611441	0.9400	0.94898	0.67	0.9773	0.920	0.0	0.950	...	0.980	0.960	0.940	0.8700	
2330	0.977273	0.91	0.923886	0.9000	0.00000	0.98	0.9650	0.990	0.0	0.990	...	0.915	1.000	0.975	0.9650	

1168 rows × 21 columns

Gambar 4.5 Hasil Normalisasi Data

Proses normalisasi berhasil diterapkan terhadap 21 variabel numerik, yang terdiri dari nilai mata pelajaran akademik dan hafalan. Hasil transformasi memperlihatkan rentang nilai antara 0 hingga 1, di mana nilai 0 merepresentasikan santri dengan nilai terendah dalam suatu fitur, sedangkan nilai 1 merepresentasikan santri dengan nilai tertinggi. Proses ini memungkinkan algoritma pembelajaran mesin terutama SVM yang sensitif terhadap skala dapat memproses fitur secara setara dan efisien. Contoh perhitungan:

Misalkan untuk menormalisasi nilai Bahasa Indonesia dari seorang santri di baris ke-2326.

a. Nilai asli (X): 92.25
b. Nilai minimum (Xmin): 52
c. Nilai maksimum (Xmax): 97

Untuk menghitungnya menggunakan Persamaan 3.1 berikut:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Subtitusi nilai menjadi:

$$X_{norm} = \frac{92.25 - 52}{97 - 52} = \frac{40.25}{45} = 0.894$$

Dari total 2.333 baris data santri yang tersedia, proses pra-pemrosesan dilakukan untuk memastikan kelengkapan dan kualitas data, khususnya pada dua aspek utama: kelengkapan nilai akademik dan kevalidan informasi hafalan.

a. Jumlah baris total: 2.333

   Ini merupakan keseluruhan entri santri sebelum dilakukan pembersihan data.

b. Jumlah baris setelah bersih: 2.268

   Ini adalah data akhir yang digunakan untuk pelatihan dan pengujian model. Data ini telah melalui tahapan konversi nilai hafalan menjadi numerik (jumlah lembar), penghapusan kolom non-penting seperti nama santri dan ID, dan *drop* semua baris yang mengandung nilai kosong (NaN) agar model tidak *error* saat proses pelatihan.

c. Jumlah kelulusan

   Siswa yang dinyatakan lulus (label 1) sebanyak 207 data dan siswa yang dinyatakan tidak lulus (label 0) sebanyak 2061 data. Hasil validasi data ditunjukkan pada Gambar 4.6.

```
Jumlah baris setelah bersih: 2268
Distribusi label kelulusan:
 kelulusan
0    2061
1     207
Name: count, dtype: int64
              nama_santri  kelulusan  jk_encoded  jumlah_hafalan_numeric
0            Ami Valentin          0           0                     0.0
1             Annisa Uana          0           0                     0.0
2  Halimah Winanti Royani          0           0                     0.0
```

Gambar 4.6 Hasil Validasi Data

4.1.4 *Split* Data

Data penelitian ini telah dibagi menjadi dua bagian utama, yakni data latih dan data uji, dengan proporsi 80% untuk pelatihan model dan 20% untuk pengujian model. Proses pembagian ini dilakukan menggunakan metode *stratified split* agar distribusi label antara kelas "lulus" dan "tidak lulus" tetap proporsional di kedua bagian data. Pembagian dilakukan menggunakan fungsi *train_test_split* dari pustaka *scikit-learn* dengan parameter *stratify* yang diisi berdasarkan kolom label. Parameter *random_state* juga ditetapkan untuk memastikan proses pembagian dapat direplikasi dengan hasil yang sama pada setiap percobaan. Hasil dari proses ini menunjukkan bahwa data latih terdiri atas 1.814 data santri, sementara data uji terdiri dari 454 data. Dari total 2.268 data bersih, sebanyak 2.061 santri dalam data latih tergolong ke dalam kategori tidak lulus, sedangkan 207 santri termasuk kategori lulus. Pada data uji, terdapat 41 santri yang lulus dan 413 santri yang tidak lulus. Hasil *split* data ditunjukkan pada Gambar 4.7.

```
Train: (1814, 35)
Test : (454, 35)
Distribusi label di train:
 kelulusan
0    1648
1     166
Name: count, dtype: int64
Distribusi label di test:
 kelulusan
0    413
1     41
Name: count, dtype: int64
```

Gambar 4.7 Hasil *Split* Data

Distribusi yang dihasilkan memperlihatkan adanya ketidakseimbangan kelas, jumlah santri yang lulus jauh lebih banyak dibandingkan yang tidak lulus. Ketimpangan ini dapat memengaruhi kinerja model klasifikasi karena algoritma

cenderung lebih memprioritaskan kelas mayoritas. Oleh karena itu, langkah selanjutnya dalam proses pra-pemrosesan adalah penanganan ketidakseimbangan data agar model dapat mengenali kedua kelas secara lebih seimbang dan akurat. Salah satu teknik yang dapat diterapkan untuk tujuan tersebut adalah *Synthetic Minority Oversampling Technique* (SMOTE).

4.1.5 Penanganan Data *Imbalanced*

Setelah dilakukan proses pembagian data, didapatkan bahwa data latih (*training set*) semula memiliki ketidakseimbangan distribusi kelas, yaitu sebanyak 1648 data pada kelas tidak lulus (label 0) dan hanya 166 data pada kelas lulus (label 1). Ketimpangan ini berpotensi menyebabkan model *machine learning* yang dilatih akan bias terhadap kelas mayoritas, karena algoritma cenderung mengabaikan kelas minoritas akibat jumlah datanya yang jauh lebih sedikit.

Untuk mengatasi masalah ini, digunakan teknik *oversampling* yaitu SMOTE (*Synthetic Minority Oversampling Technique*). SMOTE bekerja dengan menghasilkan data sintetis pada kelas minoritas berdasarkan kemiripan fitur, sehingga menambah jumlah sampel tanpa menduplikasi data yang ada secara langsung. Teknik ini membantu menciptakan representasi data yang lebih seimbang untuk proses pelatihan. Proses dilakukan menggunakan fungsi SMOTE dari pustaka *imblearn* dengan parameter *random_state* untuk memastikan replikasi hasil. Proses ini hanya diterapkan pada data latih untuk menghindari kebocoran data (*data leakage*) ke data uji. Hasil penerapan SMOTE ditunjukkan pada Gambar 4.8.

```
Distribusi label sebelum SMOTE:
 kelulusan
0    1648
1     166
Name: count, dtype: int64
Distribusi label setelah SMOTE:
 kelulusan
0    1648
1    1648
```

Gambar 4.8 Hasil Penerapan SMOTE

Penerapan SMOTE menghasilkan 1648 data sintetis ke kelas lulus (dari sebelumnya hanya 166), sehingga kini jumlahnya sama dengan kelas tidak lulus yang ada sebanyak 1648 data. Dengan keseimbangan ini, model *machine learning* yang dilatih diharapkan dapat mengenali pola dari kedua kelas secara lebih adil dan objektif, serta meningkatkan performa prediksi pada kelas minoritas.

4.1.6 Pelatihan Model

Setelah data selesai melalui tahap pra-pemrosesan dan penanganan ketidakseimbangan kelas menggunakan metode SMOTE, langkah selanjutnya adalah melakukan pelatihan model. Pada penelitian ini, dua algoritma *machine learning* digunakan, yaitu *Support Vector Machine* (SVM) dan *Random Forest* (RF), yang masing-masing memiliki karakteristik dan keunggulan tersendiri dalam proses klasifikasi.

Pelatihan dilakukan dengan menggunakan data latih yang telah diseimbangkan, di mana model mempelajari pola hubungan antara fitur-fitur *input* dan label target kelulusan. Setelah model selesai dilatih, dilakukan proses klasifikasi, yaitu memprediksi label kelulusan pada data uji yang belum pernah dilihat oleh model. Proses ini menjadi inti dari penerapan algoritma klasifikasi,

karena menunjukkan bagaimana model mampu menggeneralisasi dari data latih ke data uji.

Proses klasifikasi dalam penelitian ini dilakukan setelah model *machine learning* (SVM dan *Random Forest*) selesai dilatih menggunakan data latih yang telah melalui tahap pra-pemrosesan dan penyeimbangan kelas dengan metode SMOTE. Setelah model belajar dari pola data latih, langkah klasifikasi dilakukan dengan memberikan data uji (X_test) sebagai *input* ke model yang telah dilatih. Model kemudian memprediksi label kelas (status kelulusan santri) berdasarkan pola yang telah dipelajarinya dari data latih. Hasil prediksi ini dibandingkan dengan label aktual (y_test) untuk mengevaluasi kinerja model, dengan menggunakan metrik seperti akurasi, *precision*, *recall*, dan *f1-score*. Proses klasifikasi ini diimplementasikan dalam kode Python menggunakan fungsi *predict* () dari pustaka *Scikit-learn*, seperti model.predict (X_test). Evaluasi performa model dilakukan dengan mengukur nilai akurasi, *precision, recall*, dan *f1-score*, berdasarkan hasil prediksi terhadap data uji.

1. *Support Vector Machine* (SVM)

Setelah melalui tahap pelatihan menggunakan algoritma *Support Vector Machine* (SVM), model diuji terhadap data uji untuk mengukur performanya dalam memprediksi kelulusan santri menggunakan SVM kernel RBF dengan parameter C bernilai 1 dan Gamma bernilai *scale*. Hasil evaluasi ditampilkan melalui metrik akurasi, *precision, recall, F1-score*, serta *confusion matrix* yang ditunjukkan pada Gambar 4.9.

```
--- SUPPORT VECTOR MACHINE ---
Akurasi: 0.6497797356828194
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.67      0.78       413
           1       0.12      0.44      0.18        41

    accuracy                           0.65       454
   macro avg       0.52      0.55      0.48       454
weighted avg       0.85      0.65      0.72       454
```

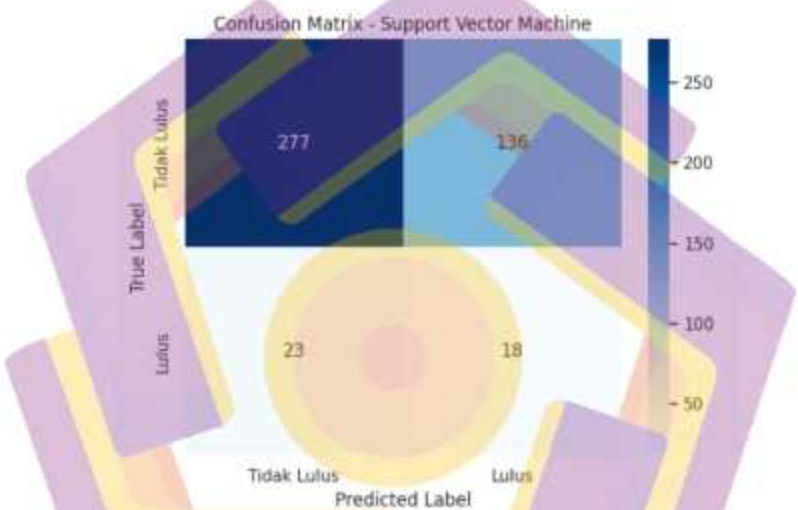
Gambar 4.9 Pelatihan Model *Support Vector Machine* (SVM)

Hasil evaluasi model *Support Vector Machine* (SVM) dengan data yang sudah dilabeli secara valid menunjukkan akurasi sebesar 64,97%. Nilai ini relatif sedang, namun masih cukup untuk menggambarkan pola umum kelulusan santri Pondok Pesantren Imam Bukhari. Pada kelas mayoritas, yaitu santri yang tidak lulus (label 0), model memberikan performa yang cukup baik. Hal ini terlihat dari *precision* sebesar 0,92, yang berarti sebagian besar santri yang diprediksi tidak lulus memang benar tidak lulus. *Recall* pada kelas ini mencapai 0,67, menunjukkan bahwa sekitar 67% santri tidak lulus berhasil dikenali dengan benar oleh model, meskipun 33% sisanya terlewat.

Sebaliknya, performa SVM pada kelas minoritas, yaitu santri yang lulus, masih belum optimal. *Precision* pada kelas ini hanya sebesar 0,12, artinya prediksi "lulus" seringkali salah sasaran. Namun, *recall* sebesar 0,44 mengindikasikan bahwa hampir setengah dari santri yang benar-benar lulus dapat dikenali oleh model. Nilai *f1-score* untuk kelas ini sebesar 0,18 juga menegaskan kelemahan SVM dalam menangani kelas minoritas.

Berdasarkan *confusion matrix*, dari 413 santri yang tidak lulus, sebanyak 277 berhasil diklasifikasikan dengan benar, sedangkan 136 salah diprediksi sebagai

lulus. Sementara itu, dari 41 santri yang lulus, 18 berhasil terdeteksi, tetapi 23 salah dikategorikan sebagai tidak lulus. Kondisi ini mengindikasikan bahwa SVM lebih condong untuk "memainkan aman" pada kelas mayoritas, sehingga lebih efektif dalam mengenali santri yang tidak lulus dibandingkan dengan santri yang lulus. *Confusion matrix* model SVM yang ditunjukkan pada Gambar 4.10.



Gambar 4.10 *Confusion Matrix* Model *Support Vector Machine* (SVM)

Hasil ini konsisten dengan karakteristik SVM yang sensitif terhadap distribusi data tidak seimbang. Walaupun sebelumnya data latih telah diseimbangkan dengan teknik SMOTE, ketimpangan jumlah data antara kelas lulus dan tidak lulus pada data uji masih berdampak pada kinerja model. Secara keseluruhan, SVM dapat digunakan ketika fokus utama penelitian adalah mendeteksi santri yang berisiko tidak lulus, namun belum cukup handal bila digunakan untuk mengidentifikasi santri yang diprediksi akan lulus.

2. *Random Forest* (RF)

Setelah melalui tahap pelatihan menggunakan algoritma *Support Vector Machine* (SVM) maka dilanjutkan dengan pelatihan menggunakan algoritma *Random Forest* (RF) dengan parameter *n_estimators* bernilai 100, *random state* bernilai 42 dan *max_depth* bernilai *none*. Hasil evaluasi ditampilkan melalui metrik akurasi, *precision, recall, F1-score*, serta *confusion matrix* yang ditunjukkan pada Gambar 4.11.

```
=== RANDOM FOREST ===
Akurasi: 0.7114537444493392
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.74      0.82       413
           1       0.15      0.46      0.22        41

    accuracy                           0.71       454
   macro avg       0.54      0.60      0.52       454
weighted avg       0.86      0.71      0.77       454
```

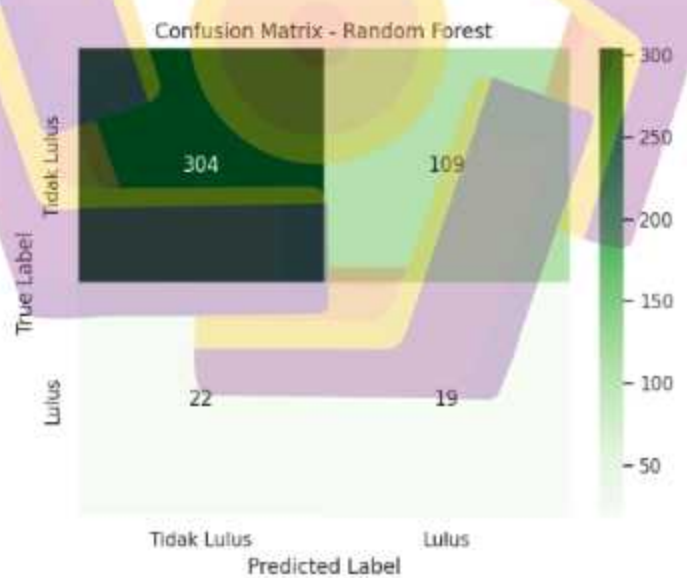
Gambar 4.11 Pelatihan Model *Random Forest* (RF)

Pengujian model Random Forest (RF) menghasilkan akurasi sebesar 71,14%, lebih tinggi dibandingkan SVM yang hanya mencapai 64,97%. Hal ini menunjukkan bahwa RF lebih stabil dalam melakukan prediksi secara keseluruhan. Pada kelas mayoritas, yaitu santri yang tidak lulus, performa RF cukup kuat dengan *precision* sebesar 0,93 dan *recall* sebesar 0,74. Artinya, sebagian besar prediksi "tidak lulus" benar, dan model mampu menangkap sekitar 74% santri yang memang tidak lulus. Nilai *f1-score* untuk kelas ini sebesar 0,82, menandakan keseimbangan yang baik antara *precision* dan *recall*.

Namun, performa RF pada kelas minoritas, yaitu santri yang lulus, masih rendah. *Precision* pada kelas ini hanya sebesar 0,15, menunjukkan bahwa banyak

prediksi "lulus" yang salah. *Recall* sebesar 0,46 juga menandakan bahwa hanya 46% dari santri yang benar-benar lulus berhasil terdeteksi oleh model. Nilai *f1-score* untuk kelas ini sangat rendah, yaitu 0,22, sehingga kemampuan RF dalam menangani kelas minoritas masih jauh dari ideal.

Berdasarkan *confusion matrix*, dari 413 santri yang tidak lulus, 304 berhasil diklasifikasikan dengan benar, sedangkan 109 salah diprediksi sebagai lulus. Pada sisi lain, dari 41 santri yang lulus, hanya 22 yang dikenali dengan benar, sementara 19 salah dikategorikan sebagai tidak lulus. Hal ini menunjukkan bahwa RF cenderung lebih "memihak" pada kelas mayoritas, sehingga kinerjanya lebih baik untuk memprediksi santri yang tidak lulus, tetapi lemah dalam mendeteksi santri yang lulus. *Confusion matrix* model RF ditunjukkan pada Gambar 4.12.



Gambar 4.12 *Confusion Matrix* Model *Random Forest* (RF)

Secara umum, hasil ini sejalan dengan karakteristik RF yang memang efektif untuk data berdimensi tinggi dan cenderung unggul dalam klasifikasi pada kelas dominan. Akan tetapi, sama seperti SVM, ketidakseimbangan data uji masih memengaruhi kinerja model. Dengan demikian, RF lebih cocok digunakan ketika tujuan utama adalah menjaga akurasi keseluruhan dan memastikan identifikasi yang tepat bagi kelompok santri yang tidak lulus.

4.1.7 Perbandingan Kinerja SVM dan RF

Tahap ini menyajikan perbandingan performa dua algoritma *machine learning*, yaitu *Support Vector Machine* (SVM) dan *Random Forest* (RF), dalam memprediksi kelulusan santri Pondok Pesantren Imam Bukhari yang akan melanjutkan studi ke Timur Tengah. Evaluasi dilakukan terhadap data uji menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score*. Selain melihat kinerja keseluruhan model, analisis juga difokuskan pada kemampuannya dalam mengklasifikasikan kedua kelas, yaitu lulus (kelas 1) dan tidak lulus (kelas 0), karena keduanya sama-sama penting dalam konteks seleksi akademik berbasis data. Skenario uji dilakukan dengan menggunakan SVM kernel RBF dengan parameter C bernilai 1 dan Gamma bernilai *scale* serta *Random Forest* (RF) dengan parameter *n_estimators* bernilai 100, *random state* bernilai 42 dan *max_depth* bernilai *none*. Perbandingan kinerja *Support Vector Machine* (SVM) dan *Random Forest* (RF) ditunjukkan pada Tabel 4.2.

Tabel 4.2 Perbandingan Kinerja SVM dan RF

Metrix	Support Vector Machine	Random Forest
Akurasi	64,98%	71,14%
Precision (kelas 0)	92%	93%
Recall (kelas 0)	67%	74%
F1-Score (kelas 0)	78%	82%
Precision (kelas 1)	12%	15%
Recall (kelas 1)	44%	46%
F1-Score (kelas 1)	18%	22%

Hasil evaluasi menunjukkan bahwa algoritma Random Forest memiliki kinerja yang lebih baik secara keseluruhan dibandingkan Support Vector Machine (SVM) dalam memprediksi kelulusan santri Pondok Pesantren Imam Bukhari untuk melanjutkan studi ke Timur Tengah.

Random Forest mencatat akurasi sebesar 71,14%, lebih tinggi dibandingkan SVM yang hanya mencapai 64,98%. Pada kelas mayoritas (santri tidak lulus), RF mampu memberikan hasil yang lebih stabil dengan *precision* 93%, *recall* 74%, dan *F1-score* 82%. Hal ini menunjukkan bahwa Random Forest cukup efektif dalam mengenali kelompok besar santri yang tidak lulus dengan tingkat kesalahan yang relatif kecil. Sebaliknya, performa pada kelas minoritas (santri lulus) masih rendah di kedua algoritma, meskipun *Random Forest* tetap lebih unggul dengan *precision* 15% dan *recall* 46% dibandingkan SVM yang hanya mencapai *precision* 12% dan *recall* 44%. Hal ini dapat dimaklumi karena jumlah santri lulus dalam data jauh lebih sedikit dibandingkan santri tidak lulus, sehingga meskipun sudah diterapkan SMOTE, ketidakseimbangan tetap memengaruhi hasil prediksi.

Dengan demikian, Random Forest lebih cocok digunakan ketika tujuan utama adalah mengidentifikasi dengan tepat kelompok santri yang tidak lulus (kelas

mayoritas) agar sistem seleksi lebih akurat secara umum. Sementara itu, SVM walaupun akurasinya lebih rendah, masih memberikan sensitivitas yang relatif baik dalam mengenali santri lulus (kelas minoritas), sehingga dapat dipertimbangkan ketika fokusnya adalah mendeteksi peluang keberhasilan.

Untuk memperoleh gambaran yang lebih jelas mengenai performa model, dilakukan pengujian dengan tiga skenario parameter yang berbeda pada algoritma *Support Vector Machine* (SVM) dan *Random Forest* (RF). Setiap skenario dirancang untuk merepresentasikan pendekatan parameterisasi yang berbeda, mulai dari konfigurasi sederhana (*baseline*), pengaturan dengan tingkat kompleksitas menengah (*balanced*), hingga kombinasi parameter yang lebih ketat untuk memaksimalkan kinerja tertentu seperti ditunjukkan pada Tabel 4.3.

Tabel 4.3 Hasil Uji Model

Uji Skenario	Parameter		Hasil	
	SVM	RF	SVM	RF
Skenario 1	kernel: rbf C: 0.1 gamma: scale	n_estimators: 200 max_depth: 10 min_samples_split: 5 min_samples_leaf: 2	1. Akurasi: 64,97% 2. Precision: • Kelas 0: 0.94 • Kelas 1: 0.14 3. Recall: • Kelas 0: 0.66 • Kelas 1: 0.59 4. F1-score: • Kelas 0: 0.77 • Kelas 1: 0.23	1. Akurasi: 69,38% 2. Precision: • Kelas 0: 0.94 • Kelas 1: 0.16 3. Recall: • Kelas 0: 0.70 • Kelas 1: 0.59 4. F1-score: • Kelas 0: 0.81 • Kelas 1: 0.26
Skenario 2	kernel: rbf C: 10 gamma: scale	n_estimators: 500 max_depth: 30 min_samples_split: 2 min_samples_leaf: 1	1. Akurasi 65,41%, 2. Precision • kelas 0 = 0,92 • kelas 1 = 0,12 3. Recall • kelas 0 = 0,68 • kelas 1 = 0,44 4. F1-score • kelas 0 = 0,78 • kelas 1 = 0,19	1. Akurasi 70,92%, 2. Precision • kelas 0 = 0,93 • kelas 1 = 0,14 3. Recall • kelas 0 = 0,74 • kelas 1 = 0,44 4. F1-score • kelas 0 = 0,82 • kelas 1 = 0,21
Skenario 3	kernel = poly, C = 10, gamma = scale	n_estimators: 150 max_depth: None min_samples_split: 4	1. Akurasi: 62,1% 2. Precision: • Kelas 0 = 0,94	1. Akurasi: 69,38% 2. Precision: • Kelas 0 = 0,93

		min_samples_leaf: 2 Bootstrap: False	• Kelas 1 = 0,14 3. Recall: • Kelas 0 = 0,62 • Kelas 1 = 0,61 4. F1-score: • Kelas 0 = 0,75 • Kelas 1 = 0,23	• Kelas 1 = 0,14 3. Recall: • Kelas 0 = 0,71 • Kelas 1 = 0,49 4. F1-score: • Kelas 0 = 0,81 • Kelas 1 = 0,22

Pengujian yang dilakukan pada tiga skenario konfigurasi model *Support Vector Machine* (SVM) dan *Random Forest* (RF) dengan parameter yang berbeda menghasilkan:

1. Skenario 1 menggunakan SVM kernel RBF dengan C rendah (0,1) dan gamma *scale*, sedangkan RF menggunakan n_estimators 200, max_depth 10, min_samples_split 5, dan min_samples_leaf 2. Pada skenario ini, menunjukkan bahwa SVM mampu mendeteksi kelas minoritas (tidak lulus) dengan *recall* yang cukup tinggi, yakni 0.59, walaupun *precision*-nya masih rendah di angka 0.14. Hal ini menghasilkan *f1-score* kelas minoritas sebesar 0.23. Sementara itu, *Random Forest* memberikan performa keseluruhan yang lebih seimbang, dengan akurasi 69,38% dan *f1-score* kelas mayoritas (lulus) sebesar 0.81. Model RF menunjukkan keunggulan pada kestabilan prediksi terhadap kelas mayoritas, sedangkan SVM lebih menonjol dalam kepekaan mendeteksi santri yang berpotensi tidak lulus.
2. Skenario 2 menggunakan SVM kernel RBF dengan C = 1.0 dan gamma *scale* (setelah penyesuaian dari *auto* untuk menghindari nilai *precision* nol), sedangkan RF mempertahankan konfigurasi sebelumnya. Hasilnya SVM mengalami peningkatan kecil pada akurasi menjadi 65,41%, tetapi *recall* untuk kelas minoritas turun menjadi 0.44. Hal ini menyebabkan sensitivitas terhadap santri yang tidak lulus melemah dibanding skenario pertama. *Random Forest* justru semakin unggul, dengan akurasi 70,92% dan *f1-score* kelas mayoritas sebesar 0.82. Meskipun *recall* pada kelas minoritas masih

rendah (0.44), RF tetap lebih stabil dalam mendeteksi santri yang lulus dengan *precision* dan *recall* yang konsisten. Dengan demikian, skenario kedua menegaskan bahwa RF dengan parameter kompleksitas tinggi lebih efektif untuk prediksi umum, sementara SVM kehilangan keunggulan pada deteksi kelas minoritas.

3. Skenario 3 menggunakan SVM kernel polynomial (degree 3) dengan C = 1.0 dan gamma *scale*, sementara RF menggunakan n_estimators 150, max_depth *None*, min_samples_split 4, min_samples_leaf 2, dan bootstrap = False. Pada skenario ini, hasil menunjukkan penurunan akurasi SVM menjadi 62,1%, dengan *f1-score* kelas minoritas hanya 0.23. Meskipun *recall* untuk kelas minoritas masih cukup tinggi (0.61), ketidakmampuan menjaga precision membuat hasil SVM kurang optimal. Di sisi lain, *Random Forest* tetap konsisten dengan akurasi 69,38% dan *f1-score* kelas mayoritas 0.81. Menariknya, pada skenario ini, RF sedikit lebih baik dalam recall kelas minoritas (0.49) dibandingkan dua skenario sebelumnya, meskipun precision tetap rendah. Dengan kata lain, ketika parameter dibuat lebih fleksibel, RF mampu menyeimbangkan prediksi antar kelas, sedangkan SVM dengan kernel polynomial justru kehilangan stabilitas performa.

Secara keseluruhan, ketiga skenario yang diuji menunjukkan pola performa yang relatif konsisten. Algoritma *Random Forest* terbukti lebih unggul dalam hal akurasi dan kestabilan, terutama pada prediksi kelas mayoritas, yaitu santri yang lulus. Sebaliknya, *Support Vector Machine* lebih menonjol dalam hal *recall* pada

kelas minoritas, yakni santri yang tidak lulus, meskipun kelemahan utamanya terletak pada *precision* yang rendah sehingga menghasilkan banyak kesalahan prediksi positif. Hasil ini menegaskan bahwa *Random Forest* lebih tepat digunakan ketika tujuan penelitian adalah menghasilkan prediksi umum dengan tingkat akurasi tinggi, sedangkan SVM dapat dipertimbangkan dalam konteks di mana deteksi awal terhadap santri yang berpotensi tidak lulus lebih diprioritaskan. Dengan demikian, kedua algoritma memiliki peran yang saling melengkapi sesuai dengan fokus evaluasi yang ingin dicapai oleh lembaga pendidikan.

4.1.8 Analis *Feature Importance*

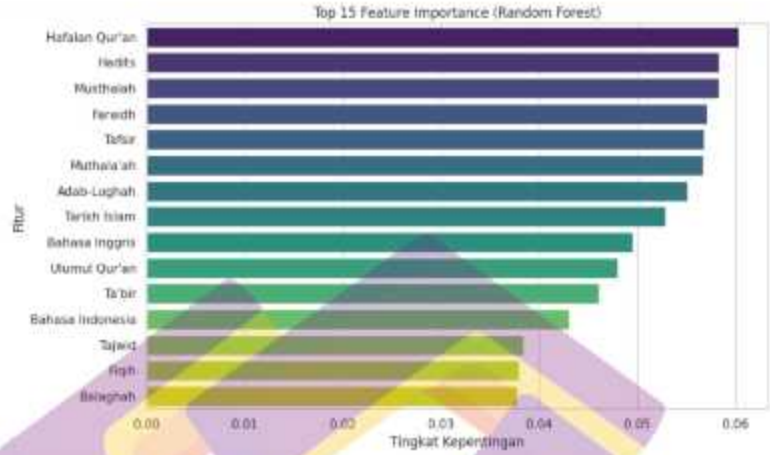
Berdasarkan hasil analisis *feature importance* dari model Random Forest, terlihat bahwa variabel Hafalan Qur'an menempati posisi teratas sebagai faktor paling dominan dalam menentukan prediksi kelulusan santri. Hal ini menunjukkan bahwa kemampuan menghafal Al-Qur'an menjadi kriteria utama yang sangat memengaruhi keberhasilan santri untuk melanjutkan studi ke Timur Tengah.

Sebelum hasil *feature importance* diinterpretasikan, dilakukan proses seleksi fitur untuk memastikan bahwa fitur yang digunakan dalam analisis benar-benar relevan dengan tujuan penelitian. Seleksi ini dilakukan dengan meninjau korelasi antar variabel serta mempertimbangkan konteks akademik dari setiap fitur. Fitur-fitur yang memiliki korelasi tinggi atau kontribusi yang sangat kecil terhadap variabel target dieliminasi agar model tidak bias terhadap atribut tertentu. Langkah ini menyebabkan perubahan pada hasil *importance factor*, di mana variabel dengan pengaruh rendah sebelumnya dapat tereduksi, dan variabel dominan menjadi lebih

menonjol. Dengan demikian, hasil *feature importance* yang ditampilkan pada penelitian ini merupakan hasil akhir setelah pemurnian fitur yang relevan, sehingga mencerminkan pengaruh nyata setiap faktor terhadap kelulusan santri.

Selain hafalan, mata pelajaran berbasis keagamaan seperti Hadits, Musthalah, Faraidh, dan Tafsir juga menempati urutan tinggi dalam kontribusinya. Kondisi ini menguatkan bahwa kompetensi pada bidang ilmu agama menjadi aspek fundamental dalam seleksi akademik, sesuai dengan karakteristik studi di Timur Tengah yang menitikberatkan pada penguasaan ilmu keislaman. Di sisi lain, pelajaran umum seperti Bahasa Inggris dan Bahasa Indonesia tetap masuk ke dalam daftar penting, meskipun kontribusinya lebih rendah dibandingkan mata pelajaran agama. Kehadiran mata pelajaran ini menunjukkan bahwa meskipun bobot seleksi lebih berat pada ilmu agama, kemampuan bahasa tetap memberikan pengaruh terhadap keberhasilan akademik santri. Menariknya, beberapa mata pelajaran seperti Balaghah dan Fiqih muncul dengan tingkat kepentingan yang relatif lebih rendah dibandingkan pelajaran agama lain, yang bisa jadi disebabkan oleh distribusi nilai yang lebih seragam atau tingkat penguasaan santri yang cenderung rata-rata.

Secara keseluruhan, analisis ini menggambarkan bahwa hafalan Al-Qur'an dan pemahaman ilmu keislaman adalah faktor inti yang menentukan kelulusan. Sementara itu, pelajaran umum berperan sebagai faktor penunjang yang tetap relevan, namun tidak sepenting aspek agama. Temuan ini sejalan dengan tujuan pendidikan pesantren dan kriteria seleksi untuk studi ke Timur Tengah yang menekankan keunggulan pada bidang keilmuan Islam. Hasil analis *feature importance* ditunjukkan pada Gambar 4.13.



Gambar 4.13 Hasil Analis *Feature Importance*

4.2 Pembahasan

Hasil penelitian menunjukkan bahwa baik *Random Forest* maupun *Support Vector Machine* (SVM) mampu memberikan performa klasifikasi, namun dengan pola yang berbeda pada masing-masing metrik evaluasi. *Random Forest* memperoleh akurasi yang lebih tinggi sebesar 71,14% dibandingkan SVM yang mencapai 64,98%. Model *Random Forest* juga menunjukkan kekuatan yang lebih baik dalam mengklasifikasikan santri lulus (kelas 0), dengan *precision* sebesar 93%, *recall* 74%, dan *F1-score* 82%. Hal ini menegaskan kemampuan *Random Forest* dalam memberikan prediksi yang relatif stabil pada kelas mayoritas, meskipun proporsi santri lulus dalam data jauh lebih dominan.

Sebaliknya, meskipun akurasi SVM lebih rendah, algoritma ini memperlihatkan kemampuan yang lebih kompetitif dalam mendeteksi santri tidak lulus (kelas 1). SVM mencatat *recall* sebesar 44% untuk kelas minoritas ini,

dibandingkan *Random Forest* yang hanya mencapai 46%. Walaupun nilai precision pada kelas 1 masih rendah (12% untuk SVM dan 15% untuk RF), temuan ini tetap menegaskan sensitivitas SVM dalam mengenali kelompok minoritas yang rentan tidak lulus, meski dengan konsekuensi tingginya *false positive*.

Analisis *feature importance* dari *Random Forest* mengungkapkan bahwa faktor-faktor dominan dalam memengaruhi prediksi kelulusan adalah hafalan Al-Qur'an, Hadits, Musthalah, dan Faraidh. Variabel-variabel ini menggambarkan bahwa kompetensi agama, khususnya hafalan dan pemahaman ilmu keislaman, menjadi kunci utama dalam menentukan kelulusan santri untuk studi ke Timur Tengah. Di sisi lain, mata pelajaran umum seperti Bahasa Inggris dan Bahasa Indonesia tetap memberikan kontribusi meskipun tidak sebesar mata pelajaran agama.

Jika dibandingkan dengan penelitian Nurfadilla et al. (2022), hasil ini kembali menegaskan keunggulan *Random Forest* dalam hal akurasi dan kestabilan model, meskipun tingkat akurasi pada penelitian ini (71,14%) lebih rendah. Hal ini kemungkinan besar disebabkan oleh kompleksitas data santri yang mencakup nilai akademik, kemampuan bahasa, serta hafalan Al-Qur'an. Sementara itu, hasil SVM yang lebih baik dalam mengenali santri tidak lulus sejalan dengan temuan Lukman dan Herlinda (2024), yang menunjukkan bahwa SVM memiliki sensitivitas tinggi terhadap kasus minoritas dalam data tidak seimbang.

Secara umum, penelitian ini memperlihatkan bahwa Random Forest lebih efektif ketika tujuan penelitian adalah memaksimalkan akurasi keseluruhan prediksi kelulusan, sementara SVM lebih tepat digunakan ketika fokus diarahkan

pada deteksi dini santri yang berisiko tidak lulus. Kedua model ini memiliki kekuatan yang saling melengkapi dan dapat dipilih sesuai dengan kebutuhan kebijakan akademik pondok pesantren.

BAB 5

PENUTUP

5.1. Kesimpulan

Berdasarkan hasil dan pembahasan yang telah dilakukan maka dapat diambil kesimpulan sebagai berikut:

1. Penanganan dataset yang imbalanced dilakukan menggunakan teknik SMOTE (*Synthetic Minority Oversampling Technique*). Strategi ini berhasil menyeimbangkan distribusi kelas antara santri yang lulus dan tidak lulus sehingga model dapat belajar lebih baik tanpa bias ke kelas mayoritas. Selain itu, melalui analisis *feature importance Random Forest*, diperoleh bahwa fitur dengan kontribusi terbesar terhadap prediksi kelulusan adalah nilai hafalan Al-Qur'an, Hadits, Musthalah, dan Faraidh, sedangkan mata pelajaran umum seperti Bahasa Inggris dan Bahasa Indonesia berperan lebih kecil. Hal ini menunjukkan bahwa fitur terbaik untuk prediksi berkaitan erat dengan kompetensi agama yang menjadi fokus seleksi studi ke Timur Tengah.
2. Pemilihan fitur yang tepat dan pengaturan *hyperparameter* berpengaruh signifikan terhadap kinerja model. Pada SVM, variasi parameter C dan gamma memengaruhi kemampuan model dalam mengklasifikasikan santri yang tidak lulus, di mana nilai C rendah cenderung meningkatkan recall pada kelas minoritas. Pada *Random Forest*, pengaturan jumlah estimator, kedalaman pohon (*max_depth*), serta ukuran minimal *split* dan *leaf*

berpengaruh terhadap keseimbangan antara akurasi keseluruhan dan kemampuan mengenali kelas minoritas. Hasil ini menegaskan bahwa kombinasi fitur yang relevan dengan setting parameter yang optimal sangat penting untuk meningkatkan performa prediksi.

3. Perbandingan kinerja antara SVM dan *Random Forest* menunjukkan pola yang konsisten. *Random Forest* unggul dalam akurasi keseluruhan (sekitar 71%) dan *F1-score* kelas mayoritas, sehingga lebih tepat digunakan untuk memprediksi santri yang berpotensi lulus. Sebaliknya, SVM lebih sensitif dalam mendeteksi santri yang tidak lulus, terbukti dari nilai *recall* yang lebih tinggi pada kelas minoritas dibandingkan *Random Forest*. Dengan demikian, pemilihan algoritma dapat disesuaikan dengan tujuan: RF lebih tepat jika fokus pada akurasi umum, sedangkan SVM lebih sesuai ketika tujuan utama adalah deteksi dini santri yang berisiko gagal lulus.

5.2. Saran

Berdasarkan kesimpulan di atas penulis memberikan saran yaitu:

1. Implementasi sistem *real-time*: Hasil model prediksi ini dapat dijadikan dasar untuk mengembangkan sistem pendukung keputusan berbasis web yang dapat digunakan oleh pihak pondok dalam proses seleksi internal.
2. Evaluasi model lebih lanjut: Diperlukan uji validasi silang (*cross-validation*) serta uji data dari pondok pesantren yang berbeda untuk memastikan generalisasi model.

DAFTAR PUSTAKA

F. A. Abdelfattah, O. S. Obeidat, Y. A. Salahat, M. B. BinBakr, and A. A. Al Sultan, "The predictive validity of entrance scores and short-term performance for long-term success in engineering education," *JARHE*, vol. 14, no. 4, pp. 1272–1285, Dec. 2022, doi: 10.1108/JARHE-04-2021-0126.

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.

Md. A. A. Walid, S. M. M. Ahmed, M. Zeyad, S. M. S. Galib, and M. Nesa, "Analysis of machine learning strategies for prediction of passing undergraduate admission test," *International Journal of Information Management Data Insights*, vol. 2, no. 2, p. 100111, Nov. 2022, doi: 10.1016/j.jjimei.2022.100111.

Z. Nurfadilla, "IMPLEMENTASI DATA MINING UNTUK MEMPREDIKSI KELULUSAN MAHASISWA TEPAT WAKTU MENGGUNAKAN RANDOM FOREST," vol. 1, no. 1, 2022.

C. Wulandari, T. H. B. Aviani, and R. Saputra, "Penerapan Algortima Support Vector Machine (SVM) Untuk Prediksi Tingkat Kelulusan Siswa SMA," vol. 4, no. 4, 2024.

E. Novianto, S. Suhirman, and D. Prasetyo, "PERBANDINGAN METODE KLASIFIKASI RANDOM FOREST DAN SUPPORT VECTOR MACHINE DALAM MEMPREDIKSI CAPAIAN STUDI MAHASISWA," vol. 9, no. 4, 2024.

A. K. Darmawan, I. Yudhisari, A. Anwari, and M. Makruf, "Pola Prediksi Kelulusan Siswa Madrasah Aliyah Swasta dengan Support Vector Machine dan Random Forest," vol. 12, 2023.

L. Lukman and H. Herlinda, "Prediksi Kelulusan Siswa dengan Metode Support Vector Machine (SVM) di SMK Adiluhur," *STRING*, vol. 9, no. 1, p. 115, Aug. 2024, doi: 10.30998/string.v9i1.23355.

A. F. A. Naibaho and A. Zahra, "PREDIKSI KELULUSAN SISWA SEKOLAH MENENGAH PERTAMA MENGGUNAKAN MACHINE LEARNING," *JITET*, vol. 11, no. 3, Jul. 2023, doi: 10.23960/jitet.v11i3.3056.

A. Fatunnisa and H. Marcos, "Prediksi Kelulusan Tepat Waktu Siswa SMK Teknik Komputer Menggunakan Algoritma Random Forest," *JAMIKA*, vol. 14, no. 1, pp. 101–111, Apr. 2024, doi: 10.34010/jamika.v14i1.12114.

J. Zeniarja, A. Salam, and F. A. Ma'ruf, "Seleksi Fitur dan Perbandingan Algoritma Klasifikasi untuk Prediksi Kelulusan Mahasiswa," *JRE*, vol. 18, no. 2, Jul. 2022, doi: 10.17529/jre.v18i2.24047.

Oon Wira Yuda, Darmawan Tuti, Lim Sheih Yee, and Susanti, "Penerapan Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Random Forest," *SATIN*, vol. 8, no. 2, pp. 122–131, Dec. 2022, doi: 10.33372/stn.v8i2.885.

Ambarwari, A., Adrian, Q. J., & Herdiyeni, Y. (2021). Analisis Pengaruh Data Scaling Terhadap Performa Algoritme Machine Learning untuk Identifikasi Tanaman. *Rekayasa Sistem dan Teknologi Informasi (RESTI)*, 117-122.

LAMPIRAN