PREDIKSI SANTRI PUTUS SEKOLAH DI PONDOK PESANTREN SIDOGIRI MENGGUNAKAN METODE RANDOM FOREST DAN DECISION TREE

TESIS

Oleh:
MUHAMMAD MAHSUN
NIM. 220605210013

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025

PREDIKSI SANTRI PUTUS SEKOLAH DI PONDOK PESANTREN SIDOGIRI MENGGUNAKAN METODE RANDOM FOREST DAN DECISION TREE

TESIS

Diajukan kepada:
Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang
Untuk Memenuhi Salah Satu Persyaratan Dalam
Memperoleh Gelar Magister Komputer (M.Kom)

Oleh:
MUHAMMAD MAHSUN
NIM. 220605210013

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2025

PREDIKSI SANTRI PUTUS SEKOLAH DI PONDOK PESANTREN SIDOGIRI MENGGUNAKAN METODE RANDOM FOREST DAN DECISION TREE

TESIS

Diajukan kepada:
Fakultas Sains dan Teknologi
Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang
Untuk Memenuhi Salah Satu Persyaratan Dalam
Memperoleh Gelar Magister Komputer (M.Kom)

Oleh:
MUHAMMAD MAHSUN
NIM. 220605210013

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025

PREDIKSI SANTRI PUTUS SEKOLAH DI PONDOK PESANTREN SIDOGIRI MENGGUNAKAN METODE RANDOM FOREST DAN DECISION TREE

TESIS

Oleh:
MUHAMMAD MAHSUN
NIM. 220605210013

Telah diperiksa dan disetujui untuk diuji:
Tanggal: 10 Desember 2025

Pembimbing I,	Pembimbing II,
Dr. Ir. Mokhamad Amin Hariyadi, M.T	Prof. Dr. Sri Harini, M.Si
NIP. 19670018 200501 1 001	NIP. 19731014 200112 2 002

Mengetahui,
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Prof. Dr. H. Muhammad Faisal, S.Kom., M.T
NIP. 19740510 200501 1 007

PREDIKSI SANTRI PUTUS SEKOLAH DI PONDOK PESANTREN SIDOGIRI MENGGUNAKAN METODE RANDOM FOREST DAN DECISION TREE

TESIS

Oleh:
MUHAMMAD MAHSUN
NIM. 220605210013

Telah Dipertahankan di Depan Dewan Penguji Thesis
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Magister Komputer (M.Kom)
Tanggal: 10 Desember 2025

Susunan Dewan Penguji			**Tanda Tangan**
Penguji I	:	Dr. Yunifa Miftachul Arif, M.T NIP. 19830616 201101 1 004	( )
Penguji II	:	Prof. Dr. Ir. Muhammad Faisal, S.Kom., M.T NIP. 19740510 200501 1 007	( )
Pembimbing I	:	Dr. Ir. Mokhamad Amin Hariyadi, M.T NIP. 19670018 200501 1 001	( )
Pembimbing II	:	Prof. Dr. Sri Harini, M. Si NIP. 19731014 200112 2 002	( )

Mengetahui dan Mengesahkan
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Prof. Dr. Ir. Muhammad Faisal, S. Kom., M.T
NIP. 19740510 200501 1 007

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan dibawah ini:

Nama	: Muhammad Mahsun
NIM	: 220605210013
Program Studi	: Magister Informatika
Fakultas	: Sains dan Teknologi
Judul Tesis	: "Prediksi Santri Putus Sekolah di Pondok Pesantren Sidogiri Menggunakan Metode Random Forest dan Decision Tree"

Menyatakan dengan sebenarnya bahwa Tesis yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambilalihan data, tulisan atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan Tesis ini hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 20 Desember 2025

Yang membuat pernyataan,

Muhammad Mahsun

NIM. 220605210013

MOTTO

العِلْمُ بِلا عَمَلٍ كَالشَّجَرِ بِلا ثَمَرٍ

Ilmu pengetahuan yang tidak disertai pengamalan,
layaknya pohon yang tidak berbuah.

PERSEMBAHAN

ٱلْحَمْدُ لِلَّهِ رَبِّ ٱلْعَالَمِينَ

Puji syukur atas kehadirat Allah ﷻ shalawat serta salam kepada Rasulullah ﷺ. Dengan iringan doa dan rasa hormat yang mendalam, karya ini penulis dedikasikan sepenuhnya kepada:

Istri tercinta Auliana Putri Zakiah, yang selalu memberi semangat, yang sering mendo'akan kebaikan, serta selalu memotivasi.

Orang tua tercinta, Bapak (Alm.) Muhammad Mustahal, Mohammad Irham Zuhdi, Umi Ummul Bariroh, serta Ibu Dewi Safiah, atas limpahan doa, bimbingan, dan dukungan yang tak terhingga. Semoga Allah ﷻ membalas seluruh pengorbanan dengan sebaik-baik balasan. *Al-Fatihah* untuk almarhum Bapak.

Saudara dan saudari tercinta, khususnya Muhammad Makmun dan Muhammad Faris, atas doa dan dukungan yang senantiasa menguatkan.

Dosen pembimbing, Dr. Ir. Mokhamad Amin Hariyadi, M.T. dan Prof. Dr. Sri Harini, M.Si., atas bimbingan, arahan, dan pengalaman keilmuan yang sangat berharga selama proses penelitian.

Segenap sivitas akademika Program Studi Magister Informatika, khususnya seluruh dosen, serta seluruh rekan mahasiswa Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang, atas kebersamaan dan perjuangan selama masa studi.

Penulis mengucapkan *jazakumullah khairan katsira*. Semoga Allah ﷻ senantiasa memudahkan setiap langkah perjuangan dan membalasnya dengan keberkahan yang berlipat ganda. *Aamiin, ya Mujibassailin*.

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh

Alhamdulillāh, segala puji bagi Allah ﷻ atas rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan studi pada Program Studi Magister Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang, serta merampungkan tesis ini dengan baik.

Penulis menyampaikan terima kasih dan doa *jazākumullāh ahsanal jazā'* kepada seluruh pihak yang telah membantu, membimbing, dan mendukung selama proses penyusunan tesis ini. Secara khusus, ucapan terima kasih penulis sampaikan kepada:

1. Bapak Dr. Mokhamad Amin Hariyadi, M.T**.**, dan Ibu Prof. Dr. Hj. Sri Harini, M.Si**.,** selaku dosen pembimbing tesis, yang telah memberikan bimbingan, dorongan, serta wawasan ilmiah yang sangat berharga, sehingga penulis dapat menyelesaikan penelitian ini dengan baik dan terarah.
2. Bapak Dr. Yunifa Miftachul Arif, M.T**.,** dan Bapak Prof. Dr. Muhammad Faisal, M.T**.**, selaku dosen penguji tesis yang telah memberikan masukan, arahan, dan koreksi yang sangat berharga bagi penyempurnaan penelitian ini.
3. Ketua pengurus Pusat Ikatan Alumni Santri Sidogiri, H. Achmad Sa'dulloh Abd. Alim, segenap pengurus urusan TMTB & Dai, serta Lembaga Amil Zakat Sidogiri yang telah memberikan kepercayaan, dukungan moril, maupun materiil melalui program beasiswa studi magister.
4. Seluruh civitas akademika Program Studi Magister Informatika, khususnya Bapak/Ibu dosen yang telah memberikan ilmu, wawasan, dan bimbingan kepada penulis selama menempuh studi.
5. Keluarga tercinta yang selalu memberikan doa, dukungan, dan semangat, sehingga menjadi motivasi terbesar bagi penulis dalam menyelesaikan tesis ini

Penelitian ini memiliki tentu keterbatasan. Oleh sebab itu, masukan, kritik, dan saran yang membangun penulis harapkan untuk penyempurnaan pada penelitian selanjutnya. Penulis berharap tesis ini dapat memberikan manfaat bagi pembaca, menjadi referensi bagi penelitian di masa mendatang, serta menjadi amal yang senantiasa mengalir bagi penulis. *Āmīn Yā Rabbal 'Ālamīn.*

*W*assalamu'alaikum *Warahmatullahi Wabarakatuh*

Malang, 20 Desember 2025

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Mahsun, Muhammad. 2025. **Prediksi Santri Putus Sekolah di Pondok Pesantren Sidogiri menggunakan *Random Forest* dan *Decision Tree.*** Tesis. Program Studi Magister Informatika Fakultas Sains Dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. Ir. Mokhamad Amin Hariyadi, M.T.(II) Prof. Dr. Sri. Harini M. Si

Kata kunci: Prediksi Santri Putus Sekolah, *Machine Learning, Random Forest*, *Decision Tree,* Pondok Pesantren, Madrasah

Fenomena putus sekolah (*dropout*) merupakan isu krusial yang berdampak negatif terhadap kinerja institusi pendidikan formal maupun pesantren, stabilitas sosial, serta pembangunan sumber daya manusia. Oleh karena itu, mendeteksi dini terhadap siswa berisiko tinggi mengalami putus sekolah menjadi langkah preventif yang strategis. penelitian ini bertujuan menganalisa model prediksi yang akurat menggunakan pendekatan *Machine Learning*, dengan melakukan evaluasi komparatif terhadap algoritma *Random Forest* dan *Decision Tree*. *Dataset* penelitian diambil dari data siswa Madrasah Miftahul Ulum Pondok Pesantren Sidogiri dengan jumlah 1.763 data siswa. Hasil tahapan eksperimen model *Random Forest* memberikan performa terbaik dengan akurasi 84%, presisi 84%, recall 80%, dan F1-score 81%. Model dilatih dengan 4 skenario dan diuji dengan tuning parameter serta menggunakan matrik untuk mengevaulis akurasi model, untuk memastikan hasilnya konsisten. Metrik tersebut mengindikasikan bahwa model bekerja secara seimbang antara sensitivitas dan ketepatan prediksi, serta efektif dalam mengidentifikasi faktor internal dan eksternal yang berkontribusi terhadap risiko *dropout*. Berdasarkan hasil evalulasi model, *Random Forest* direkomendasikan sebagai instrumen pendukung keputusan untuk memfasilitasi intervensi yang lebih tepat sasaran, seperti dukungan akademik, ekonomi, maupun bimbingan pembinaan santri. Penelitian ini memiliki keterbatasan karena model hanya diuji pada lembaga Madrasah Miftahul Ulum Pondok Pesantren Sidogiri, sehingga penerapannya pada konteks lain perlu dikaji lebih lanjut. Penelitian lainnya diharapkan mencoba model lain dan menghasilkan model prediksi yang lebih akurat, adaptif, dan dapat digunakan sebagai instrumen pendukung keputusan yang lebih komprehensif bagi lembaga pendidikan madrasah maupun pesantren dalam mencegah terjadinya santri putus sekolah sebelum waktunya lulus.

ABSTRACT

Mahsun, Muhammad. 2025. **Prediction of Student Dropout at Sidogiri Islamic Boarding School Using Random Forest and Decision Tree.** Thesis. Master of Informatics Study Program, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University Malang. Supervisor: (I) Dr. Ir. Mokhamad Amin Hariyadi, M.T. (II) Prof. Dr. Sri. Harini M. Si

Keywords: Student Dropout Prediction, Machine Learning, Random Forest, Decision Tree, Islamic Boarding School, Madrasah

The phenomenon of student dropout is a critical issue that negatively affects the performance of educational institutions, both formal schools and Islamic boarding schools (pesantren), as well as social stability and human resource development. Therefore, early detection of students at high risk of dropping out is a strategic preventive measure. This study aims to analyze accurate predictive models using a machine learning approach by conducting a comparative evaluation of the Random Forest and Decision Tree algorithms. The dataset was obtained from the student records of Madrasah Miftahul Ulum at Sidogiri Islamic Boarding School, comprising 1,763 student data entries. The experimental results indicate that the Random Forest model achieved the best performance, with an accuracy of 84%, precision of 84%, recall of 80%, and an F1-score of 81%. The model was trained using four experimental scenarios and evaluated through parameter tuning and performance metrics to ensure consistency and reliability of the results. These metrics demonstrate that the model performs in a balanced manner, with respect to both sensitivity and predictive accuracy, and is effective in identifying both internal and external factors contributing to dropout risk. Based on the model evaluation, Random Forest is recommended as a decision-support tool to facilitate more targeted interventions, such as academic support, economic assistance, and student counselling programs. This study is limited in scope, as the model was tested only within Madrasah Miftahul Ulum at Sidogiri Islamic Boarding School; therefore, its applicability to other contexts requires further investigation. Future studies are encouraged to explore additional models to develop more accurate and adaptive prediction systems that can serve as comprehensive decision-support tools for madrasah and Islamic boarding schools in preventing student dropout before graduation

مستخلص البحث

محسون، محمد . ٢٠٢٥. التنبؤ بانقطاع الطلاب عن الدراسة في المعهد سيداقري السلفى باستخدام خوارزميتي الغابة العشوائية وشجرة القرار. رسالة الماجستير. قسم المعلومات، كلية العلوم والتكنولوجيا بجامعة مولانا مالك إبراهيم الإسلامية الحكومية مولانا مالك إبراهيم مالانج. المشرف الأول: د. محمد أمين هريادي، الماجستير؛ المشرف الثاني: د. سري هاريني ، الأستاذة ,الماجستير.

الكلمات المفتاحية : التنبؤ بانقطاع الطلاب، التعلم الآلي، الغابة العشوائية، شجرة القرار،المعهد الإسلامي، المدرسة.

تُعدّ ظاهرة انقطاع الطلاب عن الدراسة (التسرب) قضيةً بالغة الأهمية لما لها من آثار سلبية على أداء المؤسسات التعليمية، سواء المدارس النظامية أو المعاهد الإسلامية (البيسانترن)، فضلاً عن تأثيرها في الاستقرار الاجتماعي وتنمية الموارد البشرية. ومن ثمّ، فإن الكشف المبكر عن الطلاب المعرضين بدرجة عالية لخطر الانقطاع يُعد إجراءً وقائيًا استراتيجيًا. تهدف هذه الدراسة إلى تحليل نماذج تنبؤية دقيقة (Random Forest) باستخدام منهجية التعلم الآلي، من خلال إجراء تقييم مقارن بين خوارزميتي الغابة العشوائية و شجرة القرار (Decision Tree).

تم الحصول على مجموعة البيانات من سجلات طلاب مدرسة مفتاح العلوم التابعة لمعهد سيدوغيري ،الإسلامي، وبلغت 1,763 سجلاً طلابيًا. وأظهرت نتائج التجارب أن نموذج الغابة العشوائية حقق أفضل أداء F1- وقيمة ،80% (Recall) والاستدعاء ،84% (Precision) حيث بلغت الدقة 84%، والدقة الإيجابية بنسبة 81%. وقد تم تدريب النموذج عبر أربعة سيناريوهات تجريبية، مع ضبط المعلمات وتقييم score الأداء باستخدام مقاييس متعددة لضمان اتساق النتائج وموثوقيتها. وتشير هذه المؤشرات إلى أن النموذج يعمل بشكل متوازن بين الحساسية ودقة التنبؤ، كما يتمتع بقدرة فعّالة على تحديد العوامل الداخلية والخارجية التي تسهم في زيادة خطر الانقطاع.

وبناءً على نتائج التقييم، توصي الدراسة باعتماد خوارزمية الغابة العشوائية كأداة داعمة لاتخاذ القرار من أجل تنفيذ تدخلات أكثر دقة واستهدافًا، مثل الدعم الأكاديمي، والمساعدات المالية، وبرامج الإرشاد الطلابي. ومع ذلك، يقتصر نطاق هذه الدراسة على مدرسة مفتاح العلوم بمعهد سيدوغيري الإسلامي، مما يستلزم إجراء أبحاث إضافية لاختبار قابلية تطبيق النموذج في سياقات تعليمية أخرى. وتدعو الدراسات المستقبلية إلى استكشاف نماذج إضافية بهدف تطوير أنظمة تنبؤ أكثر دقة وتكيفًا، يمكن استخدامها كأدوات شاملة لدعم القرار في المدارس والمعاهد الإسلامية للحد من ظاهرة التسرب الدراسي قبل التخرج.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Dibelahan dunia, putus sekolah menjadi masalah serius yang perlu ditangani dengan serius pula bagi penyelenggara pendidikan. Bukan hanya di Spanyol, tetapi masalah putus sekolah juga terjadi di beberapa negara, Estonia, Inggris Raya, Latvia, Bangladesh, Korea Selatan pada penelitian Tayebi *et al.* (2021)

Menurut Masserini *et al*. (2021) di negara Eropa banyak siswa putus sekolah sebelum menyelesaikan program pendidikan. Di Negara Denmark, hanya sekitar 80% siswa mampu menyelesaikan sekolahnya sampai lulus, sementara di Italia hanya 46% dan faktor utama yang menyebabkan siswa putus sekolah adalah kondisi sosial ekonomi. Putus sekolah atau dropout merupakan salah satu permasalahan yang paling umum di dunia pendidikan serta memerlukan perhatian lebih. Bukan hanya di tingkat sekolah menengah, putus sekolah juga terjadi di Sebagian universitas. Siswa putus sekolah yang tinggi akan berdampak buruk pada universitas serta memperburuk reputasi universitas. Bagaimana strategi untuk menjaga siswa putus sekolah. Menurut Utari et al.(2020) melakukan prediksi awal siswa yang beresiko putus sekolah adalah sangat penting untuk menentukan tingkat keberhasilan suatu strategi di masing-masing universitas.

Mengurangi tingkat putus sekolah dengan melakukan penelitian adalah rencana yang baik dan perlu ditingkatkan intensitas penelitiannya. Mengetahui

faktor yang mempengaruhi siswa meninggalkan pendidikan di sekolah tujuan dari penelitian ini. Devasia et al. (2016) Memprediksi siswa dropout dengan akurasi tinggi akan membantu dalam mengidentifikasi siswa, menganalisis item data untuk membentuk ringkasan informasi yang berguna. Penelitian ini melakukan analisis berdasarkan pola data, asosiasi, hubungan antar semua data.

Pada penelitian yang dikerjakan oleh Sani *et al*. (2020) dalam memprediksi mahasiswa *dropout* dengan menggunakan model *Neural Network*, dengan objek mahasiswa, dan menghasilkan tingkat akurasi sebesar 95,86%, penelitian tersebut dilakukan pada institusi pendidikan formal dan menggunakan variabel akademik, sehingga belum mengakomodasi kompleksitas faktor non-akademik yang terdapat pada lembaga pendidikan berbasis pesantren. Sedangakan Hegde (2016) dan Fernandez *et al.* (2021) melakukan penelitian serupa dengan metode *Support Vector Machine* dengan akurasi masing-masing mencapai angka 90,24% dan 90,87%, dengan metode SVM model ini membutuhkan proses optimasi parameter yang kompleks

Demikian pula, Harwati et al. (2016) meneliti terkait dengan prediksi siswa putus sekolah menggunakan metode *Naive Bayes* pada data pendidikan formal dan menghasilkan akurasi sebesar 80,67%. Hasil ini memiliki keterbatasan dalam menangkap hubungan interaksi antar variabel yang umum terdapat pada data pendidikan. Sementara, Sa'ad et al. (2020) menerapkan metode *Extreme Learning Machine* (ELM) pada konteks pendidikan formal dan memperoleh akurasi sebesar 72%, sebagaimana penelitian Hoe *et al*. (2013) menggunakan

metode CHAID dengan akurasi sebesar 70,17%. yang relatif lebih rendah dibandingkan metode lainnya

Meskipun telah banyak penelitian terkait prediksi putus sekolah dengan pendekatan machine learning di lingkungan sekolah umum atau formal, masih sangat terbatas studi yang secara khusus menyoroti fenomena ini dalam konteks pendidikan pesantren, terutama pesantren klasik seperti Pondok Pesantren Sidogiri . Namun demikian, karakteristik lingkungan pesantren, termasuk sistem pembelajaran, faktor sosial-keagamaan, serta dinamika hubungan antara santri dan lembaga, berbeda dengan sekolah umum. Belum banyak model prediktif yang dirancang dengan mempertimbangkan variabel-variabel spesifik santri, seperti kepatuhan terhadap aturan pondok, keterlibatan dalam kegiatan keagamaan, dan latar belakang keluarga santri yang sangat beragam. Oleh karena itu, penelitian ini hadir untuk mengisi kesenjangan tersebut.

Dalam lima tahun terakhir, Pondok Pesantren Sidogiri mengalami peningkatan jumlah santri yang putus sekolah atau berhenti mengikuti pendidikan pesantren. Fenomena ini berpotensi mengganggu keberlangsungan proses belajar mengajar serta berdampak pada efektivitas pengelolaan pendidikan di lingkungan pesantren. Oleh karena itu, diperlukan suatu penelitian yang mampu mengidentifikasi dan memprediksi santri yang berisiko mengalami putus sekolah secara akurat dan sistematis.

Pemilihan metode analisis dalam penelitian ini menjadi aspek yang sangat penting karena berkaitan langsung dengan keandalan hasil prediksi yang dihasilkan. Penelitian ini menggunakan metode *Random Forest* (RF) sebagai

model utama, dengan *Decision Tree* (DT) sebagai metode pembanding. Pemilihan metode *Random Forest* dalam penelitian ini berdasarkan hasil kajian studi literatur yang menunjukkan bahwa metode ini mampu menghasilkan tingkat akurasi yang tinggi dalam permasalahan klasifikasi, khususnya pada data berdimensi tinggi dengan jumlah variabel yang besar serta adanya korelasi antar fitur dan metode ini bekerja dengan membangun sejumlah pohon keputusan secara acak dan menggabungkan hasil prediksinya, sehingga efektif dalam mengurangi risiko *overfitting* serta meningkatkan stabilitas dan akurasi model prediksi secara signifikan.

Menurut Purwanto *et al* (2025). Membangun model prediksi berbasis algoritma Random Forest yang telah terbukti kuat dalam klasifikasi data, menggunakan variabel yang relevan dan kontekstual dengan dunia pesantren, dan menghasilkan alat bantu berbasis data yang dapat digunakan oleh pengelola pesantren untuk melakukan intervensi lebih awal terhadap santri yang berisiko putus sekolah.

Sebagaimana termaktub di dalam Al Quran Surah Al-Taubah ayat 122 tentang peristiwa yang akan terjadi Allah berfirman:

وَمَا كَانَ الْمُؤْمِنُوْنَ لِيَنْفِرُوْا كَاۤفَّةً ۗفَلَوْلَا نَفَرَ مِنْ كُلِّ فِرْقَةٍ مِّنْهُمْ طَاۤىِٕفَةٌ لِّيَتَفَقَّهُوْا فِى الدِّيْنِ وَلِيُنْذِرُوْا ۝ قَوْمَهُمْ اِذَا رَجَعُوْٓا اِلَيْهِمْ لَعَلَّهُمْ يَحْذَرُوْنَ

"*Tidak sepatutnya orang-orang mukmin pergi semuanya (ke medan perang). Mengapa sebagian dari setiap golongan di antara mereka tidak pergi (tinggal bersama Rasulullah) untuk memperdalam pengetahuan agama mereka*

dan memberi peringatan kepada kaumnya apabila mereka telah kembali, agar mereka dapat menjaga dirinya?"(QS. Al-Taubah:122)

Ayat ini menguraikan kewajiban sebagian umat islam untuk menuntut ilmu dan pentingnya pendidikan yang berkelanjutan. Mohd *et al* (2024) perpandangan bahwa pentingnya pembelajaran sepanjang hayat (lifelong learning) bagi seseorang dalam pendidikan Islam. Dalam kontek ini menurut Usman et al.(2023). Pondok Pesantren Sidogiri memiliki tujuan secara khusus, mencetak santri untuk menuntut ilmu agama secara mendalam (tafaquh fiddin) dan menjadi santri Ibadillahis Sholihin, hamba Allah yang saleh, serta memastikan santri istikamah dalam mendalami ilmu hingga lulus dan rampung maupun tuntas.

1.2 Pernyataan Masalah

Fenomena santri putus sekolah di lingkungan pondok pesantren merupakan permasalahan yang kompleks karena dipengaruhi oleh berbagai faktor sosial, ekonomi, akademik, dan psikologis. Kompleksitas ini menimbulkan tantangan dalam mengidentifikasi santri yang berisiko putus sekolah secara dini. Oleh sebab itu, perlu pendekatan prediktif yang optimal menangani banyaknya atribut yang mempengaruhi secara efektif dan akurat. Berdasarkan latar belakang tersebut, permasalahan yang akan dikaji dalam penelitian ini dapat dirumuskan sebagai berikut:

1. Bagaimana membangun model klasifikasi yang optimal untuk memprediksi santri yang berisiko putus sekolah?
2. Bagaimana tingkat akurasi model prediksi dalam mendeteksi santri yang berpotensi putus sekolah?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah.

1. Menganalisis model klasifikasi prediktif untuk mengidentifikasi santri yang berisiko putus sekolah di Pondok Pesantren Sidogiri secara optimal.
2. Menganalisis tingkat akurasi dan kinerja model prediksi secara optimal dalam mengidentifikasi santri yang berpotensi putus sekolah

1.4 Batasan Masalah

1. Penelitian ini difokuskan pada santri yang berada di lingkungan Pondok Pesantren Sidogiri, Kraton, Pasuruan, Jawa Timur.
2. Dataset yang digunakan merupakan data santri yang mengalami putus sekolah di Pondok Pesantren Sidogiri dalam rentang waktu lima tahun, yaitu dari tahun pendidikan 2019 hingga tahun 2023.

1.5 Manfaat Penelitian

Manfaat penelitian ini diharapkan memberikan dampak positif:

1. Memberikan informasi berbasis data yang dapat digunakan oleh pimpinan pondok pesantren untuk menganalisis dan mengidentifikasi risiko santri yang berpotensi putus sekolah.

2. Mengungkap faktor-faktor dominan yang memengaruhi santri mengalami putus sekolah.
3. Mendukung upaya pencegahan dini dengan menyediakan informasi berbasis data jumlah santri yang putus sekolah.

BAB II

STUDI LITERATUR

2.1 Studi Literatur

Bab ini menjelaskan landasan teori dan argumentasi mendasar penelitian dan studi literatur mengenai penelitian yang terkait. Teori dan studi pustaka yang dipaparkan mengenai prediksi, pencegahan serta resiko siswa putus sekolah

Hasil studi Kemper *et al*. (2020) dalam pencegahan siswa putus sekolah dinilai sangat baik dan penting untuk sebuah lembaga pendidikan. Dalam penelitiannya Kemper *et al*. (2020) memaparkan temuan studinya dalam kasus data statistik pendidikan yang berfokus pada prediskisi deteksi mahsiswa dropout pada fakultas S1 Teknik Sistem (SE). Penelitian itu dilakukasn dengan menggunakan data mahasiswa setelah 6 tahun mereka mendaftarkan diri di Universitas Kolombia. Metode yang diterapkan untuk mengidentifikasi putus sekolah. Random Forest, Decission Tree, Logistik Regression dan Naive Bayes untuk menjadi prediksi terbaik. Kami menemukan Decision Tree menghasilkan akurasi yang lebih unggul daripada Logistik Regression. Namun, keduanya menghasilkan akurasi prediksi sama tinggi. Dalam penelitian ini hasil menunjukkan bahwa metode Random Forest mencapai tingkat yang tinggi dengan akurasi 97% setelahnya tiga semester.

Perez *et al*. (2018) memaparkan hasil temuannya dalam sebuah kasus pendidikan yang untuk mendeteksi mahasiswa sarjana yang *dropout* setelah 6

tahun mendaftar di Universitas Kolombia. kemudian, penliti menyajikan unsur temuan terkait kualitas data untuk meningkatkan proses pengumpulan data siswa. Identifikasi angka putus sekolah telah menjadi masalah utama bagi Universitas. Inisiatif ini dirancang oleh Center for Economic Studies (SEDE) di University of the Andes untuk menindaklanjuti masalah putus sekolah di pendidikan tinggi, hingga menghitung risiko desersi setiap siswa, dan untuk mengklasifikasikan siswa ini berdasarkan kelompok. Inisiatif ini dapat mendukung evaluasi strategi untuk setiap situasi yang mempengaruhi putus sekolah seperti status siswa, program akademik dan institusi; dan juga mempromosikan konsultasi, konsolidasi, interpretasi, dan penggunaan informasi. Dalam literatur, penelitian telah menyelidiki temuan karakteristik siswa dan konteksnya yang mempengaruhi keputusannya untuk keluar dari universitas. Dalam penelitian ini menyimpulkan bahwa putus sekolah siswa sangat terkait dengan tingkat akademik mereka (kinerja kelas dan perkembangan intelektual) dan integrasi sosial (interaksi kelompok sebaya dan interaksi fakultas) di universitas. Variabel seperti kehadiran siswa di kelas , jam yang dihabiskan untuk belajar setelah kelas, pendapatan keluarga, usia ibu dan pendidikan ibu secara signifikan terkait dengan putus sekolah siswa. Namun, masih belum ada konsensus dalam literatur tentang penyebab putus sekolah di universitas. Untuk melakukannya, kami memodelkan putus sekolah siswa menggunakan data yang dikumpulkan dari database akademik dari tahun 2004 hingga 2010. Informasi ini dapat digunakan dalam beberapa proses pendidikan seperti memprediksi pendaftaran kursus, memperkirakan

tingkat putus sekolah siswa, mendeteksi nilai tipikal dalam transkrip siswa, dan peningkatan model siswa yang akan memprediksi karakteristik atau kinerja akademik siswa. Menganalisis data akademik mahasiswa (jenis kelamin mahasiswa, usia mahasiswa, jurusan kemahasiswaan, nilai sma, dosen bergelar dosen, dan lain-lain) membangun model klasifikasi menggunakan metode pohon keputusan untuk meningkatkan kualitas sistem pendidikan tinggi. Melakukan studi kasus di mana mereka menggunakan teknik pembelajaran mesin untuk memprediksi keberhasilan siswa menggunakan fitur yang diambil dari catatan akademik pra-universitas siswa. Hasil eksperimen kami menunjukkan bahwa AUC terbaik dicapai dengan random forest, dari awal semester ketiga pendaftaran kami mendapatkan 0,91 AUC hingga semester lalu, di mana model memberikan AUC 0,97. Empat fitur diperlukan (Semester, Rata-rata SE, Gagal SE, IPK) untuk mencapai akurasi ini. Ini menyiratkan bahwa kursus yang terkait dengan SE memiliki dampak terbesar dalam prediksi putus sekolah.

Sani *et al*. (2020) berkonsentrasi pada prediksi nilai akhir atau Cumulative Grade Point Average (CGPA) mahasiswa dengan memanfaatkan algoritma klasifikasi, dia memprediksi performa mahasiswa yang mengikuti kurus online dan memprediksi siswa tingkat Sekolah Mengah Atas yang berisiko tidak lulus tepat waktu. Seperti yang informasi statistik Pendidikan Tinggi, dari tahun 2011 hingga 2015, jumlah siswa yang masuk untuk program sarjana di Universitas Negeri Malaysia adalah lebih dari 85.000 siswa setiap tahunnya. Beban keuangan yang harus dibayar keluarga akan bertambah karena

pendidikan siswa jika mereka tidak lulus sekolah. Dalam penelitian ini. Peneliti mengkomparasikan tiga model tersebut menunjukkan bahwa model yaitu *Decision Tree*, *Neural Network* dan *Random Forest*. Namun demikian, ada perbedaan yang signifikan secara statistik dalam kinerja antara metode itu, namun tidak ada perbedaan yang mencolok secara statistik antara model *Random Forest* dan model Artificial Neural Network. Metode Neural Network merupakan model terbaik dalam memprediksi jumlah siswa putus sekolah pada penelitian ini dengan akurasi 95.93%. Penelitian ini memiliki tujuan untuk melakukan membandingkan model *Mechine learning* dalam memprediksi mahasiswa, khususnya pada program sarjana di Universitas Negeri Malaysia. Penelitian ini sangat penting karena dilakukan untuk memprediksi sedini mungkin untuk mencegah siswa meninggalkan studi Pendidikan merek. Metode klasifikasi memainkan peran penting dalam prediksi dunia pendidikan, terutama dalam memprediksi kinerja akademik siswa, baik di sekolah maupun perguruan tinggi. Hasilnya dapat membantu guru untuk mengidentifikasi siswa yang berisiko putus sekolah dan memeriksa kesejahteraan mereka. Model dengan akurasi yang cukup tinggi akan digunakan dalam sistem untuk mencegah secara dini sebagai upaya mendeteksi siswa yang berisiko putus sekolah

Menurut Hegde *et al*. (2016) memprediksi putus sekolah mahasiswa sarjana merupakan tantangan besar dalam sistem pendidikan. Penemuan pengetahuan tersembunyi dapat digunakan untuk perencanaan akademik yang lebih baik dan prediksi awal siswa putus sekolah. Keputusan siswa untuk keluar dari program

akan mempengaruhi organisasi, masyarakat dan pembangunan bangsa. Meskipun banyak fakultas tertarik untuk memotivasi mahasiswanya untuk mencapai tujuan kelulusan mereka, disebabkan volume penerimaan mahasiswa yang sangat besar, fakultas tidak dapat menjangkau masing-masing mahasiswa. Jadi mengenal karakter mahasiswa dalam keberagaman adalah sebuah tantangan. Mahasiswa yang dropout dapat dideteksi sebelum bertindak, dengan menerapkan teknik klasifikasi Educational Data Mining. Dalam penelitiannya, berbagai teknik penyeimbangan data telah digunakan untuk meningkatkan akurasi prediksi dalam kelas dengan mempertahankan performa klasifikasi keseluruhan. Dalam penelitian ini data diambil dengan melakukan survei mahasiswa sarjana. Kuesioner survei disiapkan secara terpisah untuk mahasiswa S1 semester I dan Semester III. Kuisioner S1 Semester III berfokus pada emosional, CGPA Semester I dan Semester II, status kehadiran, tindakan kedisiplinan, dan jumlah mata kuliah yang gagal, latar belakang dan hubungan keluarga, akses media sosial dan kemampuan individu mahasiswa dianggap sebagai atribut utama dalam kumpulan data. Teknik *preprocessing* data memungkinkan konversi set data asli ke format algoritma Mechine Learnig yang dapat digunakan ketika data dikumpulkan dari metode survei. Metode dengan akurasi yang tinggi menggunakan metode SVM untuk klasifikasi sebesar 90,24%.

Penelitiannya Hoe *et al*. (2013) membahas tentang teknik data mining yang digunakan untuk mengidentifikasi variabel signifikan untuk mempengaruhi mahasiswa. Masalah siswa putus sekolah karena kinerja akademik yang buruk

pada awal tahun pertama pendaftaran di universitas. Data demografi siswa dan prestasi akademik masa lalu kemudian digunakan untuk mempelajari pola akademik. Pemodelan data dan alat penambangan yang digunakan untuk mengidentifikasi korelasi paling signifikan dari variabel yang terkait dengan keberhasilan akademik berdasarkan data kinerja siswa sepuluh tahun terakhir. Penelitian ini dilakukan di Sekolah Tinggi Teknologi Informasi, Universitas Tenaga Nasional. Mahasiswa yang putus sekolah akibat prestasi akademik yang buruk dapat terjadi pada awal tahun pertama pendaftaran. Seberapa akurat prediksi mereka terkait dengan kinerja aktual siswa sepanjang kehidupan akademis mereka. Penelitian melihat secara detail perspektif data mining dalam memprediksi prestasi akademik siswa. Ini adalah penambangan data yang digunakan sebagai teknik dalam membantu manusia untuk mengekstraksi informasi (pengetahuan) yang berguna. Beberapa penelitian telah dilakukan untuk memprediksi kinerja siswa berdasarkan data kinerja masa lalu. Membandingkan dua teknik data mining yaitu Jaringan Syaraf Tiruan (JST) dan kombinasi teknik klasifikasi *Clustering* dan *Decision Tree* untuk memprediksi dan mengklasifikasikan kinerja akademik siswa. Meskipun tidak ada hasil yang dipublikasikan, penelitian mereka mencoba untuk mengidentifikasi teknik prediksi dan klasifikasi yang terbaik dan akurat. Metode yang digunakan adalah Algoritma CHAID dan Hasilnya menunjukkan akurasi 70,17 %

Penelitian ini dilakukan Utari *et al*. (2020) dan hasil analisis studi kasus pada pendidikan, dengan menganalisis data menggunakan teknik data mining.

Penulis menggunakan metode klasifikasi, yang berfokus pada prediksi dropout mahasiswa sarjana dan diploma. Menganalisis data dalam jumlah besar adalah salah satu tugas yang paling sulit bagi kebanyakan orang. Reputasi universitas diukur berdasarkan persentase mahasiswa pascasarjana dan bagaimana strategi universitas untuk menjaga siswa dari dropout. Pada penelitian ini dipilih metode *Random Forest* yang termasuk supervised learning untuk digunakan dalam proses pembentukan model prediksi dropout. Dataset yang digunakan untuk membangun model dalam penelitian ini adalah dataset balanced, sehingga permasalahan kelas balanced akan diselesaikan dengan metode oversampling. Ada 32 atribut dalam dataset yang akan digunakan. Kegiatan ini dilakukan berulang-ulang karena banyak data duplikat. Selanjutnya adalah membangun model klasifikasi dropout dengan menggunakan teknik klasifikasi penambangan data yang disertai dengan metode *Random Forest* dan teknik dataset balanced yaitu SMOTE. Model klasifikasi dibangun menggunakan data training yang selanjutnya akan digunakan sebagai pembelajaran pada data testing yang telah ditentukan sedemikian rupa pada bagian sebelumnya. Sebagai hasil penelitian, Random Forest disertai dengan SMOTE dapat memberikan hasil yang terbaik dengan akurasi sebesar 93,43%. Hasil utama penelitian ini juga dapat digunakan untuk mengurangi tingkat dropout dengan memprediksi potensi siswa putus sekolah dan mengidentifikasi faktor-faktor terkait

Dalam penelitian ini Timaran *et al*. (2017) mengidentifikasi pola putus sekolah siswa dari sosial ekonomi, akademik, disiplin dan kelembagaan.

Dataset yang digunakan adalah data siswa dari program sarjana di Universitas Narino dari Pasto kota (Kolombia) dengan menggunakan teknik data mining. Data diambil dari periode semester pertama tahun 2004 dan yang semester kedua tahun 2006 lengkap dianalisis Tiga kelompok dengan sebuah periode pengamatan enam tahun sampai dengan tahun 2011. Sosial ekonomi dan profil putus sekolah mahasiswa akademik ditemukan menggunakan teknik klasifikasi berdasarkan Decision Tree. Dalam penelitian ini metode *Decision Tree* memprediksi dengan menghasilkan akurasi yang tinggi yaitu 80%.

Penelitian Harwati *et al*. (2016) menjelaskan bahwa salah satu teknik data mining adalah teknik klasifikasi yang merupakan teknik pembelajaran untuk memprediksi nilai. Beberapa teknik yang diujicoba dalam klasifikasi dan estimasi juga dapat digunakan secara tepat untuk prediksi. Data dengan distribusi non-linier biasanya dapat didefinisikan sebagai fungsi yang memetakan fitur data dari dimensi awal (rendah) ke fitur lain dengan dimensi yang lebih tinggi (bahkan jauh lebih tinggi). Namun, banyak siswa yang putus sekolah karena tidak dapat menyelesaikan studinya tepat waktu. Variabel yang mempengaruhi kasus dropout belum diteliti. Tujuan dari penelitian ini adalah untuk menemukan tingkat akurasi yang paling tinggi di antara kedua metode yang digunakan, yaitu algoritma *Naïve Bayes* dan *Support Vector Machines*. Metode dengan akurasi tertinggi akan diketahui dari bentuk pola dan parameter setiap atribut yang paling berpengaruh terhadap lama studi mahasiswa. Hasil penelitian menunjukkan bahwa metode dengan akurasi tertinggi adalah Algoritma *Naïve Bayes* dengan tingkat akurasi 80,67%. Pembahasan makalah

ini difokuskan pada variabel-variabel yang mempengaruhi masa studi mahasiswa.

Fernandez *et al*. (2021) berpandangan bahwa kesulitan mengakses data pribadi dan masalah privasi yang ditimbulkannya, memaksa institusi untuk mengandalkan data akademik siswa. Pada akhirnya mereka membuat sistem prediksi yang akurat dan andal. Dengan demikian, Rekayasa Fitur dan Rekayasa Instans seperti menangani redundans, signifikansi fitur, korelasi, fitur kardinalitas, nilai yang hilang, pembuatan atau penghapusan fitur, fusi data, penghapusan instance yang tidak berguna, binning, resampling, normalisasi, atau pengkodean diterapkan secara detail sebelum pembangunan model terkenal seperti Gradient Boosting, Random Forest, dan Support Vector Machine beserta Ensemble-nya pada tahapan yang berbeda: sebelum pendaftaran, pada akhir semester pertama, pada akhir semester kedua, akhir semester ketiga, dan akhir semester keempat. Pendekatan berbeda muncul ketika *mechine learning* digunakan untuk membangun model prediktif yang mengantisipasi siswa putus sekolah sesuai dengan jenis data yang digunakan . Ini berarti bekerja keras pada *preprocessing* data, Rekayasa Fitur, Rekayasa *Instance*, dan pembuatan model, serta validasi yang tepat untuk mendapatkan model prediksi yang akurat dan andal yang berfungsi sebagai sistem pengambilan keputusan, oleh sebab itu kontribusi utama dari makalah ini terletak pada kemungkinan menentukan probabilitas putus sekolah siswa pada berbagai tahap merupakan kunci pendidikan tinggi mulai dari bagian paling awal sebelum pendaftaran di universitas hingga semester keempat. Melalui

berbagai model prediksi yang saling berhubungan. Makan hasil evaluasinya menunjukkan bahwa algoritma *Support Machine Vector* memiliki akurasi paling tinggi sebesar 90,87% dari lainnya.

Penelitian ini menggunakan algoritma *Extreme Learning Machine,* metode pembelajaran Neural Network dan algoritma *Support Vector Machine* untuk perbandingan tingkat akurasi menggunakan data yang sama. Oleh karena itu, perlu untuk mempelajari atau memprediksi putus sekolah siswa sehingga dapat digunakan sebagai informasi yang berguna untuk memperkirakan tingkat putus sekolah siswa di tahun mendatang serta mampu mengurangi tingkat putus sekolah siswa. Menurut Sa'ad *et al* (2020) Beberapa algoritma klasifikasi penambangan data telah digunakan untuk memprediksi perilaku siswa yang berpotensi dropout termasuk Decision Tree, Naural Network, Naive Bayes, Learning berbasis instance, Regression Logistik, Support Vector Machine, K-Nearest Neighbor, Rapidminer, dan Extreme Learning Machines. Metode yang digunakan untuk memprediksi penelitian yang akan dilakukan adalah dengan menggunakan Extreme Learning Machine (ELM) dengan akurasi 72 % lebih unggul dari pada Support Vector Machine yang hanya dapat akurasi 63%.

2.2 Kerangka Teori

Data pada tabel 2.1 menunjukkan hasil studi literatur beberapa jurnal. Maka dapat ditemukan metode yang memiliki akurasi tinggi yang nantinya akan dibuat metode pada pembahasan Bab 3. Terdapat banyak metode yang diperoleh hanya saja akan diambil dua metode untuk dibuat penelitian selanjutnya

Tabel 2. 1 Kerangka Teori

Input	Proses	Output
• Examinations • Average rades Social • Students Attendance In Class • Family Income • Mother's education • Mother's age • Cost • Academic • Disciplinary Issues • Health Issue	Random Forest, Accuracy = 95.00%	Student's Dropout
	Random Forest, Accuracy = 97.00%	
	Neural Network, Accuracy= 95,86%.	
	Support Machine Vector, Accuracy= 90.24%	
	CHAID, Accuracy=70.17%	
	Random Forest, Accuracy= 93.43%	
	Decision Tree, Accuracy = 80.00 %	
	Naive Bayes, Accuracy = 80,67%	
	Support Machine Vector, Accuracy = 90,87%	
	Extreme Learning Machine, Accuracy = 72%	

Hasil Studi literarur terlihat pada *Theoretical Framework* sudah bisa dilihat diantara banyaknya metode yang muncul akan diseleksi satu metode saja untuk dibuat rujukan dalam penelitian selanjutnya. Untuk memprediksi siswa putus sekolah terdapat input yang begitu beragam dan metode yang juga beragam, namun dari banyaknya metode yang ditemukan metode *Random Forest* yang akurasinya paling tinggi.

Tabel 2. 2 Daftar Jurnal Penelitian Terdahulu

No	Nama Penulis & Tahun	Metode	Judul Jurnal	Akurasi
1	Kemper *et al*. (2020)	Random Forest	Predicting Student Dropout: A Machine Learning Approach." *European Journal of Higher Education*	95%
2	Perez *et al.* (2018)	Random Forest	Predicting Student Drop-Out Rates Using Data Mining Techniques: A Case Study	97%
3	Sani *et al.* (2020)	Neural Network	Drop-Out Prediction in Higher Education Among B40 Students	95,86%.
4	Hegde (2016)	SVM	Dimensionality Reduction Technique for Developing Undergraduate Student Dropout Model Using Principal Component Analysis through R Package	90.24%
5	Hoe *et al.* (2013)	CHAID	Analyzing Students Records to Identify Patterns of Students' Performance	70.17%

6	Utari *et al.* (2020)	Random Forest	Implementation of Data Mining for Drop-Out Prediction Using Random Forest Method	93.43%
7	Timaran *et al* (2017)	Decision Tree	Application of Decision Trees for Detection of Student Dropout Profiles	80.00 %
8	Harwati *et al.* (2016)	Naive Bayes	Drop out Estimation Students Based on the Study Period: Comparisonbetween Naïve Bayes and Support Vector Machines Algorithm Methods.	80,67%
9	Fernandez *et al*. (2021)	Support Machine Vector	A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data	90,87%
10	Sa'ad *et al*. (2020)	Extreme Learning Machine (ELM)	Student Prediction of Drop Out Using Extreme Learning Machine (ELM) Algorithm	72%

Berdasarkan rangkuman pada tabel 2.2 hasil penelitian sebelumnya, terlihat bahwa berbagai metode *machine learning* telah digunakan dalam memprediksi risiko putus sekolah, dengan performa yang bervariasi bergantung pada karakteristik *dataset*. Metode *Random Forest* muncul sebagai salah satu teknik dengan performa paling konsisten dan akurat. Kemper et al. (2020) memperoleh akurasi 95%, Perez et al. (2018) mencapai 97%, dan Utari et al. (2020) melaporkan akurasi 93,43%, menguatkan bahwa *Random Forest* memiliki kemampuan yang sangat baik dalam menangani data beragam, fitur kompleks, dan struktur non-linear.

Metode lain seperti Neural Network (Sani et al., 2020) dengan akurasi 95,86% dan Support *Vector Machine* (SVM) Hegde, (2016), Fernandez et al., (2021) dengan akurasi 90,24% dan 90,87% juga menunjukkan performa tinggi, namun cenderung membutuhkan tuning parameter yang lebih kompleks serta sensitif terhadap skala dan distribusi data. Sementara itu, metode yang lebih sederhana seperti CHAID Hoe et al., (2013) dan *Decision Tree* Timaran et al. (2017) menunjukkan akurasi lebih rendah, masing-masing 70,17% dan 80,00%, menandakan keterbatasannya dalam menangkap pola data yang lebih rumit. Hal serupa juga terlihat pada metode *Naive Bayes* Harwati et al., (2016) dan *Extreme Learning Machine* (ELM) Sa'ad et al., (2020) yang mencatat akurasi 80,67% dan 72%, sehingga kurang kompetitif dibanding algoritma berbasis *ensemble*.

Secara keseluruhan, data ini menunjukkan bahwa Random Forest merupakan metode yang paling unggul dan stabil dalam memprediksi risiko

putus sekolah karena mampu mengatasi *overfitting*, bekerja optimal pada *dataset* dengan banyak variabel, serta memberikan hasil yang konsisten pada berbagai penelitian. Temuan ini mendukung pemilihan *Random Forest* sebagai metode utama dalam penelitian untuk memprediksi santri berisiko putus sekolah.

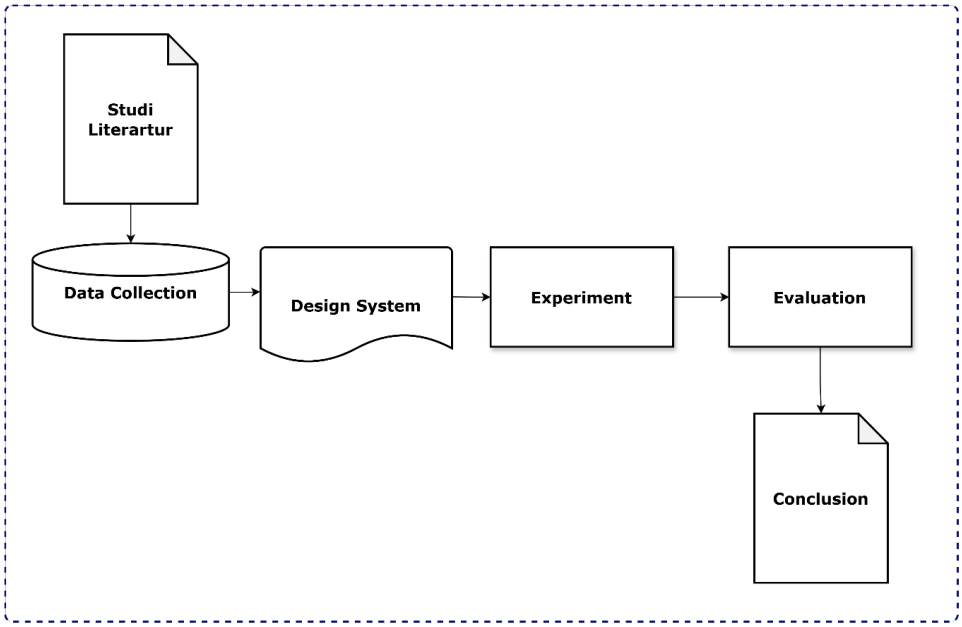
BAB III

METODE PENELITIAN

Bab ini menjelaskan langkah-langkah yang akan dilakukan dalam penelitian ini. Tahapan yang disebutkan meliputi tahap pengumpulan data, pra-pemrosesan data, pembuatan model klasifikasi dan uji kasus yang akan diimplementasikan dalam penelitian.

3.1 Prosedur Penelitian

Langkah-langkah atau prosedur dalam penelitian ini dipaparkan pada gambaran umum penelitian seperti pada diagram di bawah ini:



Gambar 3. 1 Prosedur Penelitian

Pada gambar 3.1 Prosedur penelitian mepaparkan desain penelitian, langkap pertama *Studi Literatur* untuk mengumpulkan referensi dalam memahami konteks dan teori dasar penelitian, kemudian *Data Collection*,

untuk analisis atau eksperimen dilanjut *Desain Sistem* sebagai rancangan sistem atau model yang akan diuji serta *Eksperimen* untuk menguji dan melihat kinerjanya dan *Implementasi*, menerapkan hasil eksperimen sedangkan proses terakhir adalah diskusi & kesimpulan untuk menganalisis hasil dan menarik kesimpulan untuk penelitian lebih lanjut.

3.2 Pengumpulan Data

Tahapan dalam metode pengumpulan data pada penelitian ini menggunakan beberapa atribut yang relevan dengan permasalahan yang diteliti. Data yang digunakan dalam penelitian ini merupakan data primer. Pengumpulan data primer diperoleh dari hasil pengumpulan data (data collecting) pada bagian data center Pondok Pesantren Sidogiri, berupa data santri yang mengalami putus sekolah.

Penggunaan data primer memungkinkan peneliti untuk memperoleh data yang bersifat faktual dan akurat sehingga dapat digunakan untuk mendukung penelitian ini secara langsung.

Setelah data dikumpulkan, peneliti melakukan proses normalisasi data untuk menjadikan data lebih konsisten dan terstruktur. Proses normalisasi bertujuan untuk memastikan bahwa data yang digunakan tidak mengandung redundansi (pengulangan data) dan inkonsistensi, serta memastikan dataset tersusun dalam bentuk tabel yang efisien dan meminimalkan duplikasi data.

Berdasarkan hasil pengumpulan dan pengolahan data tersebut, peneliti memperoleh sebanyak 1.763 data santri putus sekolah yang berasal dari berbagai kelas dan jenjang pendidikan. Data tersebut memiliki latar belakang

pendidikan dan sosial yang beragam, dengan rentang usia mulai dari 9 tahun hingga 35 tahun, serta berasal dari berbagai kondisi ekonomi, baik dari kalangan tidak mampu maupun kalangan mampu.

Dataset yang telah diperoleh selanjutnya digunakan sebagai objek penelitian, dengan format dan atribut data sebagaimana ditampilkan pada tabel berikut.

Tabel 3. 1 Daftar Atribut Dataset Santri Putus sekolah

No	Atribut	Tipe Data	Keterangan
1	no. urut	Numeric	≥ 0
2	no. registrasi	Numeric	≥ 0
3	alamat	Text	Character
4	asrama santri	Text	Character
5	interaksi teman asrama	Numeric	≥ 0
6	tingkatan	Numeric	≥ 0
7	kelas	Numeric	≥ 0
8	interaksi teman kelas	Numeric	≥ 0
9	taanggal registrasi	Numeric	09/9/999
10	umur	Numeric	≥ 0
11	jumlah anggota keluarga	Numeric	≥ 0
12	pendidikan ayah	Numeric	≥ 0
13	pekerjaan ayah	Numeric	≥ 0
14	pendapatan ayah	Numeric	≥ 0
15	pendidikan Ibu	Numeric	≥ 0
16	pekerjaan ibu	Numeric	≥ 0
17	pendapatan ibu	Numeric	≥ 0
18	pembayaran santri	Numeric	≥ 0
19	jarak rumah	Numeric	≥ 0
20	lama mondok	Numeric	≥ 0
21	gagal kelas	Numeric	≥ 0
22	kesehatan	Numeric	≥ 0
23	kehadiran	Numeric	≥ 0
24	jam belajar	Numeric	≥ 0
25	tempat tinggal	Text	Character
26	perguruan Tinggi	Text	Character
27	status	Text	Lulus, Putus Sekolah

Tabel 3.1 memaparkan atribut dengan detail sedangkan pengkategorian atribut dataset terdapat variabel Independen dan Dependen diuraikan pada tabel 3.2

3.2.1. Data Engineering

Tahap Data Engineering pada penelitian ini dilakukan dengan menerapkan serangkaian transformasi teknis untuk mengoptimalkan dataset yang terdiri dari 26 atribut independen dan satu variabel dependen. Proses diawali dengan reduksi dimensi melalui penghapusan atribut identitas seperti no. urut, no. registrasi, dan date registrasi yang tidak memiliki nilai korelasi terhadap target prediksi, disusul dengan imputasi data untuk menangani nilai kosong pada fitur sensitif seperti pendapatan dan pekerjaan orang tua. Selanjutnya, dilakukan transformasi fitur kategorikal menggunakan teknik Encoding untuk mengubah data tekstual seperti alamat dan asrama santri menjadi format numerik, serta penerapan skalasi fitur pada variabel kontinu seperti umur dan jarak rumah guna menyeimbangkan rentang nilai antar atribut. Seluruh rangkaian proses ini diakhiri dengan pembagian dataset untuk memastikan model Random Forest mendapatkan asupan data latih yang komprehensif sebelum dilakukan pengujian pada data uji.

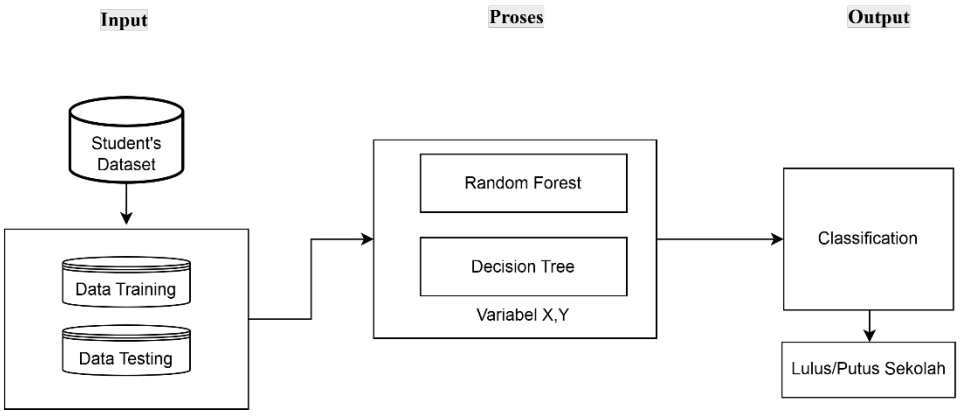
Tabel 3. 2 Variabel Independen dan Dependen

No	Atribut	Tipe Variabel
1	asrama santri	Independent
2	interaksi teman asrama	Independent
3	tingkatan	Independent
4	kelas	Independent
5	interaksi teman kelas	Independent
6	umur	Independent
7	jumlah anggota keluarga	Independent
8	pendidikan ayah	Independent
9	pekerjaan ayah	Independent
10	pendapatan ayah	Independent
11	pendidikan ibu	Independent
12	pekerjaan ibu	Independent
13	pendapatan ibu	Independent
14	pembayaran santri	Independent
15	jarak rumah	Independent
16	lama mondok	Independent
17	gagal kelas	Independent
18	kesehatan	Independent
19	kehadiran	Independent
20	jam belajar	Independent
21	tempat tinggal	Independent
22	perguruan tinggi	Independent
23	status	Dependent

Pada tebal 3.2 setelah melalui tahap seleksi awal, dataset direduksi menjadi 23 atribut yang memiliki korelasi fungsional terhadap variabel dependen. Atribut yang bersifat identitas dihapus untuk menghindari bias. Dataset akhir ini kemudian diklasifikasikan ke dalam tiga kategori utama: profil demografi, indikator ekonomi keluarga, dan performa perilaku santri, yang selanjutnya diproses melalui teknik *encoding* dan *feature extraction* untuk menghasilkan model prediksi yang akurat

3.3 Desain Sistem

Pada tahapan desain sistem secara rinci membahas bagaimana dataset santri putus sekolah (Dropout) dikelola, dilatih dan ujicoba serta mengavaluasi metode dengan beberapa langkah seperti pada gambar 3.2.



Gambar 3. 2 Desain Sistem

Pada gambar 3.2 Dataset santri dimasukkan sebagai input ke dalam sistem. Dataset kemudian akan diproses dan dibagi menjadi dua bagian, yaitu data pelatihan (*training set*) dan data pengujian (*testing set*). Setelah pembagian data, variabel-variabel dalam dataset diidentifikasi sebagai variabel X dan Y. variabel X untuk prediksi sedangkan Y sebagai target atau label yang ingin diprediksi oleh model. Proses Pemodelan dilakukan dengan model *Random Forest* dan *Decision Tree*. Hasil prediksi akan dievaluasi kembali menggunakan metrik dan evaluasi model untuk memastikan bahwa model bekerja dengan baik. Berdasarkan hasil evaluasi, sistem akan menghasilkan keputusan akhir, yaitu Putus sekolah yang berarti hasil akhir dari klasifikasi atau prediksi. Terdapat dua pilihan keputusan: “Putus sekolah” atau “Lulus”.

Dataset berisi data santri yang mencakup berbagai fitur (variabel X) seperti usia, latar belakang ekonomi, nilai akademik, kehadiran, dll., serta label target (Y), misalnya: *lulus* atau *putus sekolah*.

3.4 Implementasi Sistem

Implementasi model klasifikasi dalam penelitian ini dilakukan menggunakan bahasa pemrograman Python melalui platform Google *Colaboratory* (Colab). Pemilihan Google Colab didasarkan pada ketersediaan sumber daya komputasi berbasis *cloud* yang memungkinkan proses eksekusi algoritma dilakukan secara efisien tanpa terkendala spesifikasi perangkat keras lokal. Selain itu, platform ini memudahkan integrasi dengan berbagai pustaka open-source yang mendukung pengolahan data dan machine learning.

Proses implementasi perangkat lunak dibagi menjadi beberapa tahapan teknis sebagai berikut:

Environment sistem dikembangkan dengan memanfaatkan pustaka utama Python, antara lain:

1. Pandas & NumPy: Digunakan untuk manipulasi struktur data tabel dan operasi matriks pada dataset santri.
2. Scikit-Learn (Sklearn): Digunakan sebagai pustaka utama untuk implementasi algoritma *Random Forest* dan *Decison Tree*, proses *data splitting*, hingga kalkulasi metrik evaluasi.
3. Matplotlib & Seaborn: Digunakan untuk visualisasi data, termasuk pembuatan grafik *Confusion Matrix* dan kurva ROC.

3.5 Eksprerimen

Tahap uji coba dalam penelitian ini dilakukan untuk mengevaluasi efektivitas metode klasifikasi yang diterapkan serta memastikan bahwa model yang dibangun memiliki tingkat akurasi yang tinggi dalam mendeteksi target prediksi. Proses eksperimen ini difokuskan pada pengukuran performa model melalui metrik evaluasi yang komprehensif, meliputi nilai akurasi, *recall*, dan presisi. Melalui pengujian ini, algoritma dilatih untuk mengenali pola dalam dataset sehingga mampu memberikan hasil prediksi yang optimal sesuai dengan tujuan penelitian.

Dalam penelitian ini, peneliti menerapkan studi komparatif dengan membandingkan dua algoritma berbasis pohon keputusan, yaitu *Random Forest* dan *Decision Tree*. Untuk mendapatkan hasil pengujian yang objektif dan konsisten, eksperimen dilakukan dengan membagi dataset ke dalam empat kategori rasio data latih (*training set*) dan data uji (*testing set*) sebagai berikut:

Tabel 3. 3 Pembagian Rasio Dataset

Metode 1	**Metode 2**	**Skenario**	**Rasio**	**Data Latih**	**Data Uji**
Random Forest	Decision Tree	1	90:10	1586	178
		2	80:20	1411	353
		3	70:30	1234	530
		4	60:40	1058	706

Masing-masing data percobaann data training pada tabel 3.3 memiliki jumlah yang berbeda yang mempengaruhi prediksinya, ujicoba data training pertama, 60 data yang disiapkan 1058 dari 1763 data yang ditraining, kemudian uji coba kedua dengan 70 data lahit dan data yang tersedia 1234 dari dari

dataset 1763, data kategori, 80 dengan jumlah data 1411 dari total seluruh dataset training, dan data kategori latih level terakhir, 90 dengan jumlah data 1586, persentase data training tentunya mempengaruhi hasil prediksinya, karna pengunnaan jumlah data latih juga berefek pada hasilnya, rincian hasil ujicoba data latih metode Random Forest sebagai berikut.

3.4.1 Model Random Forest

Langkah memulai Metode Random Forest dengan melatih atau membuat pohon tree dari dataset yang sudah memiliki kelas, pohon-pohon inilah yang akan menjadi bagian dari Random Forest yang akan digunakan untuk tahap klasifikasi, selanjutnya algoritma akan memilih sampel acak dari dataset yang telah ada yang akan digunakan pada tahap identifikasi, data yang telah dipilih diklasifikasikan oleh seluruh pohon yang telah dibuat, kemudian akan didapatkan hasil predikasi dari setiap Decision Tree sehingga membentuk pohon keputusan pada metode Random Forest.

Dataset santri sebagai input ke dalam sistem. Dataset ini akan diolah untuk membentuk data pelatihan (*training set*) dan data uji (*test set*), Random Forest menggunakan teknik bootstrap sampling pada data pelatihan. Dalam bootstrap sampling, beberapa subset data diambil secara acak dari training set dan setiap subset ini akan menjadi data pelatihan yang berbeda untuk masing-masing decision tree. *voting*, hasil dari setiap *decision tree* dikumpulkan, dan keputusan akhir diambil berdasarkan mayoritas sebgai prediksi

Pohon keputusan dimulai dengan cara menghitung nilai entropy setiap Atribut sebagai penentu tingkat ketidakmurnian atribut dan nilai information gain. Untuk menghitung nilai entropy digunakan rumus seperti pada persamaan 1, sedangkan nilai information gain menggunakan persamaan 2 (K. Schouten, 2016)

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \quad (3.1)$$

Dimana S adalah himpunan kasus dan p(i|S) merupakan proporsi nilai S terhadap kelas i.

$$\text{Information Gain}(S, A) = E(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} E(S_v) \quad (3.2)$$

Dimana Values (t) merupakan semua nilai yang mungkin dalam himpunan kasus t. Sv adalah subkelas dari S dengan kelas v yang berhubungan dengan kelas t. St adalah semua nilai yang sesuai dengan t

3.4.2 Model Decision Tree

Sebagai pemanding proses pemodelan prediksi santri putus sekolah peneliti menggunakan *Decision Tree*. Prosedur dalam metode *Decision Tree* diawali dengan mengalkulasi nilai *Entropy* total. Langkah tersebut diikuti dengan penghitungan *Entropy* pada masing-masing atribut. Setelah parameter tersebut diperoleh, tahap selanjutnya adalah menentukan nilai *Information Gain* untuk setiap atribut tersebut, tahap ini terus dilakukan berulang kali sampai semua atribut dilakukan dan membentuk beberapa node pohon keputusan. Dimana hasil perhitungan dapat dilihat pada rumus berikut:

$$\text{Entropy (S)} = \sum_{i=1}^{n} n - pi * log2pi \quad (3.3)$$

$$\text{Gain(S,A)} = \text{Entropy (S)} \sum_{i=1}^{n} - \frac{|Si|}{|Si|} * \text{Entropy } (S_i) \quad (3.4)$$

Keterangan:

S : Himpunan kasus

n : Jumlahh partisi S

Pi : Proporsi dari Si terhadap S

A : Atribut

n : Jumlah partisi atribut A

|Si| : Jumlah kasus pada partisi ke-i

|S| : Jumlah kasus dalam S

Pada persiapan awal ditentukan atribut yang digunakan kemudian melakukan uji atribut dengan mencari nilai Gain tertinggi berdasarkan perhitungan entropy dari masing-masing atribut. Apabila ditemukan gain tertinggi maka gain tersebut akan menjadi root awal. Selanjutnya dilakukan penentuan cabang dengan cara yang sama dengan melihat gain tertinggi dari tiap hasil partisi

3.6 Pengukuran Performa Model

Pengukuran performa merupakan tahapan krusial dalam siklus pengembangan model *machine learning* untuk mengevaluasi kualitas, efektivitas, serta tingkat keberhasilan model klasifikasi dalam memprediksi data secara akurat. Evaluasi ini bertujuan untuk memastikan bahwa model tidak hanya memiliki akurasi yang tinggi pada data *training*, tetapi juga memiliki kemampuan kemampuan beradaptasi yang baik saat dihadapkan pada data baru (*unseen data*). Pada penelitian ini, evaluasi dilakukan secara komprehensif

menggunakan dua pendekatan utama, yaitu *Confusion Matrix* dan *Area Under the Curve - Receiver Operating Characteristic*

3.6.1 *Confusion Matrix*

Confusion Matrix merupakan instrumen evaluasi berbentuk tabel kontingensi yang digunakan untuk memvisualisasikan kinerja algoritma klasifikasi dengan membandingkan nilai aktual dari dataset terhadap nilai hasil prediksi model. Matriks ini menyediakan informasi detail mengenai sebaran prediksi yang dilakukan oleh model ke dalam empat kategori utama:

Tabel 3. 4 *Confusion Matrix*

Prediksi	Aktual	
	Positif	Positif
Positif	TP	FP
Negatif	FN	TN

Berdasarkan tabel 3.4 *Confusion Matrix*, beberapa metrik performa dapat dihitung. Akurasi model, yang merupakan perbandingan antara prediksi yang benar dengan total prediksi

4. True Positive (TP): Merupakan jumlah data positif yang berhasil diprediksi secara tepat oleh model sesuai dengan kelas aktualnya.
5. True Negative (TN): Merupakan jumlah data negatif yang berhasil diprediksi secara tepat oleh model sesuai dengan kelas aktualnya.
6. False Positive (FP): Merupakan jumlah data negatif yang salah diprediksi sebagai data positif (dikenal sebagai Kesalahan Tipe I).

7. False Negative (FN): Merupakan jumlah data positif yang salah diprediksi sebagai data negatif (dikenal sebagai Kesalahan Tipe II).

Melalui parameter tersebut, penelitian ini dapat mengukur performa model secara lebih objektif melalui berbagai metrik turunan seperti *Accuracy*, *Precision*, *Recall* (Sensitivitas), dan *F1-Score*. Hal ini sangat penting untuk mendeteksi apakah model memiliki kecenderungan bias terhadap salah satu kelas tertentu seperti penjelasn tabel 3.5

Tabel 3. 5 Rumus Evaluasi Performa Metode

Performance Metric	*Rumus*
Akurasi	$\frac{TP + TN}{TP + TN + FP + FN} \times 100$
Recall	$\frac{TP}{TP + FN} \times 100$
Presisi	$\frac{TP}{TP + FP} \times 100$

Berdasarkan empat parameter dasar dalam *Confusion Matrix* (TP, TN, FP, dan FN), dilakukan penghitungan metrik performa guna mendapatkan hasil evaluasi yang komprehensif. Pengukuran tidak hanya berfokus pada Akurasi sebagai nilai global, tetapi juga melibatkan *Recall* untuk melihat kemampuan model mendeteksi kelas target, serta Presisi untuk mengukur tingkat ketepatan prediksi. Selain itu, *F1-Score* digunakan sebagai parameter penyeimbang apabila terdapat distribusi data yang tidak merata. Seluruh metrik ini menjadi tolok ukur utama dalam memvalidasi apakah model *Random Forest* dan *Decision Tree* yang dibangun telah mencapai kinerja optimal pada pembagian dataset

3.6.2 *Area Under the Curve Receiver Operating Characteristic*

Pendekatan kedua adalah analisis menggunakan kurva *Receiver Operating Characteristic* (ROC) dan perhitungan nilai *Area Under the Curve* (AUC). Kurva ROC merupakan representasi grafis yang menunjukkan hubungan dinamis antara *True Positive Rate* (TPR) atau sensitivitas pada sumbu-Y, terhadap *False Positive Rate* (FPR) pada sumbu-X. Visualisasi ini merepresentasikan performa model pada berbagai tingkat ambang batas klasifikasi (classification thresholds).

Metrik *Area Under the Curve* (AUC) digunakan sebagai nilai skalar untuk merangkum seluruh informasi dari kurva ROC menjadi satu indikator kuantitatif yang menunjukkan kemampuan diskriminasi model. Rentang nilai AUC berada di antara 0.0 hingga 1.0, di mana nilai yang mendekati 1.0 mengindikasikan bahwa model memiliki kemampuan yang sangat baik dalam memisahkan kelas positif dan negatif secara presisi. Dalam konteks penelitian ini, penggunaan AUC-ROC menjadi standar validasi untuk membuktikan bahwa konfigurasi parameter n_estimators dan max_depth pada pembagian data telah menghasilkan model dengan performa yang optimal dan reliabel.

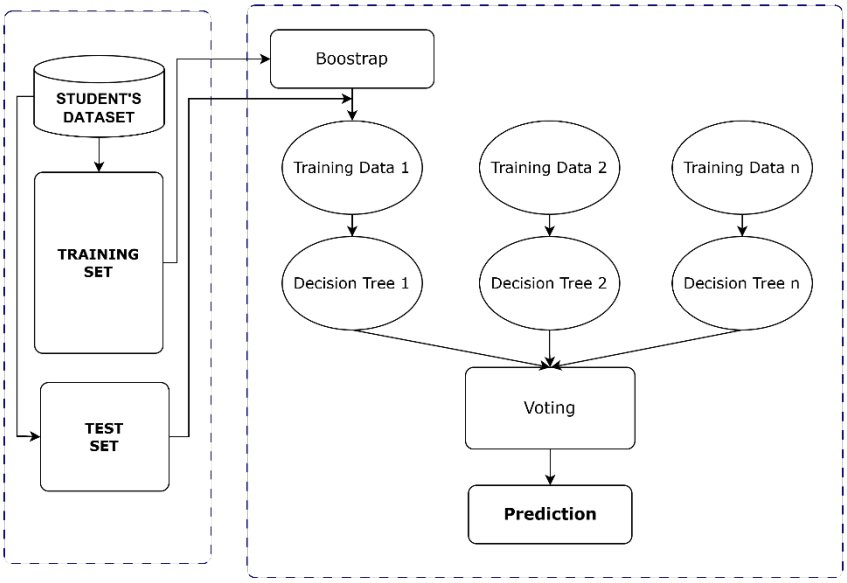
BAB IV

METODE RANDOM FOREST

4.1 Desain Metode

Pada tahap ini, sistem prediksi dirancang untuk mengklasifikasikan status santri di Pondok Pesantren Sidogiri berdasarkan beberapa atribut seperti umur, tingkat pendidikan, pekerjaan orang tua, serta riwayat interaksi di pondok. Model yang digunakan adalah *Random Forest Classifier* karena kemampuannya dalam menangani data kategorikal dan kompleksitas hubungan antar fitur.

Langkah memulai metode *Random Forest* dengan melatih atau membuat pohon tree dari dataset yang sudah memiliki kelas, pohon-pohon inilah yang akan menjadi bagian dari Random Forest yang akan digunakan untuk tahap klasifikasi, selanjutnya algoritma akan memilih sampel acak dari dataset yang telah ada yang akan digunakan pada tahap identifikasi, data yang telah dipilih diklasifikasikan oleh seluruh pohon yang telah dibuat, kemudian akan didapatkan hasil predikasi dari setiap Decision Tree sehingga membentuk pohon keputusan pada metode *Random Forest* seperti pada gambar 4.1.



Gambar 4. 1 Architecture of the Random Forest

Pada gambar 4.1 Dataset santri sebagai input ke dalam sistem. Dataset ini akan diolah untuk membentuk data pelatihan (*training set*) dan data uji (*test set*), Random Forest menggunakan teknik bootstrap sampling pada data pelatihan. Dalam bootstrap sampling, beberapa subset data diambil secara acak dari training set dan setiap subset ini akan menjadi data pelatihan yang berbeda untuk masing-masing decision tree. *voting*, hasil dari setiap *decision tree* dikumpulkan, dan keputusan akhir diambil berdasarkan mayoritas sebgai prediksi

Pohon keputusan dimulai dengan cara menghitung nilai entropy setiap Atribut sebagai penentu tingkat ketidakmurnian atribut dan nilai information gain. Untuk menghitung nilai entropy digunakan rumus seperti pada persamaan 1, sedangkan nilai information gain menggunakan persamaan 2 (K. Schouten, (2016)

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \quad (4.1)$$

$$\text{Information Gain}(S, A) = E(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} E(S_v) \quad (4.2)$$

Dimana S adalah himpunan kasus dan p(i|S) merupakan proporsi nilai S terhadap kelas i.

Dimana Values (t) merupakan semua nilai yang mungkin dalam himpunan kasus t. Sv adalah subkelas dari S dengan kelas v yang berhubungan dengan kelas t. St adalah semua nilai yang sesuai dengan t

Alur proses prediksi santri putus sekolah menggunakan metode Random Forest. Algoritma ini termasuk dalam metode ensemble learning, di mana model terdiri dari sekumpulan pohon keputusan (decision tree) yang dibentuk berdasarkan data pelatihan yang telah dibootstrap.

Taser *et al* (2021) menerapkan pendekatan *bagging* dan *boosting* menggunakan enam algoritma berbasis pohon keputusan untuk prediksi. Hasilnya menunjukkan bahwa pendekatan ensemble (*bagging* dan *boosting*) memberikan akurasi yang lebih tinggi dibandingkan dengan penerapan individual dari algoritma pohon keputusan tersebut.

4.2 Implementas Model Random Forest

Untuk menguji performa metode *Random* Forest dalam memprediksi santri putus sekolah, dilakukan pengujian berdasarkan perbandingan data *training* dan data *testing*, serta variasi jumlah pohon keputusan (*n_estimators*).

Tujuan dari pengujian ini, untuk melihat pengaruh rasio data latih dan data uji, serta jumlah pohon yang digunakan, terhadap akurasi dan kestabilan model.

Menurut Bichri *et al* (2024) melakukan pengujian melalui empat skenario berbeda berdasarkan rasio pembagian data *training* dan data *testing* dapat mengetahui pengaruh variasi jumlah pohon terhadap akurasi dan stabilitas model pada setiap skenario.

Selanjutnya Random Forest dikustomisasi dengan *n_estimators* sebanyak 50, 100,150 dan 500 untuk menghitung jumlah pohon keputusan, kemudian menggunakan *random_state* 42 untuk mengendalikan proses pengacakan. Parameter *n_estimators* pada algoritma *Random Forest* digunakan untuk mengatur jumlah pohon keputusan serta memastikan konsistensi proses pelatihan model. Semakin besar nilai *n_estimators*, jumlah pohon yang dibangun oleh model akan semakin banyak, yang berpotensi meningkatkan akurasi karena model menjadi lebih kompleks dan stabil. Namun, peningkatan ini juga disertai dengan bertambahnya waktu komputasi. Oleh karena itu, variasi nilai ini bertujuan untuk mengamati titik optimal antara akurasi dan efisiensi.

Sementara itu, parameter *random_state* diatur 42 untuk mengevaluasi stabilitas model terhadap variasi pembagian data pelatihan dan pengujian yang bersifat acak. Pengujian kombinasi dengan parameter ini memberikan gambaran yang lebih komprehensif terhadap performa model *Random Forest* dalam memprediksi status santri.

```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.1, random_state=42
)
model = RandomForestClassifier(
    n_estimators=100,
    random_state=42,
    max_depth=5,
    min_samples_split=5,
    min_samples_leaf=2,
    max_features=5,
    oob_score=True
)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

print(classification_report(y_test, y_pred))
```

Implementasi dilakukan menggunakan bahasa pemrograman *Python* dengan *library scikit-learn*. *Dataset* yang dibuat bahan penelitian merupakan data santri yang telah melewati proses *clening* dari data yang bernilai kosong (missing value) dan telah dilakukan *encoding* pada variabel target (status) menjadi dua kelas, yaitu Lulus (0) dan Putus sekolah (1).

Potongan kode berikut menunjukkan tahapan implementasi model Random Forest:

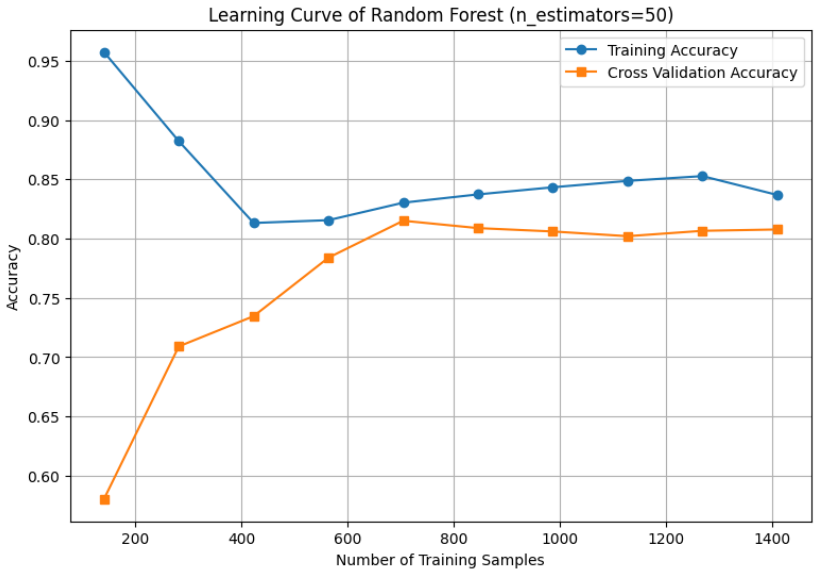
```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.1, random_state=42
)
```

Penggalan kode tersebut menggambarkan penggunaan model *Random Forest Classifier* dari *library scikit-learn* untuk memprediksi status santri, apakah Putus Sekolah atau Lulus. Bagian *test_size*=0.1 disesuaikan berdasarkan rasio data yang berbeda untuk setiap pengujian. Sementara itu, bagian "*random_state*=42" mengacu pada parameter yang digunakan untuk mengatur proses pengacakan dalam pembagian data dan pembangunan pohon keputusan. Parameter ini membantu agar hasil pelatihan model tetap konsisten dan dapat diulang.

4.2.1 Pelatihan Model Skenario 1

Dalam tahap eksperimen prediksi santri putus sekolah di pondok pesantren sidogiri, Pada model Skenario 1, pengujian model *Random Forest* dilakukan dengan membagi dataset menjadi data latih dan data uji, menggunakan rasio 90:10 sekitar 1586 data latih dan 177 data uji. Eksperimen pengujian dengan nilai parameter *n_estimators* 50 serta menetapkan *random_state* 42.



Gambar 4. 2 *Learning Curve* Model Random Forest (n_estimators=50)

Grafik *Lerning Curve* pada Gambar 4.2 menjelaskan performa algoritma *Random Forest* pelatihan model Skenario 1 dengan *n_estimators*=50 mulai dari jumlah sampel kecil hingga sekitar 1.400 lebih data latih. Seiring bertambahnya jumlah data latih, *Training Accuracy* cenderung menurun dan stabil di angka 0,85, sementara *Cross Validation Accuracy* meningkat dan stabil di angka 0,81.

Kedua kurva menunjukkan tanda pendekatan nilai setelah melewati 700 sampel, yang mengindikasikan bahwa penambahan data hingga 1.586 akan membuat model semakin stabil dan mengurangi risiko overfitting. Model pada pelatihan skenario 1 dengan total data latih tersebut, model memiliki performa yang cukup baik dengan akurasi di atas 80% dalam menentukan santri “lulus” dan “putus sekolah” di pondok pesantren sidogiri.

Tabel 4. 1 Variabel berpengaruh pada Model skenario 1

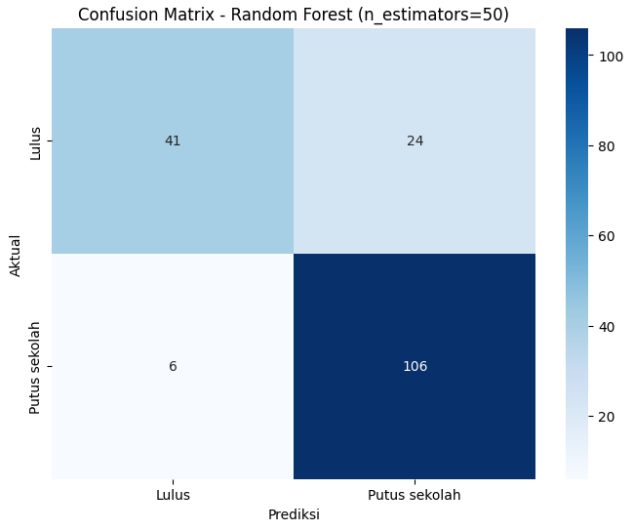
No	Variabel	Nilai
1	Kelas	0.199
2	Kehadiran	0.127
3	Umur	0.126
4	Study time	0.106
5	Gagal kelas	0.099
6	Tingkat	0.071
7	Interaksi teman asrama	0.061
8	Interaksi teman kelas	0.045
9	Jam belajar	0.027
10	Kesehatan	0.026

Hasil analisis feature importance pada Tabel 4.1 menunjukkan bahwa fitur kelas memiliki pengaruh terbesar dalam menentukan prediksi status santri, dengan nilai kontribusi sebesar 0.199. Faktor berikutnya yang juga berperan signifikan adalah kehadiran (0.127) dan umur (0.126), yang menegaskan bahwa kedisiplinan hadir dan kondisi usia menjadi indikator kuat terhadap risiko putus sekolah. Selain itu, variabel akademik seperti study time (0.106) dan gagal kelas (0.099) turut memberikan pengaruh penting terhadap kestabilan belajar santri. Faktor sosial seperti interaksi teman asrama (0.061) dan interaksi teman kelas (0.045) juga berkontribusi meski dalam porsi yang lebih kecil, menunjukkan bahwa lingkungan pergaulan tetap relevan dalam membentuk perilaku belajar. Variabel jam belajar dan kesehatan memiliki kontribusi paling rendah, namun masih memberikan informasi tambahan bagi model. Secara keseluruhan, hasil ini menegaskan bahwa kombinasi aspek

akademik, kedisiplinan, dan faktor sosial memberikan pengaruh terhadap efektivitas prediksi Random Forest dalam skenario 1.

Setelah proses pelatihan model menggunakan data latih, tahap selanjutnya adalah pengujian model menggunakan data uji**.** Pengujian ini bertujuan untuk mengetahui kemampuan model dalam mengklasifikasikan status santri ke dalam kategori *Lulus* dan *Putus Sekolah* pada data yang belum pernah digunakan dalam proses pelatihan. Evaluasi kinerja model pada tahap ini dilakukan dengan menggunakan Confusion *matrix*, yang digunakan untuk menggambarkan jumlah prediksi benar dan salah yang dihasilkan oleh model pada masing-masing kelas.

Sedangkan hasil dari pengujian model skenario 1 dalam melakukan prediksi santri “lulus” dan “putus sekolah” di pondok pesantren sidogiri disajikan pada Gambar 4.3.



Gambar 4. 3 Hasil Prediksi Random Forest Model Skenario 1

Hasil prediksi terhadap santri di pondok pesantren sidogiri pada *Confusion Matrix* gambar 4.3 menunjukkan bahwa model pada pengujian data uji 177

mampu meprediksi Lulus dan Putus Sekolah dengan baik. Model pada data uji ini mampu memprediksi Lulus 42 dengan benar, sementara 23 data salah diklasifikasikan sebagai Putus Sekolah. Model menunjukkan performa yang optimal dalam klasifikasi putus sekolah dengan 102 data. Sementara 10 data salah diprediksi sebagai *Lulus*. Hasil ini menunjukkan model lebih akurat dalam mengenali santri yang *putus sekolah* dibandingkan santri yang *lulus*, terlihat dari selisih kesalahan prediksi yang relatif kecil pada kelas tersebut.

Setelah menguji model dengan *Confusion matrix* selanjutnya menggambarkan kemampuan kinerja model secara lebih kuantitatif, evaluasi selanjutnya dilakukan menggunakan metrik *Accuracy, precision, recall, dan F1-score*. Metrik-metrik tersebut digunakan untuk mengukur tingkat keakuratan, kelengkapan, serta keseimbangan kinerja model dalam mengklasifikasikan status santri pada data *testing*

Tabel 4. 2 Hasil Pengujian Model *Random Forest* (*n_estimators* = 50)

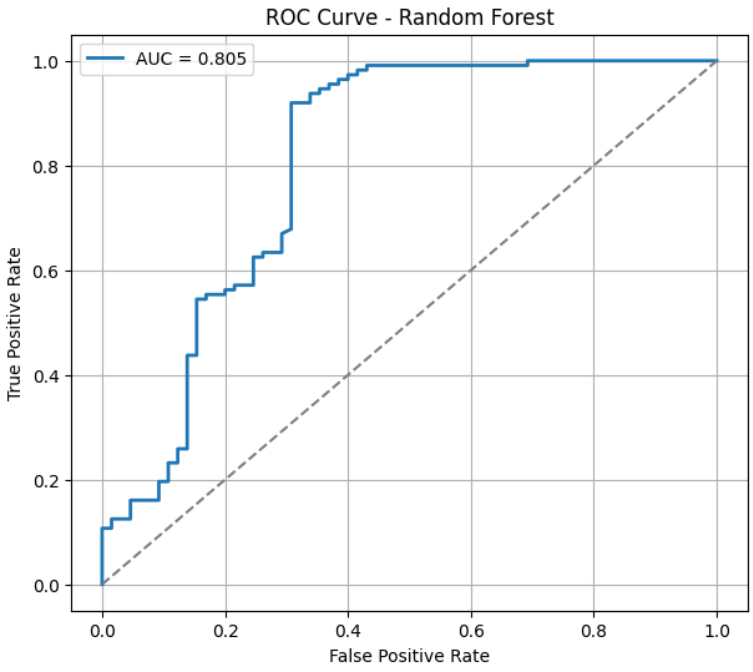
Kategori	***Accuracy***	***Precision***	***Recall***	***F1-Score***	***Support***
Lulus	0.83	0.87	0.63	0.73	65
Putus Sekolah		0.82	0.95	0.88	112
Macro Avg		0.84	0.79	0.80	177
Weighted Avg		0.84	0.83	0.82	177

Hasil Pengujian model dalam prediksi santri putus sekolah di pondok pesantren sidogiri berhasil memprediksi dengan benar 83 dari 100 data yang diuji. Sementara pada kategori Putus Sekolah model sangat baik dalam

mengenalinya, dari nilai *Precision* 0,82 dan *Recall* 0,95 yang tinggi. Sedangkan pategori Lulus nilai *Precision* 0,87 nilai *Recall* lebih rendah 0,63.

Nilai *Macro Average* sebesar 0,80 menunjukkan bahwa model memiliki performa yang stabil dan seimbang tidak hanya condong pada satu kategori.

Selanjutnya, untuk memvalidasi tingkat akurasi dan kemampuan generalisasi model terhadap data uji, kemudian fase evaluasi dilakukan untuk mengukur kualitas prediksi model dengan merujuk pada instrumen Kurva *Receiver Operating Characteristic* (ROC)



Gambar 4. 4 Kurva ROC dan Nilai AUC Model Skenario 1

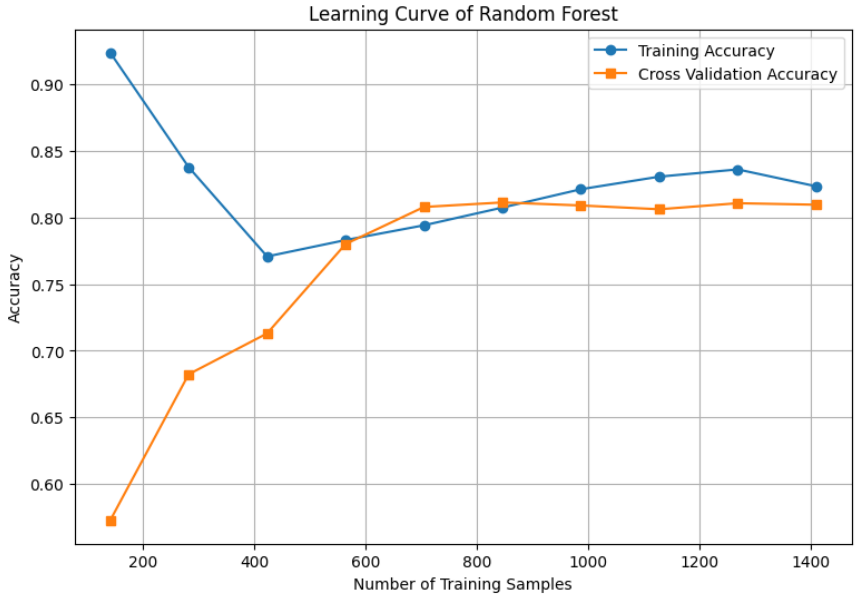
Berdasarkan gambar 4.4, kurva *ROC (Receiver Operating Characteristic)* bergerak sangat mendekati sudut kiri atas. model memiliki *True Positive Rate* ,kemampuan mendeteksi kelas positif yang sangat tinggi dengan *False Positive Rate*, kesalahan deteksi yang sangat rendah.

Sementara Skor *Area Under the Curve* (AUC) 0.99, nilai ini hampir menyentuh angka sempurna 1.00. Artinya, terdapat probabilitas sebesar 99% bahwa model mampu membedakan dengan benar antara kelas positif dan kelas negatif. Meskipun hanya menggunakan nilai parameter *max_depth*=5 dan *n_estimators*=50 sangat optimal.

4.2.2 Pelatihan Model Skenario 2

Pada Pelatihan prediksi santri putus sekolah di pondok pesantren sidogiri pada model Skenario 2, pelatihan dilakukan dengan pembagian data sebesar 80 untuk data training dan 20 untuk data testing sekitar 1411 data latih dan 353 data uji. Proporsi ini dirancang untuk memberikan lebih banyak data pelatihan kepada model *Random Forest* sehingga model dapat mempelajari pola klasifikasi secara lebih komprehensif. Eksperimen pengujian dengan nilai parameter `n_estimators` 100 serta menetapkan `Max_depth` 3. Dengan jumlah data latih yang lebih besar, diharapkan Random Forest mampu membangun

pohon-pohon keputusan yang lebih representatif dan menghasilkan prediksi yang lebih stabil.



Gambar 4. 5 *Learning Curve* Model *Random Forest* (n_estimators=100)

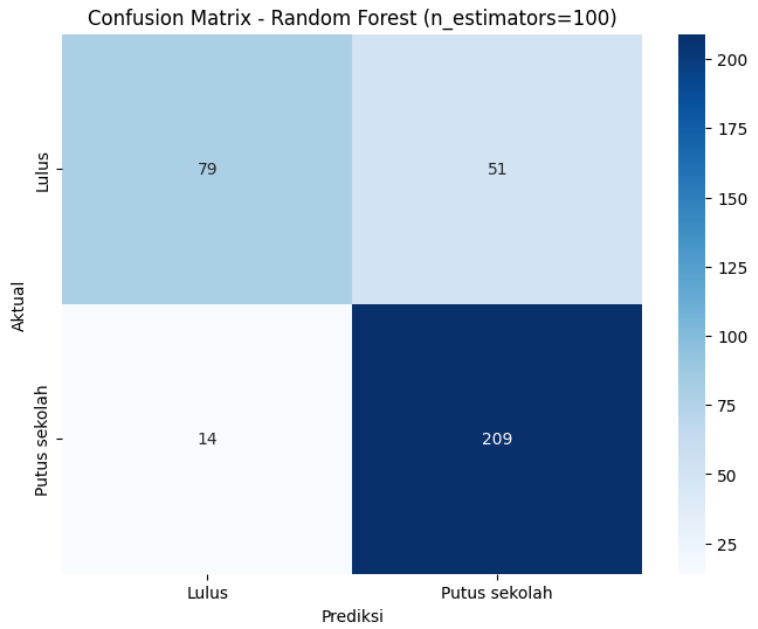
Grafik learning curve Pada Gambar 4.5 menunjukkan bahwa seiring bertambahnya jumlah data latih, training accuracy mengalami penurunan pada tahap awal dan kemudian stabil pada kisaran 0,82–0,84, sedangkan cross-validation accuracy meningkat dan stabil di sekitar 0,81. Kedua kurva mulai saling mendekat setelah jumlah data latih melebihi sekitar 500 sampel. Model pada pelatihan skenario 2 dengan total data latih 1441, model memiliki performa yang cukup baik dengan akurasi di atas 80% dalam menentukan santri "lulus" dan "putus sekolah" di pondok pesantren sidogiri.

Tabel 4. 3 Variabel berpengaruh pada Model Skenairo 2

No	Variabel	Nilai
1	kehadiran	0.20
2	kelas	0.18
3	umur	0.14
4	studytime	0.12
5	gagal _kelas	0.09
6	tingkat	0.08
7	interaksi_teman_asrama	0.05
8	interaksi_teman_kelas	0.04
9	domisili	0.02

Berdasarkan hasil pada tabel 4.3 variabel yang memiliki pengaruh pada model Skenario 2 kehadiran menyumbang nilai sekitar 0,20, diikuti oleh kelas (0,18) dan umur (0,14). Selanjutnya, variabel interaksi_teman_kelas (0,04), domisili (0,02), dan pendapatan_ibu (0,01) memiliki kontribusi yang sangat kecil terhadap pengaruh prediksi status santri.

Setelah proses pelatihan model skenario 2 menggunakan data *training*, tahap selanjutnya adalah melakukan pengujian model menggunakan data testing. Pengujian ini menggunakan *Confusion matrix*, yang digunakan untuk menggambarkan jumlah prediksi benar dan salah yang dihasilkan oleh model pada setiap kategori.



Gambar 4. 6 Hasil Prediksi Random Forest Model Skenario 2

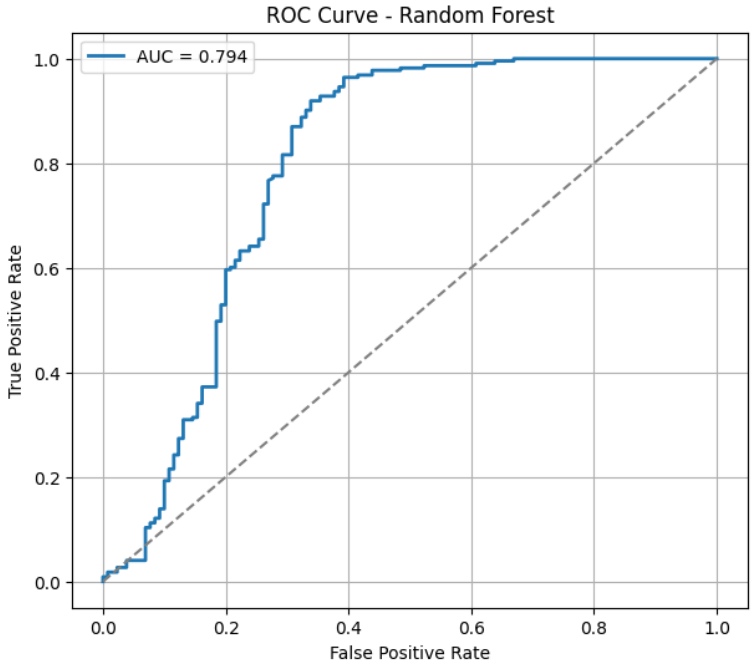
Hasil prediksi terhadap santri di pondok pesantren sidogiri pada *Confusion Matrix* gambar 4.6 menunjukkan bahwa model pada pengujian data uji 353 mampu meprediksi Lulus dan Putus Sekolah dengan baik. Model Random Forest dengan *n_estimators*=100 menunjukkan performa yang sangat baik dalam mengklasifikasikan santri dengan status Putus Sekolah, dengan tingkat keberhasilan sebesar 93,7% 209 data. Namun, model masih memiliki tantangan dalam memprediksi kategori Lulus, di mana terdapat 51 data yang salah terklasifikasi sebagai Putus Sekolah.

Setelah menguji model dengan *Confusion matrix* selanjutnya mengevaluasi dengan menggunakan metrik *Accuracy, precision, recall, dan F1-score*. Metrik-metrik tersebut digunakan untuk mengukur tingkat ketepatan, kelengkapan, serta keseimbangan performa model dalam mengklasifikasikan status santri pada data *testing*

Tabel 4. 4 Hasil Pengujian Model *Random Forest* (*n_estimators* = 100)

Kategori	***Accuracy***	***Precision***	***Recall***	***F1-Score***	***Support***
Lulus	0.81	0.87	0.62	0.72	130
Putus Sekolah		0.81	0.95	0.87	223
Macro Avg		0.84	0.78	0.80	353
Weighted Avg		0.83	0.82	0.82	353

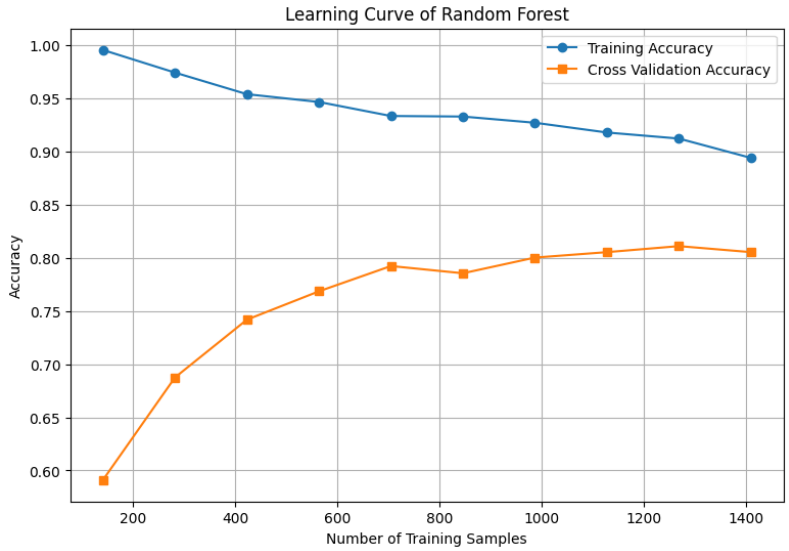
Hasil Pengujian model dalam prediksi santri putus sekolah di pondok pesantren sidogiri pada tabel 4.4. Model memiliki tingkat akurasi sebesar 81%, yang menunjukkan kemampuan prediksi yang optimal pada pengujian 353 data uji.



Gambar 4. 7 Kurva ROC dan Nilai AUC Model Skenario 2

4.2.3 Pelatihan Model Skenario 3

Pada peltihan model Skenario 3, pengujian dilakukan dengan menggunakan pembagian data sebesar 70% sebagai data training dan 30% sebagai data testing sekita 1234 data latih dan 530 data uji. Pembagian ini memberikan porsi data uji yang lebih besar dibandingkan Model sebelumnya, sehingga model diuji dengan variasi data yang lebih luas untuk melihat ketahanannya dalam menghadapi distribusi data yang lebih beragam. Pengujian Model skenario 3 menerapkan nilai parameter *n_estimators* 150 dan m*ax_depth* 9



Gambar 4. 8 *Learning Curve* Model Random Forest (n_estimators=150)

Training Accuracy pada kurva biru menunjukkan akurasi pelatihan yang dimulai dari nilai akurasi sempurna (1.00) dan mengalami penurunan bertahap hingga stabil di sekitar 0.90 saat jumlah sampel meningkat. Sedangkan Kurva oranye pada *Cross Validation Accuracy* menunjukkan peningkatan akurasi validasi yang signifikan seiring bertambahnya jumlah sampel, bergerak dari

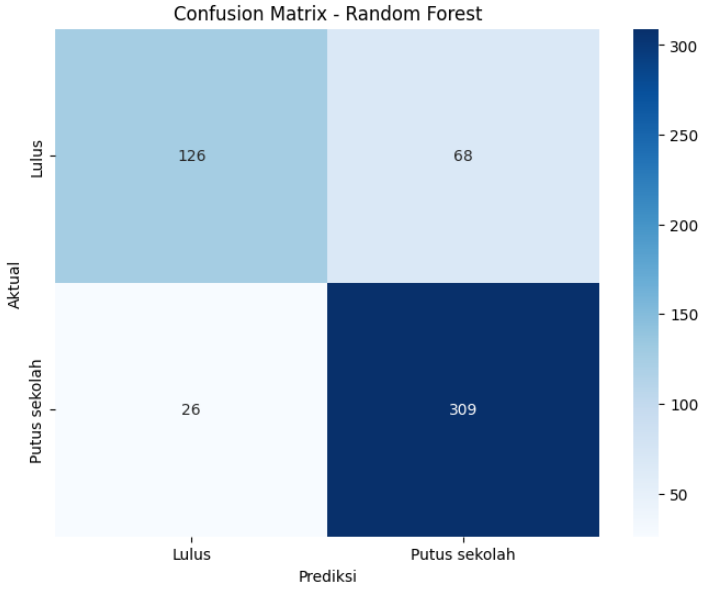
0.60 hingga mencapai titik stabil di atas 0.80. Namun jarak yang semakin mengecil antara *Training Accuracy* dan *Cross Validation Accuracy* menunjukkan bahwa model *Random Fores* pada model skenario 3 memiliki kemampuan generalisasi yang baik.

Tabel 4. 5 Variabel berpengaruh pada Model Skenario 3

No	**Variabel**	**Nilai**
1	Kehadiran	0.18
2	Kelas	0.16
3	Umur	0.13
4	Study Time	0.11
5	Gagal Kelas	0.10
6	Interaksi Teman Asrama	0.09
7	Tingkat	0.07
8	Interaksi Teman Kelas	0.06
9	Domisili	0.02

Berdasarkan Tabel 4.5, variabel kehadiran memiliki nilai pengaruh tertinggi dengan nilai 0.18, yang menunjukkan bahwa tingkat kehadiran santri merupakan faktor berpengaruh dalam prediksi status Lulus atau Putus Sekolah. Selanjutnya, variabel kelas dan umur juga memberikan kontribusi terhadap kinerja model dalam prediksi santri lulus atau putus sekolah. Sedangkan domisili memiliki pengaruh yang sangat kecil

Seusai proses pelatihan model skenario 3 dengan *n_estimators*=150, selanjutnya model diuji dengan 530 data testing . Pengujian ini melibatkan *Confusion matrix*. Hasilnya sebagai berikut.



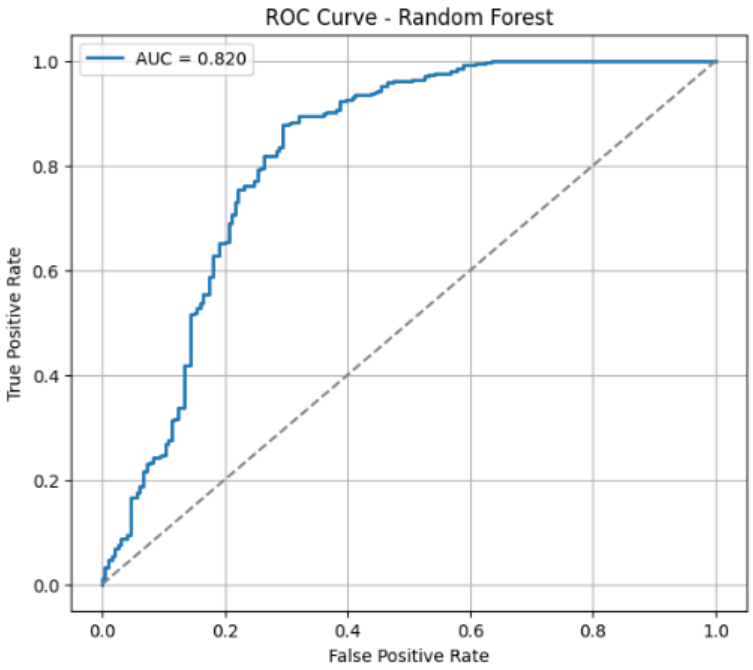
Gambar 4. 9 Hasil Prediksi *Random Forest* Model Skenario 3

Pengujian model Skenario 3 dengan *Confution Matrix* menghasilkan nilai kategori Lulus dan Putus Sekolah. Model berhasil memprediksi 309 santri putus sekolah diprediksi Putus sekolah dan 126 santri Lulus diprediksi Lulus. Dan model salah dalam memprediksi 26 santri Putus sekolah diprediksi Lulus dan 68 santri Lulus diprediksi Putus sekolah. Model menunjukkan lebih sensitif dalam mendeteksi santri Putus sekolah.

Tabel 4. 6 Hasil Pengujian Model *Random Forest* (*n_estimators* = 150)

Status	***Accuracy***	***Precision***	***Recall***	***F1-Score***	***Support***
Lulus	0.82	0.83	0.65	0.73	194
Putus Sekolah		0.82	0.92	0.87	335
Macro Avg		0.82	0.79	0.80	529
Weighted Avg		0.82	0.82	0.82	529

Hasil pengujian model skenario 3 pada tabel 4.6, model dengan nilai parameter *n_estimators* = 150 memperoleh akurasi sebesar 0,82, yang menunjukkan kinerja klasifikasi yang baik. Kelas Putus Sekolah memiliki nilai *recall* tertinggi 0,92 dan *F1-score* sebesar 0,87, menjelaskan model mampu mengidentifikasi santri putus sekolah. Sementara, kelas Lulus memiliki nilai *recall* sebesar 0,65, yang mengindikasikan masih terdapat sebagian salah dalam memprediksi santri lulus.



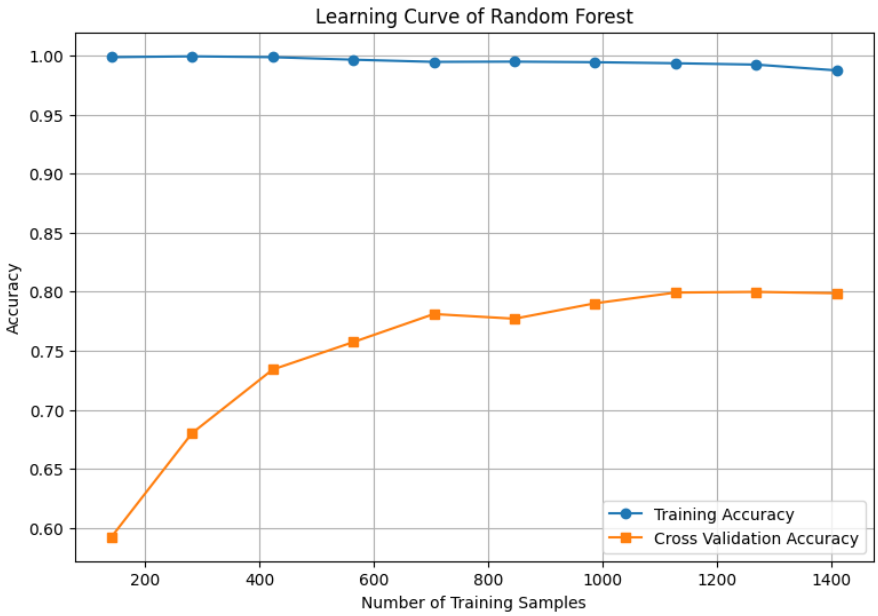
Gambar 4. 10 Kurva ROC dan Nilai AUC Model Skenario 3

Gambar ROC Curve pada skenario 3 menunjukkan kemampuan model Random Forest dalam membedakan kelas *lulus* dan *putus sekolah* dengan cukup baik. Nilai AUC sebesar 0.820 mengindikasikan bahwa performa model berada pada kategori *good classifier*, artinya model mampu melakukan pemisahan dua kelas dengan tingkat ketepatan yang tinggi. Kurva ROC yang berada jauh di atas garis diagonal (baseline) memperlihatkan bahwa tingkat

True Positive Rate lebih dominan dibanding False Positive Rate pada berbagai ambang keputusan (threshold). Dengan demikian, model Random Forest pada gambar 4.10 menunjukkan stabilitas dan reliabilitas yang kuat dalam mendeteksi santri berisiko putus sekolah.

4.2.4 Pelatihan Model Skenario 4

Kali ini tahap pelatihan pada model Skenario 4 dalam prediksi santri putus sekolah di pondok pesantren sidogiri, memakai porsi data latih 60 dan data uji 40 dengan besaran data 1058 data latih dan 706 data uji. Pengujian Model Skenario 4 menerapkan nilai parameter *n_estimators* 500 dan m*ax_depth* 20



Gambar 4. 11 Learning Curve Model Random Forest (n_estimators=500)

Grafik learning kurva dengan nilai parameter *n_estimators 500* dan *max_depth* 15 pada gambar 4.11 menunjukkan bahwa akurasi data latih *training accuracy* berada pada nilai yang sangat tinggi, mendekati 1,00, pada

seluruh variasi jumlah data latih. Hal ini menandakan bahwa model Random Forest mampu mempelajari data latih dengan sangat baik.

Sementara itu, *cross-validation accuracy* mengalami peningkatan seiring bertambahnya jumlah data latih, dari sekitar 0,59 pada jumlah data kecil hingga mencapai nilai stabil di kisaran 0,79–0,80 ketika jumlah data latih melebihi 1.000 sampel. Meskipun terjadi selisih yang cukup besar antara *training accuracy* dan *validation accuracy*. Kondisi ini mengindikasikan bahwa model cenderung mengalami *overfitting*. Namun untuk mengurangi overfitting lebih lanjut diperlukan *tuning parameter* agar kompleksitas model dapat dikendalikan dengan lebih baik.

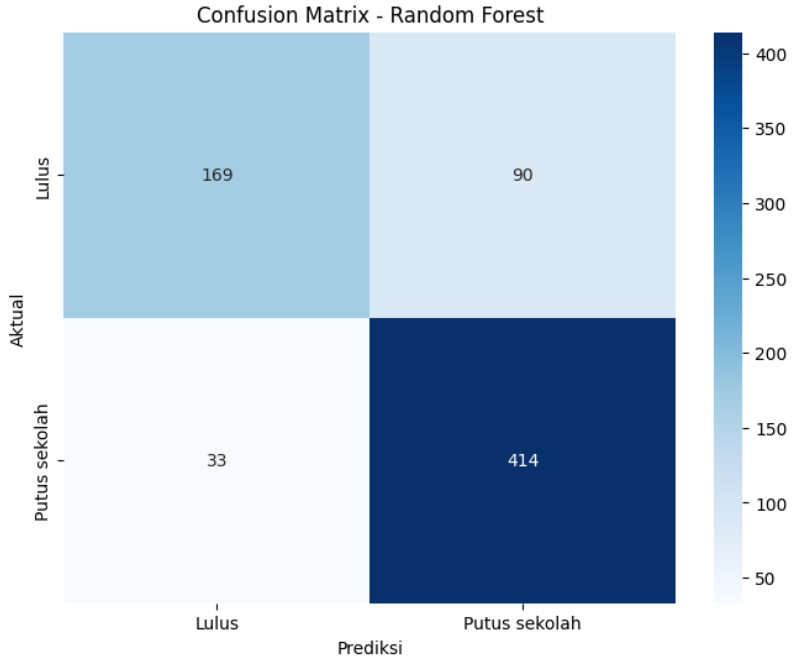
Tabel 4. 7 Variabel berpengaruh pada Model Skenario 4

No	Variabel	Nilai
1	Kelas	0.191
2	Umur	0.127
3	Gagal kelas	0.116
4	Kehadiran	0.109
5	Study time	0.105
6	Tingkat	0.067
7	Interaksi teman asrama	0.053
8	Interaksi teman kelas	0.036
9	Kesehatan	0.033
10	Jam belajar	0.024

Berdasarkan hasil analisis fitur pada model Skenario 4, variabel yang paling berpengaruh dalam proses klasifikasi adalah *kelas* dengan nilai importance sebesar 0.191, menunjukkan bahwa jenjang atau tingkat kelas santri menjadi faktor dominan dalam memprediksi status putus sekolah. Variabel berikutnya yang juga memiliki kontribusi kuat adalah *umur* (0.127), diikuti oleh

gagal_kelas (0.116) dan *kehadiran* (0.109), yang seluruhnya berkaitan erat dengan kedisiplinan dan performa akademik. Fitur *studytime* turut memberikan

Setelah proses pelatihan model skenario 4 dengan parameter *n_estimators* = 500 selesai dilakukan, tahap berikutnya adalah pengujian model menggunakan 760 data testing. Pengujian ini bertujuan untuk mengevaluasi kemampuan model dalam melakukan klasifikasi, yang dianalisis menggunakan Confusion Matrix. Adapun hasil pengujian tersebut ditunjukkan pada matriks kebingungan yang diperoleh



Gambar 4. 12 Hasil Prediksi *Random Forest* Model Skenario 4

Berdasarkan *Confusion Matrix* Gambar 4.12, model *Random Forest* mampu mengklasifikasikan data dengan cukup baik. Sebanyak 169 data lulus dan 414 data putus sekolah berhasil diprediksi dengan benar. Namun, masih terdapat 90 data lulus yang salah diprediksi sebagai putus sekolah dan 33 data putus sekolah yang salah diprediksi sebagai lulus. Hasil ini menunjukkan

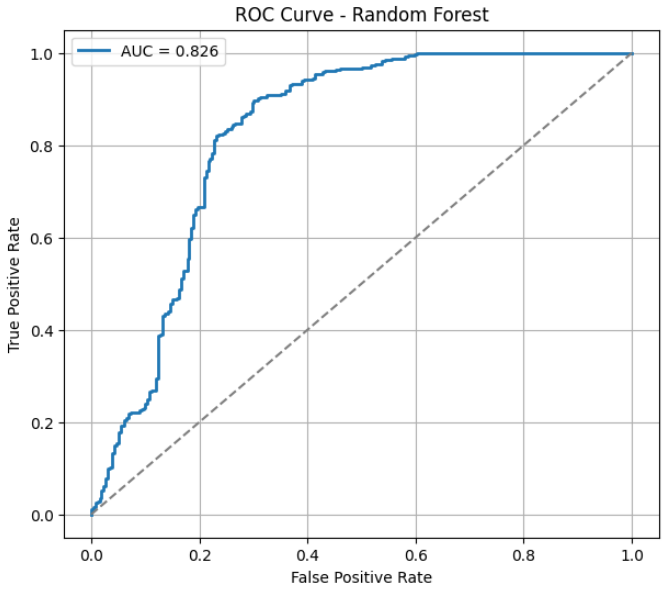
bahwa model lebih akurat dalam mendeteksi kelas putus sekolah dibandingkan kelas lulus.

Tabel 4. 8 Hasil Pengujian Model *Random Forest* (*n_estimators* = 500)

Status	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Support**
Lulus	0.84	0.86	0.66	0.75	259
Putus Sekolah		0.83	0.94	0.88	447
Macro Avg		0.84	0.80	0.81	706
Weighted Avg		0.84	0.84	0.83	706

Hasil pengujian model skenario 4, model dengan nilai parameter *n_estimators* = 500. Berdasarkan Tabel 4.8, model Skenario 4 ,dengan 40 data *testing* handal meprediksi santri putus sekolah di pondok pesantren sidogiri dengan nilai *Recall* 0.94, namun kurang sensitif dalam mendeteksi siswa yang Lulus dengan nilai *Recall* 0.66). model pada pengujian Skenario 4 cukup stabil dengan *nilai F1-Score* rata-rata di angka 0.81-0.83.

pengaruh signifikan dengan nilai 0.105, menegaskan bahwa durasi belajar santri berperan dalam keberhasilan akademik. Faktor lain seperti *tingkat*, *interaksi teman asrama*, dan *interaksi teman kelas* memiliki kontribusi moderat, sedangkan *kesehatan* dan *jam belajar* memberi pengaruh lebih kecil namun tetap relevan dalam model. Temuan ini menunjukkan bahwa kombinasi faktor akademik, sosial, dan perilaku memiliki peran penting dalam memprediksi risiko putus sekolah pada santri.



Gambar 4. 13 Grafik ROC Curve Model Skenario 4.

Gambar 4.13 menunjukkan kurva *Receiver Operating Characteristic* (ROC) dari model Random Forest pada skenario 4 dengan pembagian data 60% training dan 40% testing. Kurva ROC tampak berada jauh di atas garis diagonal, yang menunjukkan bahwa model memiliki kemampuan klasifikasi yang baik dalam membedakan antara kelas *Lulus* dan *Putus Sekolah*. Nilai *Area Under Curve* (AUC) sebesar 0.826 mengindikasikan performa model berada pada kategori *sangat baik*, karena mendekati nilai maksimum 1.0. Semakin tinggi nilai AUC, semakin kuat kemampuan model dalam mengidentifikasi kelas positif, yang dalam penelitian ini adalah santri berisiko Putus Sekolah. Dengan nilai AUC di atas 0.80, hasil ini menegaskan bahwa model Random Forest pada skenario 4 memiliki tingkat sensitivitas dan spesifisitas yang kuat serta konsisten dalam membedakan kedua kelas secara efektif.

4.3 Tuning Parameter

Pada tahanapan ini penulis melakukan tuning parameter dan Tujuan melakukan tuning parameter pada metode Random Forest adalah untuk menemukan keseimbangan optimal antara performa prediksi dan generalisasi model agar tidak terjadi *overfitting* maupun *underfitting* menurut Breiman (2001).

Tabel 4. 9 Parameter *Tuning Random Forest*

Parameter	Keterangan
n_estimators	Menentukan jumlah pohon, semakin banyak pohon, semakin stabil prediksi, namun waktu komputasi meningkat.
max_depth	Menentukan kedalaman maksimum setiap pohon keputusan, untuk mengontrol kompleksitas model
min_samples_split	Jumlah minimum sampel yang diperlukan untuk membagi sebuah node.
min_samples_leaf	Jumlah minimum sampel pada node daun. Parameter ini membantu mengurangi varians dan overfitting.
max_features	Jumlah fitur yang dipilih secara acak pada setiap proses pemisahan node. Digunakan untuk meningkatkan keragaman antar pohon.
criterion	Fungsi untuk mengukur kualitas pemisahan node.

Dengan menyetel parameter dan nilai seperti *n_estimators*, *max_depth*, dan *min_samples_split*, dapat mengontrol kompleksitas setiap pohon keputusan sehingga model tidak hanya sekadar menghafal data latihan, tetapi juga mampu memberikan prediksi yang akurat ketika menghadapi data baru yang belum pernah dilihat sebelumnya Probst *et al.* (2019). Selain itu, proses tuning ini berfungsi untuk mengoptimalkan penggunaan sumber daya komputasi dan menyesuaikan sensitivitas model terhadap ketidakseimbangan kelas (*imbalanced* data) melalui pengaturan bobot, sehingga hasil akhir model

menjadi lebih stabil dan reliabel dalam mengklasifikasikan data yang tidak seimbang Pratama (2020)."

Langkah selanjutnya mengimplementasikan tuning parameter dengan nilai yang telah ditentukan secara otomatis menggunakan *GridSearchCV* pada metode *Random Forest* dalam meprediski santri putus sekolah

```
=== PARAMETER TERBAIK RANDOM FOREST ===
{'criterion': 'entropy', 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 100}

========================================
AKURASI DATA TRAINING : 0.8966
AKURASI DATA TESTING  : 0.8192
========================================

[LAPORAN KLASIFIKASI DATA TRAINING]
              precision    recall  f1-score   support

       Lulus       0.95      0.76      0.84       582
Putus sekolah      0.88      0.98      0.92      1004

     accuracy                           0.90      1586
    macro avg       0.91      0.87      0.88      1586
 weighted avg       0.90      0.90      0.89      1586
```

Gambar 4. 14 Hasil Evaluasi *Random Forest* dengan Tuning Parameter

Hasil pengujian model *Random Forest* setelah melakukan tuning parameter sebagaimana disajikan pada Gambar 4.14 menunjukkan bahwa model dengan parameter terbaik *criterion=entropy*, *n_estimators*=100, *min_samples_split*=10, *min_samples_leaf*=1 menghasilkan akurasi data *training* sebesar 89,66% dan akurasi data *testing* sebesar 81,92%**.** Perbedaan nilai akurasi tersebut mengindikasikan adanya overfitting ringan, namun masih dalam batas wajar. Berdasarkan klasifikasi data *training*, model memiliki kinerja yang baik dalam mengenali kelas *putus sekolah* dengan *recall* sangat tinggi 0,98**,** sementara performa pada kelas *lulus* juga tergolong baik dengan *f1-score* sebesar 0,84**.** Secara keseluruhan, nilai weighted average f1-score sebesar 0,89 menunjukkan bahwa model cukup stabil dan efektif dalam melakukan klasifikasi potensi putus sekolah santri.

Untuk mendapatkan performa model Random Forest yang optimal dalam memprediksi status Lulus dan Putus Sekolah, maka dilakukan serangkaian pengujian menggunakan metode *parameter tuning*. Proses ini bertujuan untuk mencari kombinasi parameter terbaik, seperti jumlah pohon (*n_estimators*), kedalaman pohon (*max_depth*), serta kriteria pembagian data (*criterion*), yang dapat menghasilkan tingkat akurasi tertinggi. Berdasarkan hasil uji coba *tuning paramater* terdapat 10 kombinasi parameter sebagai berikut

Tabel 4. 10 Hasil Tuning Parameter Model *Random Forest*

No	n_estimators	max_depth	min_samples_split	min_samples_leaf	criterion	Accuracy
1	100	None	10	1	entropy	**0,8336**
2	50	None	5	1	entropy	0,8323
3	100	20	10	1	entropy	0,8323
4	50	None	10	1	entropy	0,8317
5	50	20	5	1	gini	0,8317
6	150	20	10	1	entropy	0,8317
7	100	20	5	1	entropy	0,8310
8	100	20	2	2	entropy	0,8310
9	150	20	2	2	entropy	0,8310
10	100	None	2	2	gini	0,8310

Hasil pengujian tuning parameter menggunakan *GridSearchCV*, diperoleh sepuluh kombinasi parameter hasil tuning model Random Forest sebagaimana disajikan pada Tabel 4.10. Kombinasi terbaik diperoleh pada parameter *n_estimators*=100**,** *max_depth*=None**,** min_samples_split=10**,** min_samples_leaf=1**,** dan *criterion=entropy***,** dengan nilai mean accuracy cross-validation sebesar 83,36%**.** Hasil ini menunjukkan bahwa penggunaan

fungsi pemisah entropy dengan jumlah pohon yang cukup mampu memberikan kinerja klasifikasi yang optimal dan stabil.

4.4 Pengujian Model Random Forest

Setelah tahap pelatihan rampung, setiap model *Random Forest* diuji menggunakan data pengujian untuk mengukur efektivitas prediksi santri putus sekolah di pondok pesantren sidogiri dalam kategori Lulus dan Putus Skolah. Penilaian performa dilakukan melalui analisis matriks evaluasi, mencakup *accuracy*, *precision*, *recall*, dan *F1-Score*, guna mendapatkan gambaran kinerja model yang optimal dari berbagai kategori.

Tabel 4. 11 Hasil Pengujian Model *Random Forest*

Skenario	Status	Accuracy	Precision	Recall	F1-Score	Support
1	Lulus	0.83	0.87	0.63	0.73	65
	Putus Sekolah		0.82	0.95	0.88	112
2	Lulus	0.81	0.87	0.62	0.72	130
	Putus Sekolah		0.81	0.95	0.87	223
3	Lulus	0.82	0.83	0.65	0.73	194
	Putus Sekolah		0.82	0.92	0.87	335
4	Lulus	0.84	0.86	0.66	0.75	259
	Putus Sekolah		0.83	0.94	0.88	447

Hasil pengujian model pada tabel 4.11 dari skenario 1- 4. Performa tertinggi, skenario 4 tercatat sebagai model dengan kinerja paling optimal secara keseluruhan, dengan capaian *Accuracy* sebesar 0,84. Pada skenario 4, model menunjukkan keseimbangan yang sangat baik dengan nilai *F1-Score* tertinggi di kedua kelas, yakni 0,75 untuk kelas Lulus dan 0,88 untuk kelas Putus Sekolah. Sedangkan efektivitas deteksi kelas putus sekolah secara

konsisten di seluruh skenario 1-4, model *Random Forest* menunjukkan kemampuan yang sangat superior dalam mengidentifikasi santri pada kategori Putus Sekolah. Hal ini dibuktikan dengan nilai *Precision* yang mencapai rentang 0,82 hingga 0,83 dan *Recall* stabil pada angka 0,92 hingga 0,95. Tingginya nilai *recall* mencapai angka maksimal 0,95 pada Skenario 1 dan 2 menandakan bahwa model mampu meminimalkan risiko luputnya deteksi pada santri yang berpotensi putus sekolah.

Meskipun akurasi keseluruhan tinggi, metrik *recall* pada kelas Lulus berada pada rentang 0,62 hingga 0,66. Hal ini mengindikasikan bahwa terdapat tantangan bagi model dalam mengenali secara lengkap seluruh santri yang lulus dibandingkan dengan ketajamannya pada kelas Putus Sekolah.

Stabilitas berdasarkan support data, skenario 4 memiliki jumlah Support data paling besar yaitu 706 data, namun tetap mampu mempertahankan nilai akurasi tertinggi. Hal ini menunjukkan bahwa Model *Random Fo*rest memiliki stabilitas dan ketahanan yang baik terhadap peningkatan volume data uji tanpa mengorbankan kualitas prediksi.

Berdasarkan metrik evaluasi tersebut, skenario 4 ditetapkan sebagai konfigurasi terbaik untuk metode *Random F*orest. Capaian akurasi 0,84 menjadikannya salah satu representasi model yang sangat handal untuk dijadikan dasar perbandingan terhadap metode klasifikasi lainnya dalam penelitian ini.

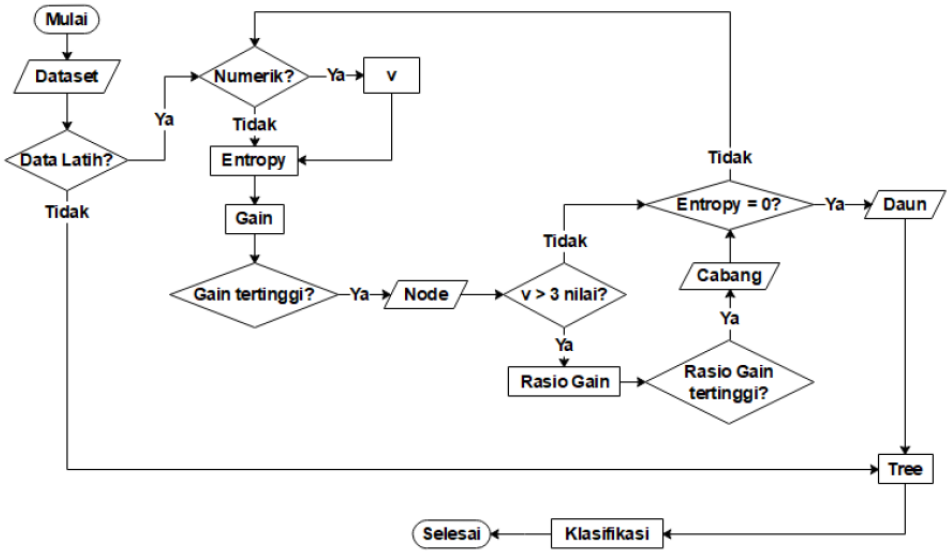
BAB V

METODE DECISION TREE

5.1 Desain Metode

Untuk memperoleh hasil prediksi yang lebih optimal, penelitian ini menggunakan dua metode algoritma yaitu Decision Tree (DT) metode kedua setelah metode Random Forest (RF). Metode Decision Tree dipilih karena mampu memberikan aturan keputusan yang jelas dan mudah diinterpretasikan, sementara metode Random Forest dipilih karena mempunyai akurasi yang lebih tinggi serta lebih stabil terhadap variasi data. Dengan membandingkan kedua metode ini diharapkan diperoleh hasil prediksi yang lebih baik.

Sedangkan desain diagram alur metode Decision Tree disajikan dalam bentuk flowchart pada gambar berikut 5.1.



Gambar 5. 1 *Flow Chart* Algoritma De*cision Tree*

Gambar 5.1 menunjukkan alur kerja algoritma *Decision Tree* dalam membentuk model klasifikasi. Proses diawali dengan pemanggilan *dataset* yang telah ditetapkan sebagai data latih, kemudian dilakukan identifikasi terhadap jenis atribut apakah bersifat numerik atau kategorikal. Untuk atribut kategorikal, sistem menghitung nilai *entropy* dan *information gain* guna menentukan atribut dengan nilai tertinggi yang akan dijadikan *node* utama. Jika atribut memiliki lebih dari tiga nilai, digunakan perhitungan *gain ratio* agar pembagian data lebih proporsional. Proses pembentukan *node* dan cabang terus berlanjut hingga diperoleh nilai *entropy* sebesar nol, yang menandakan bahwa data telah homogen dan terbentuk node daun sebagai hasil klasifikasi akhir. Tahapan ini menggambarkan mekanisme dasar pengambilan keputusan dalam algoritma *Decision Tree* sebagaimana dijelaskan oleh Aaboub F, *et al.* (2019).

Penentuan atribut yang akan menjadi akar dimulai dengan menghitung nilai *gain* dari setiap atribut. Atribut dengan nilai *gain* tertinggi kemudian dipilih sebagai akar utama. Sebelum proses perhitungan *gain* dilakukan, terlebih dahulu dihitung nilai *entropy* menggunakan rumus sebagai berikut:

$$Entropy\ (S) = \sum_{i=1}^{n} -pi * log2\ pi \qquad (5.1)$$

dengan:

S : Himpunan kasus

n : Jumlahh partisi S

Pi : Proporsi dari Si terhadap S

$$Gain(S,A) = Entropy\ (S) \sum_{i=1}^{n} - \frac{|Si|}{|Si|} * Entropy\ (S_i) \qquad (5.2)$$

dengan:

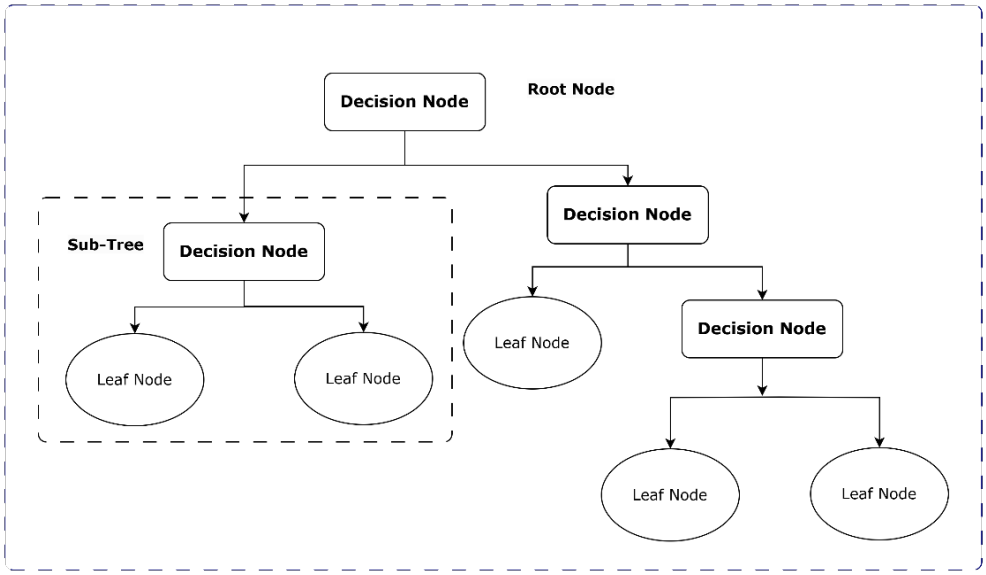
A : Atribut

n : Jumlah partisi atribut A

|Si| : Jumlah kasus pada partisi ke-i

|S| : Jumlah kasus dalam S

Pada persiapan awal ditentukan atribut yang digunakan kemudian melakukan uji atribut dengan mencari nilai *Gain* tertinggi berdasarkan perhitungan *entropy* dari masing-masing atribut. Apabila ditemukan gain tertinggi maka gain tersebut akan menjadi *root* awal. Selanjutnya dilakukan penentuan cabang dengan cara yang sama dengan melihat gain tertinggi dari tiap hasil partisi seperti pada gambar 5.2



Gambar 5. 2 *Architecture Decision Tree Classification Algorithm*

5.2 Konfigurasi Model

Pada tahap ini dilakukan proses konfigurasi awal sebelum model *Decision Tree Classifier* dijalankan. Konfigurasi ini bertujuan untuk menentukan parameter-parameter dasar yang berpengaruh terhadap hasil pemodelan, seperti penentuan variabel target, pembagian data, serta pengaturan *random seed* agar proses dapat direproduksi secara konsisten.:

```
# Konfigurasi Model
TARGET     = "status"               # variabel target yang akan diprediksi
DROP_COLS  = ["tanggal hijriah", "tanggal masehi"]  # kolom yang tidak digunakan dalam pe
TEST_SIZE  = 0.25                   # proporsi data uji sebesar 25%
RANDOM_SEED = 42                    # nilai seed untuk menjaga konsistensi hasil
CLASS_WEIGHT_BALANCED = False       # tidak menggunakan pembobotan kelas
```

Gambar 5. 3 Potongan kode Desicion Treen

Variabel target ditetapkan pada kolom *status*, yang berisi label klasifikasi santri yaitu Lulus atau Putus Sekolah. Pemilihan kolom ini sesuai dengan tujuan penelitian, yaitu memprediksi kemungkinan santri akan menyelesaikan pendidikannya atau tidak, lulus atau putus sekolah.

5.3 Implementasi Model Decision Tree

Pelatihan dengan metode *Decision Tree* untuk memprediksi status santri dengan kategori “Putus Sekolah” atau “Lulus”. Data yang digunakan merupakan data santri dengan atribut meliputi umur, pendidikan orang tua, pekerjaan orang tua, serta tingkat interaksi santri di pondok. Atribut-atribut ini dipilih karena dianggap berpengaruh terhadap kemungkinan santri bertahan hingga lulus atau berisiko mengalami putus sekolah.

Pada tahap ini dilakukan proses pembagian *dataset* ke dalam 4 skenario rasio dengan rasio, 60:40, 70:30, 80:20, dan 90:10, dengan perbandingan

proporsi antara data yang digunakan untuk pelatihan (*training*) dan data yang digunakan untuk pengujian (*testing*). Pengujian keempat skenario ini dijabarkan dalam tabel 5.1 beserta pembagian *dataset* data *training* dan data *testing*

Tabel 5. 1 Pembagian data *training* dan *testing*

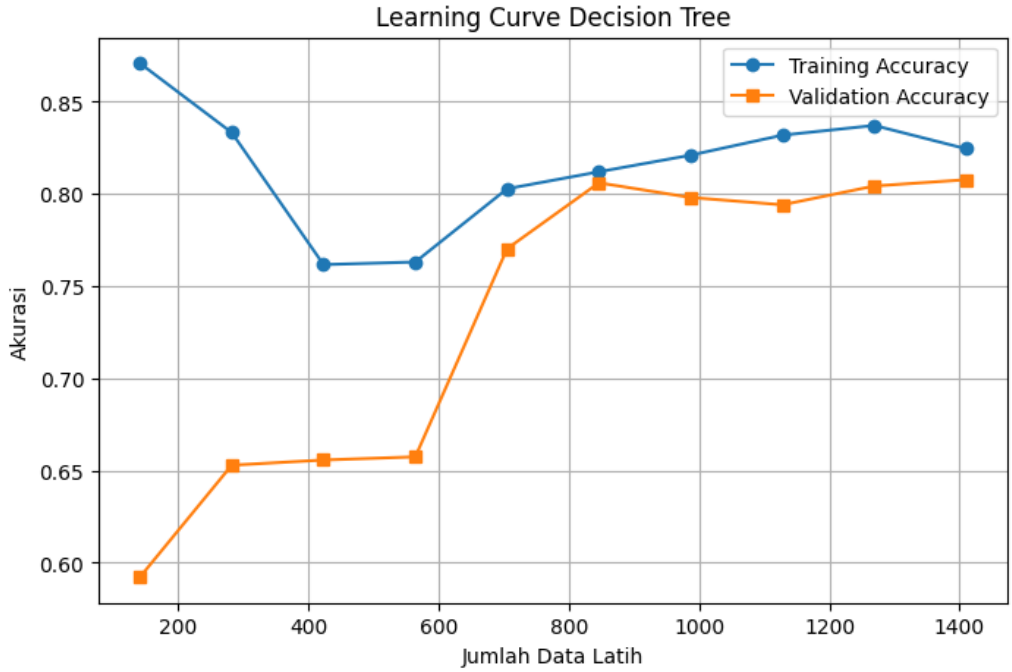
Pelatihan	**Rasio**	**Data Training**	**Data Testing**
Skenario 1	90 : 10	1.586	178
Skenario 2	80 : 20	1.411	353
Skenario 3	70 : 30	1.234	530
Skenario 4	60 : 40	1.058	706

Selanjutnya, algoritma *Decision Tree* diterapkan pada masing-masing skenario. Model dibangun menggunakan data pelatihan untuk menghasilkan struktur pohon keputusan berdasarkan perhitungan *entropy* dan *information gain*, kemudian dievaluasi dengan data pengujian guna mengukur tingkat kinerjanya.

Setiap hasil pengujian disimpan dalam suatu struktur data yang memuat informasi mengenai rasio pembagian, jumlah data pelatihan dan pengujian, serta indikator performa model seperti akurasi, *precision*, *recall*, dan *f1-score.* Melalui pendekatan ini, peneliti dapat melakukan perbandingan performa model secara sistematis, sehingga diperoleh konfigurasi pembagian data yang paling sesuai dan akurat untuk memprediksi status santri (Lulus atau Putus Sekolah) berdasarkan variabel-variabel yang digunakan.

5.3.1 Pelatihan Model Skanario 1

Pada pelatihan model skenario 1, model *Decision Tree* dilatih menggunakan parameter *criterion gini*, *max_depth* 3, *min_samples_split* 20, *min_samples_leaf* 10, dan *random_state* 42 dengan pembagian data *training* dan data *testing* menggunakan rasio 90:10. Hasil pelatihan model scenario 1 kemudian divisualisasikan dalam bentuk diagram *Learning Curve* dan diagram keputusan, sebagaimana ditampilkan pada gambar di atas, yang menggambarkan struktur aturan klasifikasi berdasarkan atribut yang digunakan.



Gambar 5. 4 *Learning Curve* Model *Decision Tree* Skenario 1

Hasil pelatihan model scenario 1 pada gambar *learning curve* 5.4 menunjukkan bahwa pada jumlah data latih yang kecil, hasil *training accuracy* sangat tinggi. Setelah jumlah data latih mencapai kisaran lebih dari 800 data, kedua kurva relatif stabi. Sementara seiring bertambahnya jumlah data latih,

nilai *training accuracy* meningkat, serta keduanya stabil di kisaran 0,80–0,83. Model pada pelatihan 1 semakin mampu melakukan generalisasi dengan baik dalam mengklasifikasikan prediksi santri putus sekolah di pondok pesantren sidogiri ke dalam kelas lulus dan putus sekolah**.** Pada pelatihan skenario 1 model *Decision Tree* cukup efektif dan stabil untuk memprediksi risiko santri putus sekolah berdasarkan data yang tersedia.

Tabel 5. 2 Nilai *Entropy* dan *Information Gain* – Rasio 90:10

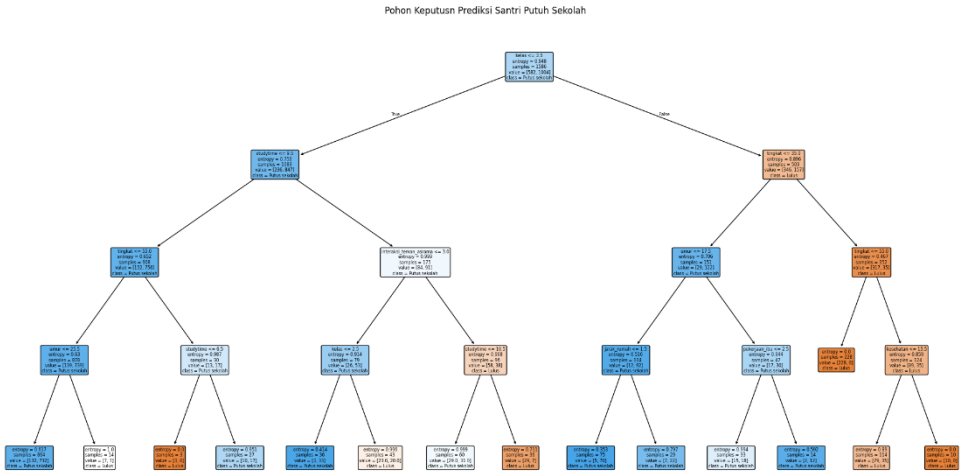
Node	Feature	Entropy	Information Gain
0	kelas	0.948310	0.147800
1	studytime	0.756339	0.048524
2	tingkat	0.651724	0.009672
3	umur	0.630262	0.007402
6	studytime	0.987138	0.131277
9	interaksi_teman_asrama	0.998846	0.054961
10	kelas	0.914019	0.183053
13	studytime	0.968461	0.077458
16	tingkat	0.895614	0.356808
17	umur	0.705746	0.056537
18	jarak_rumah	0.515947	0.038789
21	pekerjaan_ibu	0.944087	0.069908
24	tingkat	0.467192	0.164762
26	kesehatan	0.858509	0.040559

Nilai *Entropy* dan Information Gain pada setiap *node* dalam pohon keputusan dibangun dengan rasio data training 90:10. Simpul Akar, atau *Root Node* pada Node 0 sebagai proses awal dimulai atribut kelas yang memiliki nilai *Entropy* tertinggi sebesar 0,948310. Ini adalah titik dengan ketidakpastian terbesar di mana data masih sangat bercampur antara santri yang Lulus dan Putus Sekolah. Sementara Nilai *entropy* pada tabel 5.2 menerangkan faktor

akademik (kelas, tingkat, dan waktu belajar) adalah fitur yang paling sering muncul di bagian awal dan tengah pohon (*Node* 0-7), yang berarti menjadi atribut kunci. Sementara faktor sosial (interaksi teman, jarak rumah, pekerjaan ibu) muncul sebagai penyaring akhir untuk memperjelas prediksi.

Munculnya fitur seperti kelas (Node 0 dan 6), tingkat (Node 2, 8, dan 12), serta studytime (Node 1, 4, dan 7) pada beberapa *node* yang berbeda disebabkan oleh perhitungan Information Gain yang dilakukan secara lokal. Algoritma mengevaluasi kembali subset data pada setiap cabang baru; sehingga, atribut yang sama dapat terpilih kembali di level yang lebih dalam apabila atribut tersebut masih memberikan kontribusi pengurangan entropi paling signifikan dibandingkan atribut lainnya pada subset data tersebut.

Selanjutnya dibentuk struktur pohon keputusan sebagaimana ditunjukkan pada Gambar 5.5. Pohon keputusan ini menggambarkan alur pengambilan keputusan model dalam memprediksi santri putus sekolah, di mana setiap node merepresentasikan atribut pemisah, setiap cabang menunjukkan kondisi atribut, dan setiap leaf node menghasilkan keputusan akhir berupa status santri lulus dan putus sekolah.

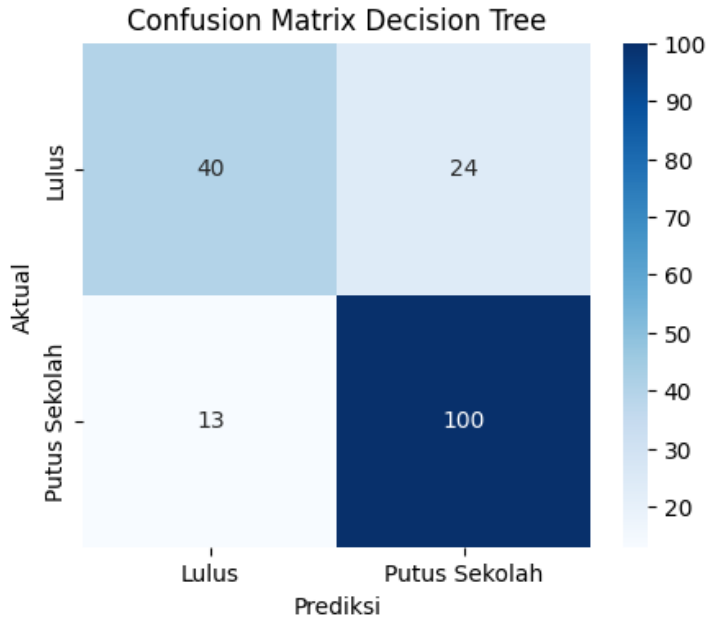


Gambar 5. 5 Struktur Pohon Keputusan Model Skenario 1

Struktur pohon keputusan pada gambar 5.5 Model menentukan bahwa atribut Kelas ≤ 3.5 adalah pemisah paling krusial bagi 1.586 data latih dengan entropi awal 0,948 dan menempatkan atribut kelas sebagai *root node* karena merupakan faktor pembeda paling signifikan dalam memprediksi status santri. Jika atribut Kelas ≤ 3.5, mayoritas data mengarah pada prediksi Putus Sekolah dan atribut akademik seperti studytime dan tingkat menjadi penyaring berikutnya. Sedangkan Jika Kelas ≥ 3.5, data cenderung mengarah pada prediksi Lulus. Alur kemudian bercabang ke internal node seperti studytime dan tingkat, yang menunjukkan bahwa performa akademik menjadi indikator utama dalam proses klasifikasi awal.

Setelah model selesai dilatih dan menghasilkan *learning curve*, nilai entropy dan *information gain*, serta struktur pohon keputusan, tahap selanjutnya adalah melakukan pengujian model menggunakan 178 data testing yang belum pernah digunakan pada proses pelatihan. Pengujian ini bertujuan untuk mengevaluasi kemampuan model dalam membedakan setiap kelas

secara tepat. Oleh karena itu, digunakan Confusion Matrix sebagai alat evaluasi, yang berfungsi untuk menunjukkan jumlah prediksi benar dan salah pada masing-masing kelas, termasuk kesalahan prediksi ketika santri yang sebenarnya lulus diprediksi putus sekolah, maupun santri yang sebenarnya putus sekolah tetapi diprediksi lulus.



Gambar 5. 6 *Confusion Matrix* Model Skenario 1

Berdasarkan *confusion matrix*, pelatihan model skenario 1 matrik ini mampu mengklasifikasikan data dengan cukup baik. Sebanyak 40 data santri lulus dan 100 data santri putus sekolah berhasil diprediksi dengan benar. Namun, masih ada 24 data santri lulus yang salah diprediksi sebagai putus sekolah serta 13 data santri putus sekolah yang salah diprediksi sebagai lulus. Matrik ini menunjukkan kemampuan yang lebih baik dalam prediksi kelas putus sekolah dibandingkan kelas lulus, sehingga efektif digunakan untuk mengidentifikasi santri yang berisiko putus sekolah.

Selepas model dilatih menggunakan data pelatihan (90%), langkah krusial berikutnya adalah melakukan evaluasi menggunakan data pengujian (10%) untuk mengukur sejauh mana model mampu memprediksi data santri yang belum pernah dilihat sebelumnya.

Berikut adalah penjelasan mengenai evaluasi model berdasarkan hasil pengujian model *Decision Tree* skenario 1 disajikan pada table berikut.

Tabel 5. 3 Hasil Pengujian Model *Decision Tree* Skenario 1

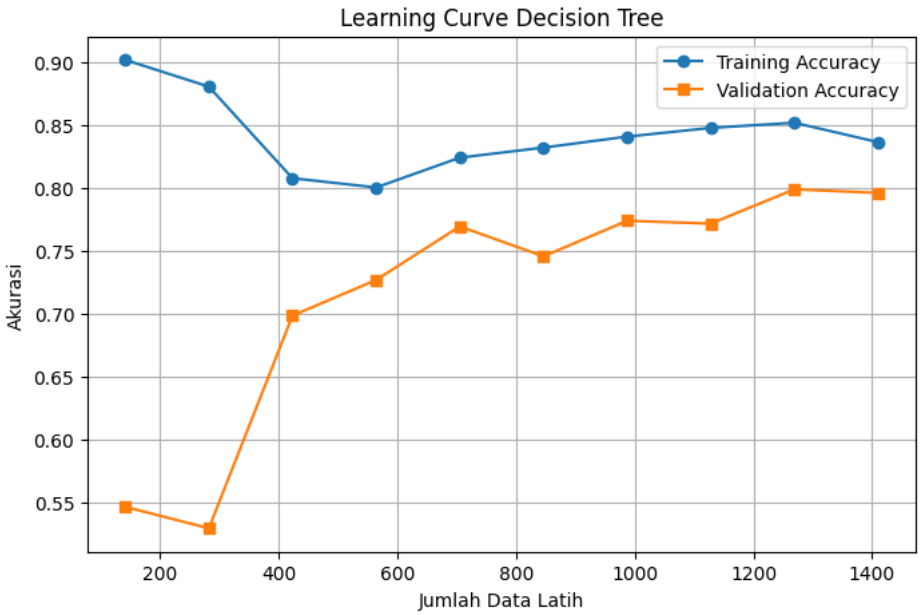
Kategori	***Accuracy***	***Precision***	***Recall***	***F1-Score***	***Support***
Lulus	0.79	0.75	0.62	0.68	65
Putus Sekolah		0.81	0.88	0.84	113
Macro Avg		0.78	0.75	0.76	177
Weighted Avg		0.79	0.79	0.79	177

Pengujian model skenario 1 model memiliki nilai akurasi 0,79, 177 data santri di kelompok penguji, model berhasil prediksi dengan benar sebanyak 140 santri baik yang Lulus maupun Putus Sekolah. Ketika model memprediksi data santri Putus Sekolah, tingkat presisinya mencapai 0.81. Model sangat sensitif dengan *Recall* 0.88 dari total data santri yang kenyataannya memang putus sekolah. Sejalan dengan tujuan penelitian ini, memprediksi risiko santri putus sekolah di pondok pesantren sidogiri. Sedangkan kategori Lulus Memiliki performa di bawah kategori Putus Sekolah dengan nila *F1-Score* 0,68. Hal ini dipengaruhi oleh jumlah data yang lebih sedikit 65 data dibandingkan kategori Putus Sekolah sebesar 113 data.

5.3.2 Pelatihan Model Skanario 2

Pada skenario 2, pelatihan dilakukan dengan pembagian data sebesar 80% untuk data *training* dan 20% untuk data *testing*. Proporsi ini bertujuan untuk memberikan data pelatihan yang lebih banyak kepada model agar mampu mempelajari pola klasifikasi dengan lebih baik.

Pengujian pada skenario ini model Decision Tree pada penelitian ini dibangun menggunakan parameter criterion *entropy*, dengan nilai *max_depth* sebesar 4, *min_samples_leaf* sebesar 10, dan *min_samples_split* sebesar 2. Selanjutnya, dilakukan analisis menggunakan diagram *learning curve* untuk mengetahui performa model terhadap variasi jumlah data pelatihan



Gambar 5. 7 *Learning Curve* Model *Decision Tree* Skenario 2

Grafik *Learning Curve* pada Skenario 2 (80:20) model memiliki kemampuan generalisasi yang baik, nilai *Training Accuracy* stabil di kisaran 0,84–0,85 dan *validation Accuracy* nilai meningkat signifikan hingga mencapai angka 0,80.

Training Accuracy mulai menurun dari nilai 0,90 ini merupakan proses wajar di mana model beralih dari sekadar menghafal data menjadi mempelajari pola yang lebih luas, sementara meningkatnya nilai *validation Accuracy* ketika jumlah data latih menambah sehingga memperkuat kemampuan model dalam memprediksi data baru. Model pada skenario 2 ini bebas dari masalah *overfitting*, sehingga struktur pohon keputusan yang terbentuk dianggap stabil dan handal untuk digunakan dalam prediksi santri Lulus dan Putus Sekolah.

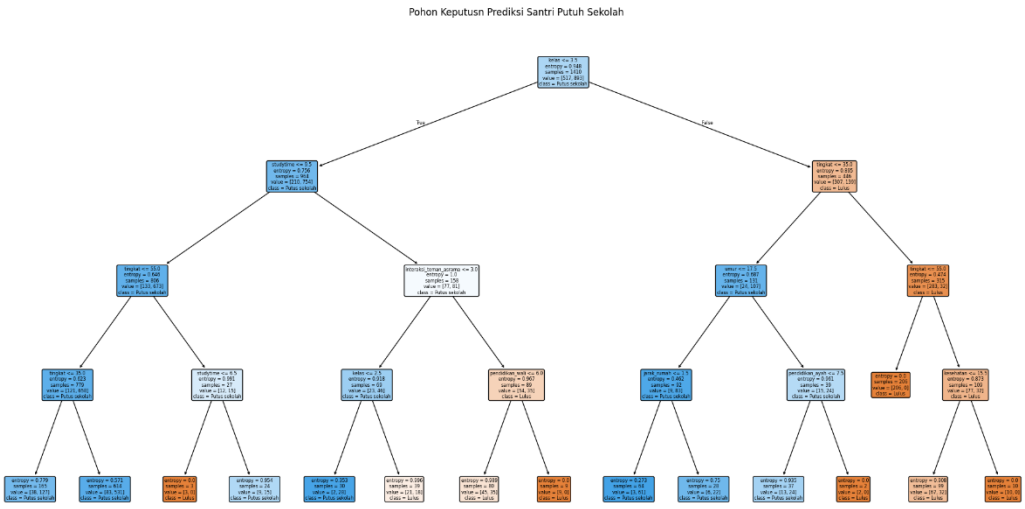
Tabel 5. 4 Nilai *Entropy* dan *Information Gain* – Rasio 80:20

Node	Feature	Entropy	Information Gain
0	kelas	0.948078	0.800135
1	studytime	0.756208	0.704085
2	tingkat	0.646167	0.635343
3	tingkat	0.623013	0.615332
6	studytime	0.991076	0.848386
9	interaksi_teman_asrama	0.999538	0.945657
10	kelas	0.918296	0.716437
13	pendidikan_wali	0.966870	0.888719
16	tingkat	0.895079	0.536580
17	umur	0.687041	0.610674
18	jarak_rumah	0.462066	0.418030
21	pendidikan_ayah	0.961237	0.887307
24	tingkat	0.474007	0.302189
26	kesehatan	0.873298	0.824572

Nilai *Entropy* dan Information Gain pada tabel 5.4 didominasi Faktor Akademik berupa atribut kelas, *study time*, dan atribut tingkat secara konsisten terpilih sebagai fitur pemisah di *Node* 0-3, yang menunjukkan bahwa variabel akademik merupakan prediktor paling kuat dalam menentukan status santri. Sedangkan *Root node*, *Node* 0 dimulai dengan Entropi 0,948 dan *Information*

Gain 0,800, yaitu atribut kelas sangat efektif dalam mengurangi ketidakpastian data di awal proses klasifikasi. Pada simpul yang lebih dalam seperti Node 9, 13, dan 18, model mulai melibatkan faktor sosial dan keluarga seperti interaksi teman asrama, pendidikan wali, dan jarak rumah untuk mempertajam prediksi pada subset data yang lebih kecil. Nilai entropi terendah terlihat pada *Node* 18 yaitu 0,462 dan *Node* 24 yaitu 0,474, yang berarti pada titik tersebut data sudah semakin murni menuju keputusan akhir antara lulus atau putus sekolah. Pada tabel 5.3 Satu fitur dapat muncul berkali-kali karena *Decision Tree* bekerja secara lokal, di mana perhitungan *Information Gain* dilakukan ulang pada setiap *node* menggunakan subset data yang berbeda. Salah satunya, fitur seperti tingkat muncul berulang pad *Node* 2, 3, 8, dan 12 dan kelas *Node* 0 dan 6, node tersebut dapat kembali terpilih apabila pada subset data tersebut masih memberikan pengurangan *entropy* terbesar dibandingkan fitur lainnya.

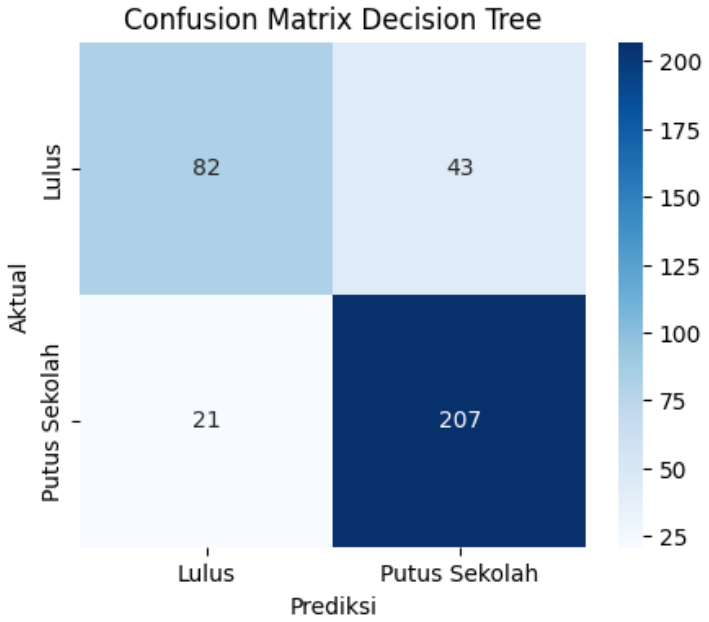
Tahap berikutnya membentuk struktur pohon keputusan yang disajikan pada Gambar 5.6, untuk menggambarkan proses klasifikasi santri berdasarkan atribut-atribut terpilih. Pada pohon keputusan, atribut sebagai *node*, cabang menyatakan kondisi dari atribut tersebut, dan *leaf node* memberikan hasil prediksi Lulus dan Putuh Sekolah.



Gambar 5. 8 Struktur Pohon Keputusan Model Skenario 2

Struktur Pohon keputusan Model Skenario 2 menempatkan fitur Kelas sebagai *Root Node* dengan nilai *Entropy* 0,948, yang menjadikannya faktor paling kritis dalam menentukan probabilitas kelulusan santri. Di level awal, interaksi antara fitur Kelas ≤ 3,5 dan Studytime ≤ 9,5 menjadi pembeda utama yang secara dominan mengarahkan prediksi pada status Putus Sekolah. Kekuatan model ini juga terlihat pada fitur Tingkat yang muncul berulang kali di Node 2, 3, 8, dan 12, menegaskan posisinya sebagai prediktor yang sangat kuat dalam memisahkan kelompok yang akan Lulus atau Putus Sekolah.

Setelah proses pelatihan model selesai dan diperoleh hasil berupa *learning curve*, nilai *entropy* dan *information gain*, serta struktur pohon keputusan, langkah berikutnya adalah melakukan evaluasi kinerja model dengan menggunakan 353 data uji. pengujian menggunakan *Confusion Matrix* sebagai metode evaluasi model prediksi yang akurat dan valid pada setiap kelas.



Gambar 5. 9 *Confusion Matrix* Model Skenario 2

Matriks pada skenario 2 menunjukkan kemampuan yang jauh lebih tajam dalam memprediksi kelas putus sekolah dibandingkan dengan kelas lulus. terlihat dari jumlah prediksi benar pada kelas putus sekolah 207 data yang mendominasi total data. Dengan tingkat keberhasilan menangkap data kasus putus sekolah yang tinggi, model ini sangat efektif dan reliabel untuk digunakan sebagai sistem peringatan dini dalam mengidentifikasi santri yang berisiko putus sekolah, sehingga langkah pencegahan dapat dilakukan lebih awal. Sebanyak 82 data santri lulus dan 207 data santri putus sekolah berhasil diprediksi dengan benar sesuai dengan data aktualnya.

Sedangkan masih terdapat 43 data santri lulus yang salah diprediksi sebagai putus sekolah (False Positive), serta 21 data santri putus sekolah yang salah diprediksi sebagai lulus (False Negative).

Setelah fase pelatihan model diselesaikan dengan porsi 80% data, fase selanjutnya adalah validasi melalui pengujian 20% data . Langkah ini bertujuan

untuk menguji keandalan dan kemampuan generalisasi model dalam mengklasifikasikan data santri yang belum pernah ditemui sebelumnya dalam proses pembelajaran.

Melalui skenario pengujian kedua ini, efektivitas algoritma Decision Tree diukur secara objektif. Hasil evaluasi mendalam mengenai performa model tersebut dirangkum secara detail pada tabel berikut:

Tabel 5. 5 Hasil Pengujian Model *Decision Tree* Skenario 2

Status	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Support**
Lulus	0.82	0.80	0.66	0.72	125
Putus Sekolah		0.83	0.91	0.87	228
Macro Avg		0.81	0.78	0.79	335
Weighted Avg		0.82	0.82	0.81	335

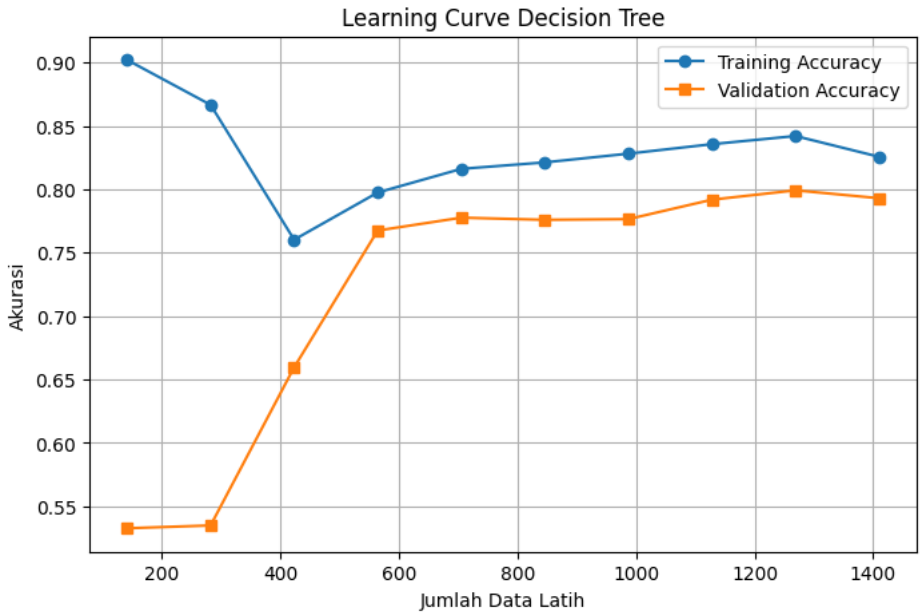
Tabel 5.5 menjelaskan bahwa nilai presisi untuk kelas Lulus sebesar 0,85, yang berarti dari seluruh santri yang diprediksi lulus, sebanyak 85% benar-benar lulus sesuai dengan data aktual. Sementara itu, nilai presisi untuk kelas Putus Sekolah sebesar 0,78, yang menunjukkan bahwa dari seluruh santri yang diprediksi putus sekolah, sebanyak 78% benar-benar putus sekolah.

Nilai recall untuk kelas Lulus sebesar 0,53, yang berarti model hanya mampu mengenali 53% dari seluruh santri yang sebenarnya lulus. Dengan demikian, masih terdapat sebagian santri yang sebenarnya lulus namun diklasifikasikan salah sebagai putus sekolah. Sebaliknya, nilai recall untuk kelas Putus Sekolah mencapai 0,95, yang menunjukkan bahwa model mampu mengenali hampir seluruh santri yang benar-benar putus sekolah.

5.3.3 Pelatihan Model Skanario 3

Pada implementasi model pada skenario 3, distribusi dataset diatur dengan proporsi 70% sebagai data training dan 30% sebagai data *testing*. Alokasi dataset ini dipilih secara strategis untuk memperkaya referensi model dalam mengekstraksi pola-pola klasifikasi secara lebih mendalam dan komprehensif.

Konfigurasi algoritma Decision Tree pada tahap ini dioptimalkan menggunakan parameter spesifik, dengan max depth 4, *min_samples_leaf* sebesar 15 dan *min_samples_split* 5. Untuk memvalidasi efektivitas skenario ini, dilakukan analisis melalui visualisasi *Learning Curve*. Pendekatan ini bertujuan untuk memantau bagaimana performa model berkembang seiring dengan penambahan jumlah data pelatihan, sekaligus mendeteksi potensi terjadinya overfitting atau underfitting pada hasil akhir.



Gambar 5. 10 *Learning Curve* Model *Decision Tree* Skenario 3

Grafik *learning curve* Gambar 5.10 memberikan gambaran komprehensif mengenai hubungan antara penambahan volume data latih dengan tingkat akurasi yang dicapai model. Pada awal fase dengan jumlah data latih yang sedikit, akurasi pelatihan berada di titik tertinggi (sekitar 0.90). Seiring bertambahnya jumlah data latih di atas 600, kemampuan model untuk memprediksi santri yang lulus dan santri yang berisiko putus sekolah meningkat data hingga mencapai angka 1400, kurva cenderung mengalami kesesimbangan di kisaran 0.82 hingga 0.84. Sedangkan Kurva validasi menunjukkan peningkatan yang sangat signifikan seiring bertambahnya jumlah data latih. Dimulai dari akurasi di bawah 0.55, performa model meningkat tajam dan akhirnya mencapai titik stabil pada kisaran 0.78 hingga 0.80. antara kurva pelatihan dan kurva validasi semakin menyempit seiring dengan bertambahnya jumlah data pada titik antara jumlah data 12.000 dengan training Accuracy 0.83. Model mencapai titik optimal ketika jumlah data latih berada di atas 12.00, di mana kedua kurva mulai bergerak sejajar. Kondisi ini menggambarkan bahwa model *Decision Tree* pada pelatihan skenario 3 dalam posisi stabil.

Selanjutnya, Sebagai dasar dalam proses pembentukan struktur pohon keputusan, dilakukan perhitungan nilai *Entropy* dan *Information Gain* untuk setiap atribut. Hasil perhitungan tersebut disajikan dalam Tabel 5.3 dan digunakan untuk menentukan atribut dengan tingkat pemisahan terbaik.

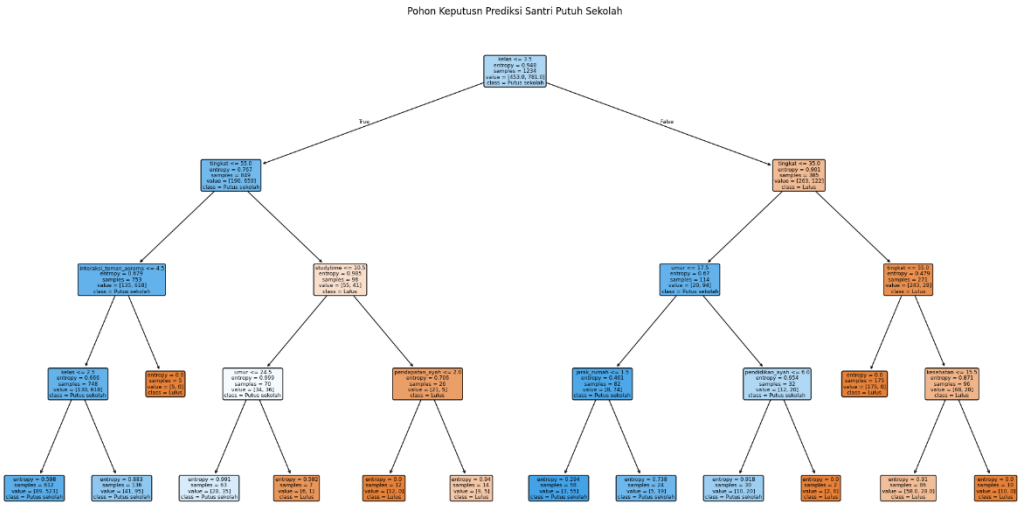
Tabel 5. 6 Nilai *Entropy* dan *Information Gain* – Rasio 70:30

Node	Feature	Entropy	Information Gain
0	kelas	0.948418	0.139601
1	tingkat	0.767033	0.053915
2	interaksi_teman_asrama	0.678506	0.016613
3	kelas	0.666317	0.016262
7	studytime	0.984604	0.064584
8	umur	0.999411	0.048275
11	pendapatan_ayah	0.706274	0.199966
14	tingkat	0.900960	0.365099
15	umur	0.669996	0.070333
16	jarak_rumah	0.461216	0.037410
19	pendidikan_ayah	0.954434	0.093532
22	tingkat	0.479435	0.170937
24	kesehatan	0.870864	0.055344

Tabel ini merinci *Information Gain* serta tingkat ketidakpastian (*Entropy*) pada setiap node keputusan untuk skenario rasio 70:30. Berdasarkan nilai pada tabel 5.6, variabel tingkat pada *node* 14 mencatatkan nilai *Information Gain* tertinggi sebesar 0.365099. Hal ini menunjukkan bahwa jenjang pendidikan atau tingkatan kelas merupakan pembeda paling krusial dalam memprediksi status santri Lulus dan Putus Sekolah. Fitur pendapatan ayah pada *node* 11 dan pendidikan ayah *node* 19 juga menunjukkan pengaruh yang signifikan dengan nilai perolehan informasi masing-masing sebesar 0.199966 dan 0.093532. Ini mengindikasikan bahwa kondisi latar belakang keluarga memiliki korelasi kuat terhadap keberlanjutan pendidikan santri.

Berdasarkan hasil pelatihan model *Decision Tree* dengan parameter *max depth* sebesar 4, *min_samples_leaf* sebesar 15, dan *min_samples_split* 5 dengan

data latih 70% dari *dataset* dengan jumlah data latih 1.234. selanjutnya dibentuk struktur pohon keputusan yang digunakan dalam proses klasifikasi



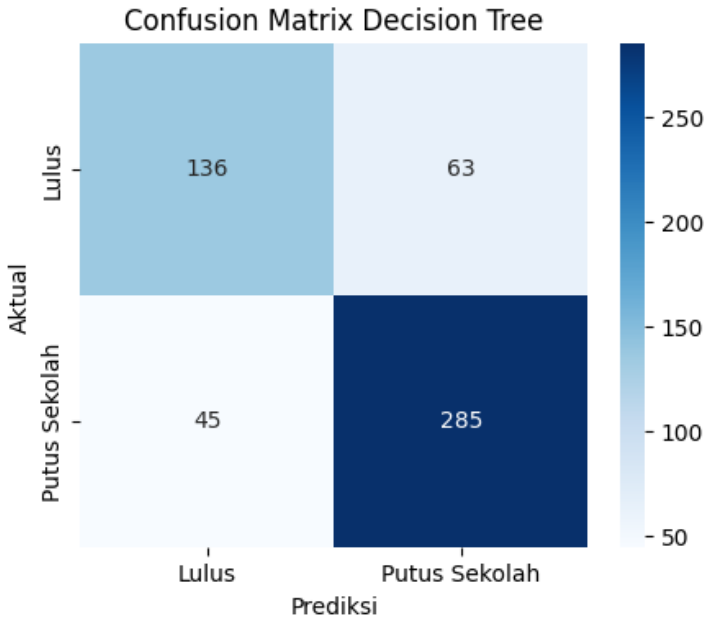
Gambar 5. 11 Struktur Pohon Keputusan Model Skenario 3

Pada Model Skenario 3 ini, model dilatih menggunakan 1.234 sampel data. Meskipun jumlah data latih lebih sedikit dibandingkan skenario sebelumnya, pola utama yang terbentuk tetap menunjukkan konsistensi pada variabel akademik. *Node* 0 sebgai *root n*ode. Atribut kelas dengan aturan ≤ 3.5 tetap menjadi pemisah utama dengan nilai Entropi 0,948 menunjukkan tingkat ketidakpastian awal yang tinggi bagi 1.234 santri sebelum dilakukan pemisahan. Setelah dipisahkan oleh kelas, atribut tingkat ≤ 55.0 dan Atribut interaksi teman asrama ≤ 4.5 pada *node* 1 dan 2 menjadi faktor krusial berikutnya yang menentukan dominasi status santri putus sekolah.

Sedangkan entropi tertinggi terjadi pada n*ode* 5 yaitu atribut umur sebesar 0,999, yang menandakan bahwa pada kelompok 70 santri tersebut, sebaran antara yang lulus dan putus sekolah hampir seimbang, sehingga model

memerlukan kriteria tambahan untuk memutuskannya. Sementara Entropi terendah terdapat pada *node* 9 yaitu atribut jarak rumah sebesar 0,461 dan *node* 11 atribut tingkat sebesar 0,479.

Setelah seluruh parameter model seperti *learning curve* dan struktur pohon dan entropy dan *information gain* terbentuk, kinerja sistem diuji menggunakan data pengujian sebanyak 530 sampel. Pengujian ini menerapkan metode *Confusion Matrix* guna memperoleh gambaran yang objektif mengenai ketepatan klasifikasi model pada masing-masing kategori, baik untuk santri lulus maupun putus sekolah.



Gambar 5. 12 *Matrix Confusion* Pengujian Data Latih Model Skenario 3

Hasil evaluasi menggunakan *Confusion Matrix* pada model Skenario 3 (70:30) terhadap 530 data uji menunjukkan tingkat validitas yang tinggi. Model berhasil mengklasifikasikan 285 santri ke dalam kategori 'Putus Sekolah' secara tepat dan 136 santri pada kategori 'Lulus'. Dengan total 421 prediksi

yang akurat, model ini membuktikan ketangguhannya dalam mengenali pola data santri meskipun menggunakan porsi data pelatihan yang lebih sedikit dibandingkan skenario sebelumnya. Hal ini menegaskan bahwa atribut akademik seperti kelas dan tingkat yang muncul pada pohon keputusan pada Gambar 5.12 memiliki relevansi kuat terhadap kondisi aktual santri.

Kemudian setelah melalui fase pelatihan model diselesaikan menggunakan proporsi data 70%, tahapan berikutnya adalah validasi melalui pengujian terhadap 30% sisa data. Langkah ini krusial untuk mengukur keandalan serta kemampuan generalisasi model dalam mengklasifikasikan data santri yang belum pernah terlibat dalam proses pembelajaran sebelumnya. Melalui skenario pengujian ini, efektivitas model *Decision Tree* dievaluasi secara objektif, di mana hasil pengukuran performa model tersebut dirangkum secara mendalam pada tabel berikut.

Tabel 5. 7 Hasil Pengujian Model *Decision Tree* Skenario 3

Status	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Support**
Lulus	0.80	0.75	0.68	0.72	199
Putus Sekolah		0.82	0.86	0.84	330
Macro Avg		0.79	0.77	0.78	529
Weighted Avg		0.79	0.80	0.79	529

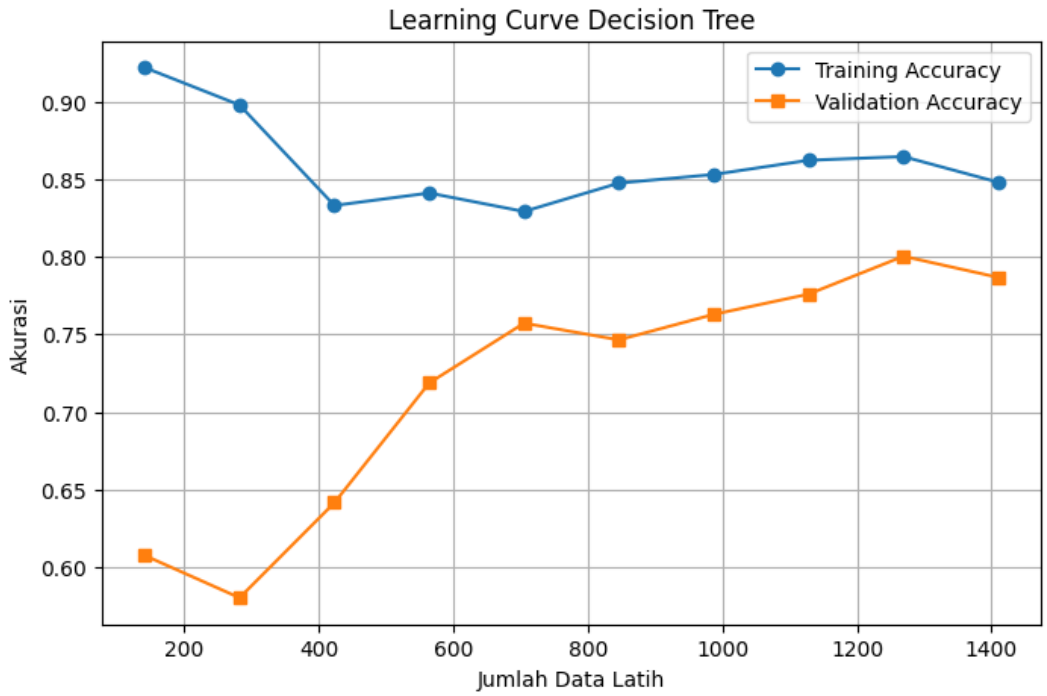
Hasil pengujian model skenario 3 dengan rasio 70:30 dan pembatasan parameter *max depth* pada level 4 menghasilkan performa yang seimbang. Model mencapai tingkat akurasi 0,80 untuk kategori Lulus dan 0,82 untuk kategori Putus Sekolah. Secara rata-rata akurasi model berada pada angka 0,79**.** Model pada skenario ini terlihat pada nilai *Recall* kategori Putus Sekolah yang

mencapai 0,86, menandakan efektivitas model dalam menangkap sebagian besar data santri putu sekolah pada 330 sampel data uji kategori tersebut. Meskipun porsi data latih dikurangi menjadi 70%, penerapan parameter *min samples leaf* sebesar 15 terbukti mampu menjaga stabilitas model dalam menggeneralisasi pola data santri secara objektif dan valid.

5.3.4 Pelatihan Model Skanario 4

Pada skenario ini, porsi data uji ditingkatkan secara signifikan untuk menguji ketangguhan model pada volume data yang lebih besar. Data latih 60% Sekitar 1.058 sampel digunakan oleh model untuk mempelajari pola prediksi data santri lulus dan putus sekolah. sedangkan data uji 40% Sekitar 706 sampel digunakan untuk melakukan validasi akhir. Porsi ini merupakan tantangan terbesar bagi model karena referensi belajar yang lebih sedikit dibandingkan skenario-skenario sebelumnya

Pada tahap ini, Model *Decision Tree* dikonfigurasi menggunakan *Criterion Entropy* dengan parameter kontrol berupa *max_depth* sebesar 6, *min_samples_leaf* sebanyak 8, dan *min_samples_split* sejumlah 6. Penyesuaian parameter ini dilakukan untuk mengoptimalkan kemampuan model dalam mengenali pola data yang lebih spesifik meskipun porsi data pelatihan berkurang, sekaligus menjaga agar model tetap memiliki kemampuan eneralisasi yang stabil saat dihadapkan pada 706 data uji yang belum pernah dipelajari sebelumnya. Pada model skenario 4 visualisasi *Learning Curve* sebagai dilakukan untuk memantau perilaku pelatihan model.



Gambar 5. 13 *Learning Curve* Model *Decision Tree* Skenario 4

Berdasarkan hasil visualisasi *learning curve* gambar 5.13, bagaimana model *Decision Tree* beradaptasi terhadap pembagian data pelatihan sebesar 60% untuk mengenali pola santri yang lulus maupun putus sekolah. Pada awal fase pada *training accuracy* jumlah data < 400, model menunjukkan akurasi yang sangat tinggi di atas 0.90, namun kemudian mengalami penyesuaian hingga stabil di kisaran 0.83 - 0.85. Hal ini menandakan bahwa penggunaan parameter *max_depth* sebesar 6 efektif dalam mencegah model menghafal data secara berlebihan (*overfitting*), sehingga tetap fleksibel terhadap variasi data baru. jarak terlihat antara kurva pelatihan dan kurva validasi semakin menyempit saat jumlah data latih mendekati angka maksimal yaitu 1400. Fenomena ini menunjukkan bahwa model telah mencapai titik stabilitas yang baik. Dengan konfigurasi parameter yang digunakan, model *Decision Tree* mampu membedakan karakteristik santri yang lulus dan putus sekolah.

Setelah mengevaluasi performa model melalui *learning curve*, langkah krusial berikutnya adalah membedah mekanisme internal pohon keputusan dalam mengolah data. Guna memberikan transparansi terhadap proses klasifikasi, tahapan ini akan menyajikan nilai *Entropy* dan *Information Gain* untuk setiap atribut yang digunakan.

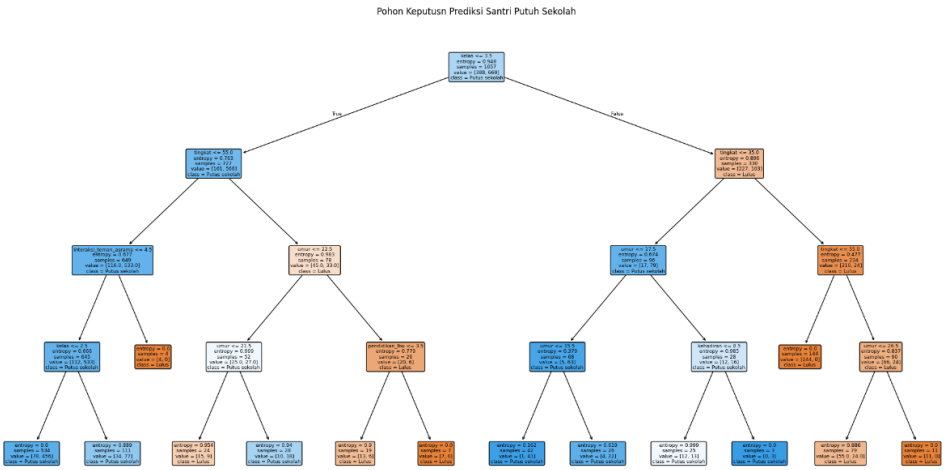
Penyajian data pada tabel berikut berfungsi untuk memetakan variabel mana yang memiliki kontribusi paling signifikan dalam membedakan santri yang lulus dan putus sekolah. Dengan memahami besaran nilai perolehan informasi dapat mengidentifikasi fitur utama yang menjadi dasar pengambilan keputusan model dalam membagi data ke dalam kelompok yang lebih homogen.

Tabel 5. 8 Nilai *Entropy* dan *Information Gain* – Rasio 60:40

Node	Feature	Entropy	Information Gain
0	kelas	0.948401	0.144124
1	tingkat	0.762821	0.052733
2	interaksi_teman_asrama	0.677305	0.015438
3	kelas	0.665972	0.016334
7	umur	0.982859	0.057120
8	umur	0.998933	0.052117
11	pendidikan_Ibu	0.779350	0.121845
14	tingkat	0.895607	0.361348
15	umur	0.673655	0.117867
16	umur	0.378959	0.041876
19	kehadiran	0.985228	0.093402
22	tingkat	0.477071	0.155286
24	umur	0.836641	0.059032

Berdasarkan hasil pengolahan data pada model skenario 1 rasio 60:40, Tabel 5.8 menyajikan rincian nilai *Entropy* dan *Information Gain* yang menjadi basis pembentukan pohon keputusan, di mana atribut tingkat pada *node* 14 muncul sebagai fitur paling berpengaruh dengan nilai perolehan informasi tertinggi sebesar 0.361348. Dominasi variabel akademik juga terlihat pada node 22 tingkat dan *node* 0 kelas yang memiliki kontribusi signifikan dalam membagi data ke dalam kelompok yang lebih homogen, sementara atribut umur muncul berulang kali di berbagai node 7, 8, 15, 16, dan 24 sebagai parameter pendukung untuk mempertajam klasifikasi. Selain itu, keterlibatan faktor eksternal seperti pendidikan Ibu pada n*ode* 11 dengan nilai *Information Gain* 0.121845 serta tingkat kehadiran pada *node* 19 menunjukkan bahwa model secara sistematis mengintegrasikan aspek akademik, latar belakang keluarga, dan kedisiplinan santri untuk memprediksi santri putus sekolah secara akurat.

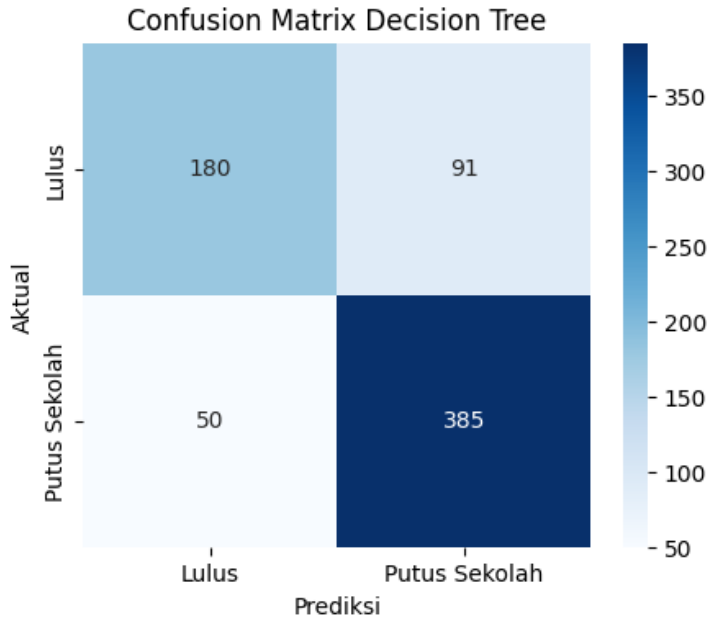
Setelah proses pelatihan selesai menggunakan 1.058 data latih 60% dari total dataset, model berhasil membentuk struktur pohon keputusan yang menjadi dasar logika klasifikasi. Dengan parameter max_depth=6, pohon ini memiliki kedalaman maksimal enam tingkatan, yang memungkinkannya menangkap hubungan antar fitur secara lebih detail dibandingkan skenario sebelumnya.



Gambar 5. 14 Struktur Pohon Keputusan Model Skenario 4

Gambar 5.14 menjelaskan bahwa *Root Node* 0 Pohon dimulai dengan fitur kelas pada ambang batas ≤ 3.5$ sebagai pembeda utama. *node* ini memiliki *entropy* tertinggi sebesar 0.948 dengan total 1.057 sampel, yang menunjukkan bahwa fitur kelas adalah variabel paling berpengaruh dalam menentukan apakah seorang santri berisiko putus sekolah atau tetap lulus. Atribut interaksi teman asrama muncul sebagai faktor penentu pada *node* 2, menunjukkan memiliki dampak terhadap dominasi kelas Putus Sekolah dengan *entropy* 0.677. Sedangkan pada *node* 6, pendidikan_ibu ≤ 3.5 menjadi indikator penting bagi santri yang cenderung masuk ke kategori Lulus. Faktor Usia: atribut umur muncul berulang kali pada *node* 8, 9, dan 12, menandakan adanya korelasi kuat antara usia santri dengan tingkat kelulusan. Dengan dominasi kelas yang bervariasi antara Lulus dan Putus Sekolah pada tiap *node*, model menunjukkan kemampuan adaptasi yang baik. Hal ini didukung oleh *validation accuracy* yang menyentuh angka 80% pada grafik pelatihan pada gambar 5.13

Setelah proses pelatihan model rampung dan struktur pohon serta parameter optimasi seperti entropy dan information gain terbentuk, langkah selanjutnya adalah melakukan pengujian formal. Proses pengujian ini menerapkan metode Confusion Matrix dengan menyiapkan data uji sebanyak 40% dari *dataset* sebesar 706 data. Penggunaan data yang belum pernah dilihat model selama pelatihan ini bertujuan untuk memperoleh gambaran yang utuh serta objektif mengenai ketepatan klasifikasi model pada masing-masing kategori, baik untuk santri lulus maupun putus sekolah.



Gambar 5. 15 *Matrix Confusion* Pengujian Data Latih Model Skenario 4

Berdasarkan hasil pengujian yang disajikan pada Gambar 5.15, dilakukan evaluasi terhadap 706 data testing mewakili 40% dari total *dataset*. Tahap akhir dari eksperimen skenario ini adalah membedah performa prediksi menggunakan metode *Confusion Matrix*. Model berhasil mendeteksi secara akurat sebanyak 385 santri yang memang benar-benar berada dalam kategori

putus sekolah.dan berhasi mengklasifikasikan 180 santri yang lulus dengan tepat. Namun, terdapat kesalahan prediksi pada 91 data santri lulus yang diprediksi putus sekolah dan 50 santri putus sekolah yang luput dari prediksi model. Hasil model menunjukkan bahwa model memiliki kemampuan yang lebih kuat dalam membedakan pola santri putus sekolah dibandingkan santri yang lulus, sesuai dengan tujuan utama klasifikasi.

Tabel 5. 9 Hasil Pengujian Model *Decision Tree* Skenario 4

Kelas	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Support**
Lulus	0.78	0.73	0.66	0.70	271
Putus Sekolah		0.80	0.85	0.83	435
Macro Avg		0.77	0.76	0.76	706
Weighted Avg		0.78	0.78	0.78	706

Berdasarkan hasil evaluasi kinerja model Decision Tree tabel 5.9, diperoleh nilai akurasi sebesar 0,78, yang menunjukkan bahwa sebagian mayoritas data dapat diklasifikasikan dengan benar. Pada kelas Lulus, model menghasilkan nilai precision sebesar 0,73, recall 0,66, dan F1-score 0,70, yang mengindikasikan bahwa masih terdapat sebagian santri lulus yang belum berhasil dikenali secara optimal.

Sementara, pada kelas Putus Sekolah, model menunjukkan performa yang lebih baik dengan nilai precision 0,80, recall 0,85, dan F1-score 0,83, menandakan kemampuan model yang tinggi dalam mengidentifikasi santri yang berpotensi putus sekolah.

5.4 Tuning Parameter Decision Tree

Pada fase ini dilakukan proses tuning parameter pada metode D*ecision Tree*. Tujuan dilakukannya *tuning parameter* adalah untuk menemukan keseimbangan optimal antara performa prediksi dan kemampuan generalisasi model, sehingga model tidak mengalami overfitting maupun underfitting, sebagaimana dikemukakan oleh Breiman (2001).

Parameter tuning *Decision Tr*ee yang digunakan dalam penelitian ini ditampilkan pada Tabel 5.10. Parameter-parameter tersebut berperan dalam mengontrol kompleksitas model dan kualitas pemisahan data

Tabel 5. 10 Parameter Tuning *Decision Tree*

Parameter	Nilai	Fungsi / Keterangan
criterion	Gini, Entropy	Menentukan metode pengukuran kualitas pemisahan node. Opsi: `gini` (Gini Impurity) dan `entropy` (Information Gain).
max_depth	3,5,7,100 dan none	Menentukan kedalaman maksimum pohon. Semakin besar nilainya, pohon semakin kompleks dan berisiko overfitting.
min_samples_split	2,5,10	Jumlah minimum sampel yang diperlukan untuk membagi (split) sebuah node internal.
min_samples_leaf	1,2,3,5	Jumlah minimum sampel yang harus ada pada node daun. Digunakan untuk meningkatkan generalisasi model.

Dengan menyetel parameter dan nilai seperti *max_depth*, dan *min_samples_split*, dapat mengontrol kompleksitas setiap pohon keputusan sehingga model tidak hanya sekadar menghafal data latihan, tetapi juga mampu memberikan prediksi yang akurat ketika menghadapi data baru yang belum

pernah dilihat sebelumnya Probst *et al.* (2019). Selain itu, proses tuning ini berfungsi untuk mengoptimalkan penggunaan sumber daya komputasi dan menyesuaikan sensitivitas model terhadap ketidakseimbangan kelas (*imbalanced* data) melalui pengaturan bobot, sehingga hasil akhir model menjadi lebih stabil dan reliabel dalam mengklasifikasikan data yang tidak seimbang Pratama (2020).

Selanjutnya, untuk memperoleh performa model *Decision Tree* yang optimal dalam memprediksi status Lulus dan Putus Sekolah, dilakukan serangkaian pengujian menggunakan metode *parameter tuning*, untuk mencari kombinasi parameter terbaik, dengan mengaur *max_depth* dan kriteria pembagian data, untuk menghasilkan tingkat akurasi maksimal. Pada fase pengujian ini proses *tuning parameter* diujicoba menggunakan metode *GridSearchCV*, sebuah teknik pencarian parameter secara sistematis dengan menguji seluruh kombinasi parameter yang telah ditentukan melalui proses *cross-validation*.

Kemudian setelah megatur seluruh kebutuhan parameter dan menentukan data uji, diperoleh 10 kombinasi parameter yang diujikan, sebagaimana disajikan pada tabel 5.11.

Tabel 5. 11 Hasil Tuning Parameter Model *Decision Tree*

No	Criterion	Max Depth	Min Samples Split	Min Samples Leaf	Accuracy
1	entropy	3	100	20	**0.819858**
2	gini	3	100	20	**0.819858**
3	gini	3	20	20	0.817730
4	gini	3	50	20	0.817730
5	entropy	3	20	20	0.817730
6	entropy	3	50	20	0.817730
7	entropy	3	100	10	0.817021
8	gini	3	100	10	0.817021
9	gini	3	50	10	0.814894
10	gini	3	20	10	0.814894

Hasil pengujian tuning parameter dengan metode *GridSearchCV*, diperoleh 10 kombinasi parameter sebagaimana diuraikan pada tabel 5.11. Model dengan performa terbaik diperoleh pada *criterion* entropy dan gini dengan max_depth = 3, min_samples_split = 100, dan min_samples_leaf = 20, yang sama-sama menghasilkan akurasi tertinggi sebesar 0,819858. Hal ini menunjukkan bahwa pohon dengan kedalaman dangkal dan jumlah sampel minimum yang besar mampu menghasilkan generalisasi yang lebih baik serta mengurangi risiko overfitting.

Langkah selanjutnya mengimplementasikan tuning parameter dengan nilai yang telah ditentukan secara otomatis menggunakan *GridSearchCV* pada metode *Decision Tree* dalam meprediski santri putus sekolah

```
=== PARAMETER TERBAIK ===
{'criterion': 'gini', 'max_depth': 3, 'min_samples_leaf': 20, 'min_samples_split': 100}

Akurasi Model Terbaik: 0.8130

              precision    recall  f1-score   support

       Lulus       0.79      0.68      0.73       130
Putus sekolah      0.83      0.89      0.86       223

    accuracy                           0.81       353
   macro avg       0.81      0.78      0.79       353
weighted avg       0.81      0.81      0.81       353
```

Gambar 5. 16 Hasil Evaluasi *Decision Tree* dengan Tuning Parameter

Hasil pengujian model *Decision Tree* setelah melakukan tuning parameter sebagaimana disajikan pada Gambar 5.16 menunjukkan bahwa model dengan parameter terbaik. Model Decision Tree dengan parameter terbaik yaitu criterion = gini, max_depth = 3, min_samples_split = 100, dan min_samples_leaf = 20 menghasilkan akurasi sebesar 81,30%. Konfigurasi ini menunjukkan bahwa pohon keputusan yang sederhana mampu memberikan performa yang stabil dan menghindari overfitting.

Berdasarkan classification report, kelas Putus Sekolah memiliki performa lebih baik dengan recall 0,89 dan F1-score 0,86, yang berarti model sangat baik dalam mendeteksi santri yang berisiko putus sekolah. Sementara itu, kelas Lulus memiliki recall 0,68, menandakan masih terdapat sebagian santri lulus yang salah dalam prediksi. menunjukkan bahwa model cukup stabil dan efektif dalam melakukan klasifikasi potensi putus sekolah santri.

5.5 Pengujian Model Decision Tree

Tahap evaluasi dilakukan dengan menyeleksi hasil prediksi optimal dari seluruh skenario pengujian metode *Decision Tree.* Mengingat setiap skenario menerapkan kombinasi parameter dan rasio pembagian *dataset* yang berbeda, maka performa model yang dihasilkan tentunya bervariasi.

Dari keempat skenario tersebut, model dengan nilai akurasi tertinggi ditentukan sebagai representasi performa terbaik untuk metode *Decision Tree*. Capaian optimal skenario selanjutnya digunakan sebagai parameter pembanding terhadap metode lain dalam penelitian ini. Ringkasan evaluasi hasil terbaik dari setiap skenario disajikan sebagai berikut

Tabel 5. 12 Evaluasi Performance Metode *Decision Tree*

Skenario	Status	Accuracy	Precision	Recall	F1-Score	Support
1	Lulus	0.79	0.75	0.62	0.68	65
	Putus Sekolah		0.81	0.88	0.84	113
2	Lulus	0.82	0.80	0.66	0.72	125
	Putus Sekolah		0.83	0.91	0.87	228
3	Lulus	0.80	0.75	0.68	0.72	199
	Putus Sekolah		0.82	0.86	0.84	330
4	Lulus	0.78	0.73	0.66	0.70	271
	Putus Sekolah		0.80	0.85	0.83	435

Berdasarkan data pada Tabel 5.12, berikut adalah penjelasan mengenai evaluasi performa metode *Decision Tree* dari empat skenario yang telah diuji. Performa tertinggi, skenario 2 menunjukkan hasil paling optimal dengan akurasi sebesar 0,82. Pada skenario ini, kemampuan model dalam

mengidentifikasi santri "Putus Sekolah" mencapai nilai *precision* tertinggi sebesar 0,83 dan *recall* 0,91, yang menghasilkan *F1-score* sebesar 0,87.

Sedangkan konsistensi klasifikasi, Secara umum, model di seluruh skenario menunjukkan performa yang lebih unggul dalam mendeteksi kelas Putus Sekolah dibandingkan kelas Lulus. Hal ini terlihat dari nilai *recall* dan *F1-score* yang secara konsisten lebih tinggi pada kategori Putus Sekolah di setiap skenario.

Sementara kestabilan model, pada skenario 3 memiliki akurasi sebesar 0,80 dengan nilai *F1-score* yang seimbang, 0,72 untuk Lulus dan 0,84 untuk Putus Sekolah, model skenario 3 menunjukkan kemampuan generalisasi yang cukup stabil dengan jumlah support data yang lebih besar dibandingkan skenario 1 dan 2. Sedangkan Skenario dengan data terbanyak yaitu skenario 4. Meskipun memiliki jumlah *support* tertinggi total 706 data, skenario 4 mencatatkan akurasi terendah sebesar 0,78. Hal ini sesuai dengan pengamatan pada confusion matrix sebelumnya yang menunjukkan adanya 91 kesalahan prediksi pada kelas Lulus dan 50 pada kelas Putus Sekolah.

Kesimpulan akhir pada model *Decision Tree* dari hasil evaluasi tersebut, Skenario 2 dipilih sebagai representasi terbaik metode *Decision Tree* karena memiliki tingkat akurasi dan keseimbangan metrik (*precision, recall, f1-score*) yang paling kompetitif. Struktur pohon pada skenario ini terbukti paling efektif dalam mengekstrak informasi dari fitur-fitur dominan seperti kelas, tingkat pendidikan, dan umur.

BAB VI

PEMBAHASAN

6.1 Performa Pengujian Random Forest dan Decision Tree

Dalam penelitian ini, proses pengujian dilakukan melalui pendekatan eksperimental dengan memvariasikan rasio pembagian data latih dan data uji, mulai dari proporsi 90:10 hingga 60:40. Langkah ini bertujuan untuk mengevaluasi stabilitas dan konsistensi performa algoritma *Decision Tree* (DT) dan *Random Forest* (RF) dalam menghadapi volume data yang berbeda.

Eksperimen dimulai dengan Rasio 90:10 pada model skenario 1, di mana model diberikan akses maksimal terhadap data pelatihan untuk mengidentifikasi pola dasar. Meskipun menunjukkan akurasi tinggi pada fase awal, pengujian pada porsi data uji yang kecil cenderung menghasilkan variansi yang fluktuatif. Selanjutnya, pada Rasio 80:20 dan 70:30 pada skenario 2 & 3, dilakukan penambahan porsi data uji secara bertahap. Hasilnya menunjukkan adanya konvergensi pada *Learning Curve*, yang menandakan peningkatan kemampuan generalisasi model seiring dengan bertambahnya sampel evaluasi.

Tahap akhir menggunakan rasio 60:40 pada Skenario 4, yang memberikan tantangan pengujian paling berat dengan porsi data uji sebesar 40% 706 data. Pada tahap ini, model diuji tingkat objektivitasnya dalam mengklasifikasikan data yang belum pernah ditemui. Melalui *Confusion Matrix*, terlihat bahwa model mampu mempertahankan akurasi yang stabil meskipun dihadapkan pada

porsi data uji yang lebih besar.Penilaian terhadap performa kedua metode tersebut dilakukan secara komprehensif berdasarkan empat indikator, *Accuracy, Precision, Recall,* dan *F1-Score*, Merupakan perpaduan antara Precision dan Recall yang memberikan ukuran keseimbangan antara ketepatan dan kelengkapan hasil prediksi

Hasil keseluruhan dari rangkaian pengujian tersebut direkapitulasi dalam tabel berikut:

Tabel 6. 1 Perbandingan Performa Keseluruhan Skenario

Model	**Skenario**	**Status**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Support**
Decision Tree	**1**	Lulus	0.79	0.75	0.62	0.68	65
		Putus Sekolah		0.81	0.88	0.84	113
	2	Lulus	0.82	0.80	0.66	0.72	125
		Putus Sekolah		0.83	0.91	0.87	228
	3	Lulus	0.80	0.75	0.68	0.72	199
		Putus Sekolah		0.82	0.86	0.84	330
	4	Lulus	0.78	0.73	0.66	0.70	271
		Putus Sekolah		0.80	0.85	0.83	435
Random Forest	1	Lulus	0.83	0.87	0.63	0.73	65
		Putus Sekolah		0.82	0.95	0.88	112
	2	Lulus	0.81	0.87	0.62	0.72	130
		Putus Sekolah		0.81	0.95	0.87	223
	3	Lulus	0.82	0.83	0.65	0.73	194
		Putus Sekolah		0.82	0.92	0.87	335
	4	Lulus	0.84	0.86	0.66	0.75	259
		Putus Sekolah		0.83	0.94	0.88	447

Berdasarkan ringkasan hasil pengujian pada tabel 6.1. Secara komprehensif, model Random Forest menunjukkan keunggulan performa yang signifikan dibandingkan dengan Decision Tree dalam mengklasifikasikan data santri. Berdasarkan hasil pengujian yang telah dipaparkan, keunggulan ini terlihat jelas pada beberapa aspek krusial:

1. Ketepatan Klasifikasi (Accuracy); Model *Random Forest* mencapai nilai akurasi tertinggi sebesar 0,84 Skenario 4, mengungguli capaian terbaik *Decision Tree* yang berada di angka 0,82 pada Skenario 2.

2. Keandalan Prediksi Positif (Precision): Pada kelas Putus Sekolah, *Random Forest* secara konsisten menghasilkan nilai *precision* yang lebih tinggi dan stabil berkisar 0,82 - 0,83 dibandingkan *Decision Tree.*

3. Kemampuan Menangkap Data (Recall): Random Forest unggul dalam meminimalisir kesalahan deteksi dini dengan nilai *recall* mencapai 0,94 hingga 0,95, memastikan sangat sedikit santri berisiko yang terlewat oleh sistem.

4. Keseimbangan Metrik (F1-Score): Sebagai perpaduan antara ketepatan dan kelengkapan, nilai *F1-Score Random Forest* yang menyentuh 0,88 pada kelas Putus Sekolah menunjukkan bahwa model ini adalah yang paling seimbang dan optimal untuk diimplementasikan.

Kesimpulan akhir, model *Random Forest* menunjukkan keunggulan performa yang signifikan dibandingkan dengan *Decision Tree* dalam memprediksi data santri baik lulus maupun putus sekolah.

6.2 Perbandingan Metode Menggunakan *Confusion Matrix*

Evaluasi mendalam dilakukan melalui pembedahan nilai matriks konfusi untuk meninjau distribusi prediksi benar dan kesalahan klasifikasi (*misclassification*) pada setiap skenario. Matriks ini menjadi instrumen krusial untuk mengukur sensitivitas model terhadap masing-masing kelas target. Hasil performa pengujian *model* berdasarkan *Confusion Matrix* sebai berikut.

Tabel 6. 2 Performa Pengujian *Model* berdasarkan *Confusion Matrix*

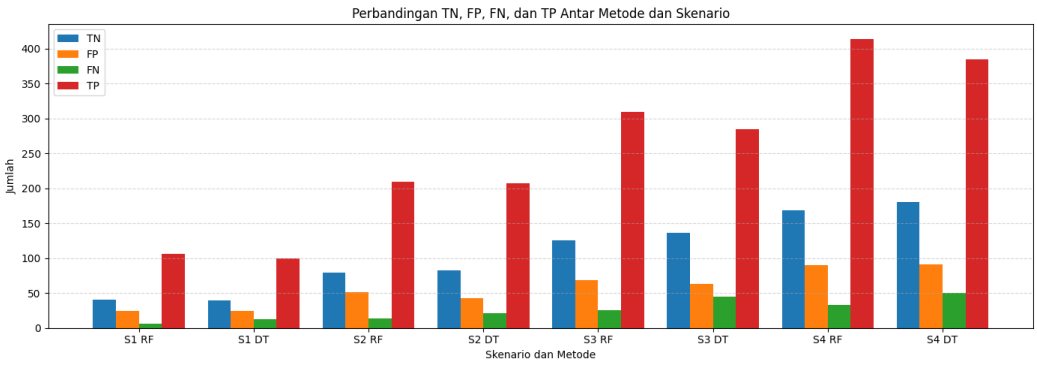
Model	Skenario	kelas	Prediksi Lulus	Prediksi Putus Sekolah
Decision Tree	1	Aktual Lulus	40	24
		Aktual Putus Sekolah	13	100
	2	Aktual Lulus	82	43
		Aktual Putus Sekolah	21	207
	3	Aktual Lulus	136	63
		Aktual Putus Sekolah	45	285
	4	Aktual Lulus	180	91
		Aktual Putus Sekolah	50	385
Random Forest	1	Aktual Lulus	41	24
		Aktual Putus Sekolah	6	106
	2	Aktual Lulus	79	51
		Aktual Putus Sekolah	14	209
	3	Aktual Lulus	126	68
		Aktual Putus Sekolah	26	309
	4	Aktual Lulus	169	90
		Aktual Putus Sekolah	33	414

Model *Random Forest* menunjukkan keunggulan yang lebih superior dalam mereduksi tingkat kesalahan klasifikasi, khususnya pada nilai *False Negative* santri putus sekolah yang diprediksi lulus Pada Skenario 4, *Random Forest*

mencatatkan angka *True Positive* tertinggi sebesar 414 data, jauh melampaui *Decision Tree* sebesar 385 data. *Random Forest* berhasil menekan angka data santri putus sekolah yang luput dari prediksi hingga hanya 33 data, sementara *Decision Tree* mencatat 50 data yang luput.

Mekanisme *ensemble* yang menggabungkan banyak pohon keputusan pada *Random Forest* terbukti lebih efektif dalam menangani ambiguitas data dibandingkan *Decision Tree* tunggal. Hal ini terlihat dari konsistensi *Random Forest* yang tetap mampu mendeteksi kategori "Aktual Putus Sekolah" dengan presisi tinggi di setiap skenario.

Secara metodologis, *Random Forest* memiliki kapabilitas generalisasi yang lebih kuat, terbukti dari rendahnya nilai *False Negative* di seluruh skenario dibandingkan *Decision Tree*. Seiring bertambahnya jumlah data uji pada Skenario 1 ke Skenario 4, *Random Forest* menunjukkan peningkatan performa yang lebih stabil, sedangkan DT mengalami peningkatan kesalahan klasifikasi *misclassification* yang lebih cepat



Gambar 6. 1 Hasil *Confusion Matrix Random Forest* dan *Decision Tree*

Gambar 6.1 menyajikan representasi visual dari distribusi metrik *Confusion Matrix*, yang mencakup *True Negative* (TN), *False Positive* (FP),

False Negative (FN), dan *True Positive* (TP) untuk seluruh skenario pengujian pada metode *Random Forest* (RF) dan *Decision Tree* (DT). Diagram batang ini memberikan simulasi komparatif mengenai efektivitas kedua model seiring dengan peningkatan volume data dari Skenario 1 hingga Skenario 4.

6.3 Implementasi Metode Terbaik

Berdasarkan hasil analisis komparatif, penelitian ini mengimplementasikan algoritma *Random Forest* sebagai metode terbaik untuk diterapkan dalam sistem prediksi status santri Lulus maupun Putus Sekolah. Pemilihan ini didasarkan pada capaian akurasi tertinggi sebesar 0,84 serta kemampuan deteksi yang sangat tajam terhadap kategori santri berisiko putus sekolah, yang dibuktikan dengan nilai *recall* mencapai 0,94 dan *F1-score* sebesar 0,88. Keunggulan model *Random Forest* dalam meminimalkan kesalahan prediksi dibandingkan *Decision Tree* menunjukkan bahwa mekanisme *ensemble learning* pada *Random Forest* lebih stabil dan andal dalam menangani kompleksitas data, sehingga sangat layak dijadikan instrumen pengambilan keputusan strategis bagi pihak pondok pesantren sidogiri dalam prediksi santri putus sekolah.

6.4 Prediksi Putus Sekolah dalam Pandangan Islam

Dalam perspektif Islam, pendidikan merupakan pilar fundamental dalam pembentukan karakter dan akhlak seorang muslim. Al-Qur'an dan hadis secara tegas menekankan urgensi menuntut ilmu sebagai jalan untuk meraih kemuliaan, sebagaimana firman Allah dalam QS. Al-Mujādalah [58]:11:

﴿ يَرْفَعِ اللَّهُ الَّذِينَ آمَنُوا مِنكُمْ وَالَّذِينَ أُوتُوا الْعِلْمَ دَرَجَاتٍ ﴾

Artinya*"Allah akan meninggikan derajat orang-orang yang beriman di antara kamu dan orang-orang yang diberi ilmu pengetahuan beberapa derajat."*

Ayat QS. Al-Mujādalah [58]:11 menegaskan bahwa Allah meninggikan derajat orang-orang yang beriman dan berilmu. Pesan ini menunjukkan bahwa pendidikan bukan hanya sarana intelektual, tetapi juga jalan spiritual untuk memperoleh kemuliaan di sisi Allah. Dalam konteks penelitian prediksi santri putus sekolah, ayat ini mengandung makna bahwa menjaga keberlanjutan pendidikan santri adalah bagian dari upaya menjaga derajat dan masa depan mereka, baik secara duniawi maupun ukhrawi. Oleh karena itu, penggunaan teknologi seperti machine learning untuk mendeteksi santri yang berisiko putus sekolah bukan sekadar aktivitas teknis, tetapi merupakan ikhtiar proaktif agar tidak ada santri yang terlepas dari kesempatan meraih ilmu dan kemuliaan yang dijanjikan Allah. Dengan mendeteksi risiko lebih awal, pesantren dapat memberikan pendampingan, bimbingan, dan intervensi yang tepat sehingga santri tetap berada pada jalur pendidikan.

Pentingnya pendidikan juga ditegaskan dalam sabda Nabi Muhammad ﷺ yang diriwayatkan oleh Ibnu Majah (hadis sahih menurut sebagian ulama):

طَلَبُ الْعِلْمِ فَرِيضَةٌ عَلَى كُلِّ مُسْلِمٍ

Artinya *"*Menuntut ilmu adalah kewajiban bagi setiap muslim.*"*

Selain itu, Rasulullah ﷺ bersabda dalam hadis sahih riwayat Muslim:

مَنْ سَلَكَ طَرِيقًا يَلْتَمِسُ فِيهِ عِلْمًا سَهَّلَ اللَّهُ لَهُ بِهِ طَرِيقًا إِلَى الْجَنَّةِ

Artinya *"Barang siapa menempuh jalan untuk mencari ilmu, Allah akan mudahkan baginya jalan menuju surga."*

Ayat dan hadis-hadis tersebut menegaskan bahwa menjaga keberlangsungan proses pendidikan merupakan kewajiban *syar'i* sekaligus jalan menuju kemuliaan dunia dan akhirat. Dalam konteks penelitian prediksi santri putus sekolah, nilai-nilai ini memberi pesan jelas bahwa upaya mencegah terjadinya putus sekolah merupakan bagian dari menjaga derajat dan masa depan santri. Ketika seorang santri berhenti menuntut ilmu, ia kehilangan kesempatan meraih keutamaan yang Allah dan Rasul-Nya janjikan.

Sejalan dengan perkembangan teknologi, Penerapan *machine learning* algoritma *Random Forest* dan *Decision Tree* merupakan manifestasi dari konsep al-akhdzu bil-asbāb, yaitu ikhtiar maksimal menggunakan sarana mutakhir untuk meraih kemaslahatan. Dalam kerangka *Maqāṣid al-Syarī'ah*, upaya ini bagian implementasi nyata dari *ḥifẓ al-'aql* (menjaga akal) dan ḥifẓ al-nasl (menjaga keberlangsungan generasi). Dengan mendeteksi risiko santri putus sekolah secara dini, pesantren menjalankan peran preventif untuk memastikan santri tidak kehilangan kesempatan meraih derajat mulia yang dijanjikan Allah bagi para penuntut ilmu

Dengan demikian, penerapan sistem prediksi berbasis data di pesantren bukan hanya sebagai bentuk pemanfaatan teknologi baru, tetapi juga sebagai cara yang didukung oleh bukti empirik untuk menjaga agar proses pendidikan

santri tetap berjalan baik. Langkah ini menjadi ikhtiar agar santri tetap fokus belajar, berkembang secara akademik, dan menjalankan nilai-nilai ilmu yang diajarkan dalam Al-Qur'an dan Sunnah Nabi ﷺ.

Dengan berkembangnya teknologi *machine learning* terkait prediksi membantu pesantren membuat keputusan dengan lebih cepat dan tepat, sehingga proses pembinaan dapat dilakukan secara lebih efektif dan memastikan santri tidak putus sekolah sebelum waktunya lulus.

BAB VII

KESIMPULAN DAN SARAN

7.1 Kesimpulan

Berdasarkan hasil penelitian, pengujian, dan analisis yang telah dilakukan mengenai prediksi santri putus sekolah di Pondok Pesantren Sidogiri, dapat disimpulkan bahwa model klasifikasi prediktif *Decision Tree* dan *Random Forest* berhasil mengidentifikasi santri yang berpotensi mengalami putus sekolah

Hasil pengujian menunjukkan bahwa model prediksi memiliki tingkat akurasi dan kinerja yang baik dalam mendeteksi santri yang berisiko putus sekolah, sehingga dapat digunakan sebagai alat bantu pengambilan keputusan. Evaluasi kinerja model berdasarkan metrik akurasi seperti *precision, recall,* atau *F1-score* menunjukkan bahwa pendekatan klasifikasi yang diterapkan efektif dalam konteks data santri Pondok Pesantren Sidogiri periode 2019–2023.

Selain Itu, penelitian ini berhasil mengimplementasikan model *Decision Tree* dan *Random Forest* untuk mendeteksi risiko santri putus sekolah dengan performa yang tinggi. Pengujian melalui empat skenario rasio data (90:10 hingga 60:40) membuktikan bahwa ketersediaan data uji yang lebih besar memberikan gambaran performa yang paling objektif dan stabil.

Perbandingan performa secara komparatif, algoritma Random Forest menunjukkan keunggulan dibandingkan Decision Tree di hampir seluruh

metrik evaluasi. Model terbaik dicapai oleh *Random Forest* dengan nilai Accuracy sebesar 0,84, Precision 0,83, Recall 0,94, dan F1-Score 0,88. Sementara itu, performa tertinggi Decision Tree hanya mencapai akurasi 0,82 pada Skenario 2.

Selain itu, penelitian ini juga mengidentifikasi beberapa faktor dominan yang berkontribusi terhadap terjadinya putus sekolah, salah satu variabel yang memiliki pengaruh dalam prediksi santri putus sekolah di pondok pesantren sidogiri adalah kehadiran, faktor kelas dan umur seperti hasil pembahasan di bab empat yang menjadi hasil pengujian *Random Forest* pada skenario ke satu sampai pengujian ke tiga. variabel kelas memiliki pengaruh terbesar dalam menentukan prediksi status santri, dengan nilai kontribusi tertinggi sebesar 0.199. Faktor berikutnya yang juga berperan signifikan adalah kehadiran (0.127) dan umur (0.126), yang menegaskan bahwa kedisiplinan hadir dan kondisi usia menjadi indikator kuat terhadap risiko putus sekolah.

7.2 Saran

Merujuk hasil penelitian ini, meskipun metode *Random Forest* menunjukkan performa yang baik dan akurat dalam memprediksi risiko siswa putus sekolah di Madrash Miftahul Ulum Pondok Pesantren Sidogiri, penelitian selanjutnya disarankan untuk mencoba dan membandingkannya dengan algoritma lain. Penggunaan metode alternatif yang memiliki mekanisme pengambilan keputusan berbeda akan membantu mengidentifikasi algoritma yang paling sesuai dengan karakteristik data dropout santri.

Selain itu, penelitian lanjutan perlu melibatkan lebih banyak instansi pesantren maupun madrasah agar pola-pola putus sekolah yang diperoleh menjadi lebih beragam, variatif dan representatif. Integrasi data tambahan seperti riwayat bimbingan konseling, tingkat kehadiran, kondisi keluarga, dan catatan perilaku juga berpotensi meningkatkan sensitivitas dan akurasi model dalam mendeteksi santri berisiko *dropout*. Melalui perluasan cakupan data dan eksplorasi algoritma tambahan tersebut, penelitian mendatang diharapkan mampu menghasilkan model prediksi yang lebih akurat, adaptif, dan dapat digunakan sebagai instrumen pendukung keputusan yang lebih komprehensif bagi lembaga pendidikan pesantren maupun madrasah dalam mencegah terjadinya putus sekolah sebelum akhirnya lulus.

DAFTAR PUSTAKA

Kemper, Lorenz, Gerrit Vorhoff, and Berthold U. Wigger. 2020. "Predicting Student Dropout: A Machine Learning Approach." *European Journal of Higher Education* 10 (1): 28–47. https://doi.org/10.1080/21568235.2020.1718520.

Pérez, Boris, Camilo Castellanos, and Darío Correal. 2018. "Predicting Student Drop-Out Rates Using Data Mining Techniques: A Case Study." In Applications of Computational Intelligence, edited by Alvaro David Orjuela-Cañón, Juan Carlos Figueroa-García, and Julián David Arias-Londoño, 833:111–25. Communications in Computer and Information Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-03023-0_10.

Sani, Nor Samsiah, Ahmad Fikri, Zulaiha Ali, Mohd Zakree, and Khairul Nadiyah. 2020. "Drop-Out Prediction in Higher Education Among B40 Students." *International Journal of Advanced Computer Science and Applications* 11 (11). https://doi.org/10.14569/IJACSA.2020.0111169.

Hegde, Vinayak. 2016. "Dimensionality Reduction Technique for Developing Undergraduate Student Dropout Model Using Principal Component Analysis through R Package." In 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 1–6. Chennai, India: IEEE. https://doi.org/10.1109/ICCIC.2016.7919670.

Hoe, Alan Cheah Kah, Mohd Sharifuddin Ahmad, Tan Chin Hooi, Mohana Shanmugam, Saraswathy Shamini Gunasekaran, Zaihisma Che Cob, and Ammuthavali Ramasamy. 2013. "Analyzing Students Records to Identify Patterns of Students' Performance." In 2013 International Conference on Research and Innovation in Information Systems (ICRIIS), 544–47. Kuala Lumpur, Malaysia: IEEE. https://doi.org/10.1109/ICRIIS.2013.6716767.

Utari, Meylani, Budi Warsito, and Retno Kusumaningrum. 2020. "Implementation of Data Mining for Drop-Out Prediction Using Random Forest Method." In 2020 8th International Conference on Information and Communication Technology (ICoICT), 1–5. Yogyakarta, Indonesia: IEEE. https://doi.org/10.1109/ICoICT49345.2020.9166276.

Timaran Pereira, Ricardo, and Javier Caicedo Zambrano. 2017. "Application of Decision Trees for Detection of Student Dropout Profiles." In 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 528–31. Cancun, Mexico: IEEE. https://doi.org/10.1109/ICMLA.2017.0-107.

Harwati, Riezky Ikha Virdyanawaty, and Agus Mansur. 2016. "Drop out Estimation Students Based on the Study Period: Comparisonbetween Naïve Bayes and Support Vector Machines Algorithm Methods." IOP

Conference Series: Materials Science and Engineering 105 (January): 012039. https://doi.org/10.1088/1757-899X/105/1/012039.

Fernandez-Garcia, Antonio Jesus, Juan Carlos Preciado, Fran Melchor, Roberto Rodriguez-Echeverria, Jose Maria Conejero, and Fernando Sanchez-Figueroa. 2021. "A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data." IEEE Access 9: 133076–90. https://doi.org/10.1109/ACCESS.2021.3115851.

Sa'ad, Muhammad Ibnu, Kusrini, and M. Syukri Mustafa. 2020. "Student Prediction of Drop Out Using Extreme Learning Machine (ELM) Algorithm." In *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*, 1–6. Manado, Indonesia: IEEE. https://doi.org/10.1109/ICORIS50180.2020.9320831.

Masserini, Lucio, and Matilde Bini. 2021. "Does Joining Social Media Groups Help to Reduce Students' Dropout within the First University Year?" Socio-Economic Planning Sciences 73 (February): 100865. https://doi.org/10.1016/j.seps.2020.100865.

Tayebi, Abdelhamid, Josefa Gomez, and Carlos Delgado. 2021. "Analysis on the Lack of Motivation and Dropout in Engineering Students in Spain." IEEE Access 9: 66253–65. https://doi.org/10.1109/ACCESS.2021.3076751.

Devasia, Tismy, Vinushree T P, and Vinayak Hegde. 2016. "Prediction of Students Performance Using Educational Data Mining." In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), 91–95. Ernakulam, India: IEEE. https://doi.org/10.1109/SAPIENCE.2016.7684167.

J. Chen et al., "Pre-evacuation Time Estimation Based Emergency Evacuation Simulation in Urban Residential Communities," IJERPH, vol. 16, no. 23, p. 4599, Nov. 2019, doi: 10.3390/ijerph16234599.

A. M. Mariano, A. B. D. M. L. Ferreira, M. R. Santos, M. L. Castilho, and A. C. F. L. C. Bastos, "Decision trees for predicting dropout in Engineering Course students in Brazil," Procedia Computer Science, vol. 214, pp. 1113–1120, 2022, doi: 10.1016/j.procs.2022.11.285.

K. Schouten, F. Frasincar, and R. Dekker, "An Information Gain-Driven Feature Study for Aspect-Based Sentiment Analysis," in Natural Language Processing and Information Systems, vol. 9612, E. Métais, F. Meziane, M. Saraee, V. Sugumaran, and S. Vadera, Eds., in Lecture Notes in Computer Science, vol. 9612. , Cham: Springer International Publishing, 2016, pp. 48–59. doi: 10.1007/978-3-319-41754-7_5.

Millatinnafi'ah, I. and Claretta, D. 2024. Strategi Pondok Pesantren Sidogiri dalam Membentuk Karakter Santri. *JIIP - Jurnal Ilmiah Ilmu Pendidikan*. 7, 9 (Sep. 2024), 9899-9905. DOI:https://doi.org/10.54371/jiip.v7i9.5869.

Usman, A., Fatkhurrohman, F., & Nasokah, N. (2023). *Konsep Tafaqquh Fiddīn dalam Sistem Pendidikan Pesantren (Kajian QS. At-Taubah: 122)*. Al Jabiri: Jurnal Ilmiah Studi Islam, 1(2), 141–154.

A. Purwanto, B. Sartono, and K. A. Notodiputro, “A Comparison Of Random Forest And Double Random Forest: Dropout Rates Of Madrasah Students In Indonesia,” Barekeng: J. Math. & App., vol. 19, no. 1, pp. 227–236, Jan. 2025, doi: 10.30598/barekengvol19iss1pp227-236.

Ma'ruf, Mohammad. (2018). Eksistensi Pondok Pesantren Sidogiri Pasuruan Dalam Mempertahankan Nilai-Nilai Salaf Di Era Globalisasi. journal EVALUASI. 1. 167. 10.32478/evaluasi.v1i2.71.

Mohd, R. F., Zulkifli, H., Hamzah, M. I., & Tamuri, A. H. (2024). Lifelong Learning among Islamic Education Teachers. International Journal of Academic Research in Business and Social Sciences, 14(8), 1064–1076. http://dx.doi.org/10.6007/IJARBSS/v14-i8/22479

Bichri, Houda & Chergui, Adil & Mustapha, Hain. (2024). Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets. International Journal of Advanced Computer Science and Applications. 15. 10.14569/IJACSA.2024.0150235.

Taser, Pelin Yildirim “Application of Bagging and Boosting Approaches Using Decision Tree-Based Algorithms in Diabetes Risk Prediction,” in The 7th International Management Information Systems Conference, MDPI, Mar. 2021, p. 6. doi: 10.3390/proceedings2021074006.

Aaboub F, Chamlal H, Ouaderhman T. Statistical analysis of various splitting criteria for decision trees *. Journal of Algorithms & Computational Technology. 2023;17. doi:10.1177/17483026231198181

Sun, H., Zhang, J., & Li, W. (2019). *Attribute Selection Based on Constraint Gain and Depth in Decision Tree Algorithms. Entropy*, 21(2), 198.

Vrigazova, B**.** (2021). *The Proportion for Splitting Data into Training and Test Set for the Bootstrap in Classification Problems*. *Business Systems Research*, 12(1), 228–242. https://doi.org/10.2478/bsrj-2021-0015

Cui, G., Liu, Z., Zhao, L., & Wang, H. (2025). *The Machine Learning Algorithm Based on Decision Tree for Sprint Pattern Recognition*. *PLOS ONE*, 20(3), e0317414. https://doi.org/10.1371/journal.pone.0317414

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Pratama, A. R. (2020). Optimasi Hyperparameter Random Forest Menggunakan GridSearchCV pada Klasifikasi Data Imbalanced. Jurnal Rekayasa Sistem & Industri, 7(1), 25-32. https://doi.org/10.25124/jrsi.v7i1.391

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), e1301. https://doi.org/10.1002/widm.1301