

MENINGKATKAN MODEL PREDIKSI KELULUSAN SANTRI TAHFIDZ DI PONDOK PESANTREN AL-KAUTSAR MENGGUNAKAN ALGORITMA RANDOM FOREST

Syahrul Sobari¹, Ade Irma Purnamasari², Agus Bahtiar³, Kaslani⁴

^{1,2,3,4}Teknik Informatika, STMIK IKMI Cirebon; Jl. Perjuangan No.10B, Karyamulya, Kesambi, Kota Cirebon, Jawa Barat 45131, Indonesia.; (0231) 490480

Received: 16 Desember 2024

Accepted: 14 Januari 2025

Published: 20 Januari 2025

Keywords:

Random Forest, Prediksi kelulusan, Tahfidz, Pondok Pesantren, Machine Learning

Correspondent Email:

syahrulsobari29@gmail.com

Abstrak. Penelitian ini bertujuan untuk meningkatkan model prediksi kelulusan santri Tahfidz di Pondok Pesantren Al-Kautsar dengan menerapkan algoritma *Random Forest*, yang dikenal memiliki kemampuan menangani data kompleks dan menghasilkan prediksi yang lebih akurat dibandingkan metode tradisional. Dengan algoritma *Random Forest*, penelitian ini berusaha mengeksplorasi keunggulan dalam memberikan prediksi yang lebih baik. Penelitian ini melibatkan beberapa tahapan penting, dimulai dari pengumpulan data santri yang berjumlah 300 dengan mencakup beberapa atribut, seperti jumlah hafalan, tingkat kehadiran, nilai ujian, dan status kelulusan sebagai target prediksi. Hasil penelitian menunjukkan bahwa algoritma *Random Forest* secara signifikan lebih unggul dibandingkan regresi linier dalam memprediksi kelulusan santri. *Random Forest* menghasilkan akurasi sebesar 99,64%, presisi 100,00%, dan recall 98,80%, sementara regresi linier hanya mencapai akurasi sebesar 77%. Analisis mendalam menunjukkan bahwa atribut jumlah hafalan dan tingkat kehadiran memiliki pengaruh terbesar terhadap prediksi kelulusan, diikuti oleh nilai ujian. Hal ini menegaskan bahwa faktor non-akademik seperti hafalan dan kehadiran memiliki peran penting dalam keberhasilan santri. Keunggulan algoritma *Random Forest* terletak pada kemampuannya menggabungkan prediksi dari banyak decision tree, sehingga menghasilkan model yang lebih stabil dan tahan terhadap noise dalam data. Selain itu, algoritma ini bisa menangani data dengan kombinasi variabel numerik dan kategorikal tanpa perlu transformasi yang signifikan, menjadikannya cocok untuk dataset dunia nyata yang seringkali tidak sempurna.

Abstract. This research aims to improve the prediction model of Tahfidz santri graduation at Al-Kautsar Islamic Boarding School by applying the *Random Forest* algorithm, which is known to have the ability to handle complex data and produce more accurate predictions than traditional methods. With the *Random Forest* algorithm, this research seeks to explore the advantages in providing better predictions. This research involves several important stages, starting from the collection of 300 santri data that includes several attributes, such as the number of memorization, attendance rate, test scores, and graduation status as prediction targets. The results show that the *Random Forest* algorithm is significantly superior to linear regression in predicting student graduation. *Random Forest* produced an accuracy of 99.64%, precision of 100.00%, and recall of 98.80%, while linear regression only achieved an accuracy of 77%. In-depth analysis shows that the attributes of memorization and attendance rate have the greatest influence on predicting

graduation, followed by test scores. This confirms that non-academic factors such as memorization and attendance have an important role in the success of students. The advantage of the Random Forest algorithm lies in its ability to combine predictions from many decision trees, resulting in a model that is more stable and resistant to noise in the data. In addition, it can handle data with a combination of numerical and categorical variables without the need for significant transformation, making it suitable for real-world datasets that are often imperfect.

1. PENDAHULUAN

Di berbagai bidang kehidupan, seperti bisnis, kesehatan, dan pendidikan, kemajuan teknologi informatika saat ini telah membawa perubahan besar. khususnya pembelajaran mesin, telah digunakan dalam pendidikan formal untuk memecahkan masalah yang kompleks, seperti menemukan dan memprediksi kinerja akademik siswa. Algoritma *Random Forest* termasuk salah satu metode yang dapat memberikan prediksi dengan hasil akurat, terutama ketika menangani data besar dan kompleks. Kemampuan kognitif, dan metode pembelajaran adalah beberapa dari banyak faktor yang memengaruhi tingkat kelulusan santri di program Tahfidz pesantren. Oleh karena itu, diharapkan penggunaan algoritma *Random Forest* sebagai model prediksi kelulusan santri Tahfidz akan membawa solusi yang berguna untuk meningkatkan kualitas pendidikan di pondok pesantren [1], [2]

Pembelajaran mesin memiliki potensi besar untuk meningkatkan prediksi kinerja akademik. [1] menekankan bahwa pembelajaran mesin dapat membuat prediksi yang lebih akurat daripada pendekatan tradisional [2] [3]. Menemukan bahwa sistem pembelajaran mesin otomatis efektif dalam memperkirakan prestasi siswa dan memungkinkan intervensi yang tepat waktu. Selain itu, [4] membahas manfaat pendidikan terstruktur bagi kemampuan kognitif siswa Tahfidz. Namun, penelitian ini belum mempelajari algoritma *Random Forest* di lingkungan pesantren, dan masih terbatas pada lingkungan pendidikan umum.

Tujuan utama dari penelitian ini adalah untuk mengembangkan model prediksi yang dapat meningkatkan akurasi dalam menentukan keberhasilan santri Tahfidz di Pondok Pesantren dengan menggunakan algoritma *Random Forest*. Penelitian ini berusaha untuk mengisi celah dalam literatur tentang

penggunaan algoritma pembelajaran mesin dalam pendidikan keagamaan, khususnya dalam konteks pendidikan Tahfidz, yang memiliki karakteristik dan kebutuhan khusus. Diharapkan bahwa penelitian ini akan menemukan komponen penting yang mempengaruhi kelulusan santri secara lebih mendalam dengan menggunakan *Random Forest*. Faktor-faktor seperti ini masih belum banyak diteliti dalam konteks pesantren [1], [4]. Penelitian, ini dapat membantu mengembangkan sistem pendidikan pesantren berbasis data, yang akan memungkinkan intervensi yang lebih tepat dan mendukung pengambilan keputusan yang berbasis bukti. Secara praktis, penelitian ini dapat digunakan oleh pengelola pesantren untuk meningkatkan efisiensi program Tahfidz, menyediakan bimbingan yang sesuai dengan kebutuhan santri, dan memaksimalkan kemampuan mereka untuk menyelesaikan program.

Jika penelitian ini berhasil mencapai tujuan utamanya, hasil dari penelitian tersebut akan sangat bermanfaat untuk meningkatkan pemahaman dan penggunaan pembelajaran mesin, khususnya algoritma *Random Forest*, dalam konteks pendidikan keagamaan seperti pondok pesantren. Hasil penelitian ini dapat membantu pesantren menemukan komponen utama yang memengaruhi keberhasilan santri Tahfidz. Kemampuan kognitif, aspek emosional, dan latar belakang demografis adalah beberapa komponen yang selama ini sulit dinilai secara menyeluruh [5]. [6] Hal ini akan memungkinkan guru menggunakan data untuk membuat keputusan yang lebih baik tentang manajemen dan peningkatan program Tahfidz. Penelitian ini membantu para peneliti di bidang informatika mengembangkan model prediktif yang dapat digunakan dalam konteks data multidimensi dan kompleks lainnya. Ini akan meningkatkan metode untuk menganalisis data pendidikan. Penelitian ini juga membuka

peluang untuk mengembangkan alat evaluasi otomatis dan prediktif yang lebih akurat. Alat ini dapat digunakan oleh praktisi pendidikan di berbagai lembaga yang sebanding. Pada skala lebih luas, hasil ini dapat mendorong adopsi teknologi prediktif yang lebih efektif dalam berbagai bidang pendidikan, berkontribusi pada peningkatan kualitas pendidikan berbasis teknologi di era digital. Secara keseluruhan, temuan ini dapat menjadi landasan bagi pengembangan teknologi yang semakin responsif terhadap kebutuhan institusi pendidikan yang dinamis dan kompleks.

2. TINJAUAN PUSTAKA

2.1 *Random Forest*

Random Forest adalah algoritma berbasis ensemble yang menggabungkan prediksi dari beberapa *decision tree* untuk meningkatkan akurasi dan stabilitas model. Algoritma ini telah digunakan dalam berbagai studi, termasuk prediksi kesehatan [7],[8]. Serta menunjukkan keunggulan algoritma *Random Forest* dalam menangani dataset kesehatan yang kompleks dengan akurasi tinggi.[9]

2.2 *Knowledge Discovery in Database (KDD)*

KDD adalah proses sistematis untuk menemukan informasi berguna dari data. Tahapan KDD meliputi seleksi data, prapemrosesan, transformasi, data mining, dan evaluasi. Penelitian ini menggunakan tahapan KDD untuk memastikan proses analisis data dilakukan secara sistematis dan hasilnya dapat diandalkan.

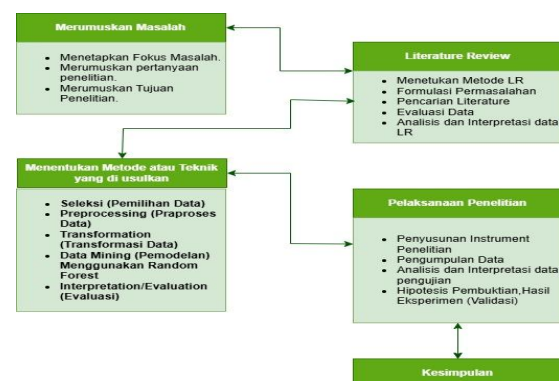
2.3 *Prediksi kelulusan menggunakan Machine learning*

Mengidentifikasi faktor-faktor signifikan yang mempengaruhi kualitas disertai menggunakan algoritma pembelajaran mesin, termasuk *Random Forest*[10]. Studi sebelumnya menunjukkan bahwa algoritma machine learning, seperti *Random Forest*, mampu meningkatkan akurasi prediksi pendidikan dibandingkan dengan metode tradisional. *Random Forest* memiliki keunggulan dalam memprediksi data yang kompleks dan non-linier[11]. Studi oleh [8] menggunakan algoritma C4.5 untuk memprediksi kelulusan

santri berdasarkan atribut seperti hafalan dan kedisiplinan, dan [1] menunjukkan bahwa *Random Forest* lebih baik menangani variabel kompleks dalam prediksi akademik [12].

3. METODE PENELITIAN

Penelitian ini menggunakan metode kuantitatif dengan pendekatan eksperimental untuk menguji efektivitas algoritma *Random Forest* dalam memprediksi Kelulusan santri Tahfidz di Pondok Pesantren Al-Kautsar. Data yang digunakan meliputi variabel seperti, nilai ujian, kehadiran, dan prestasi hafalan Al-Qur'an. Analisis proses meliputi pembersihan data, pemilihan fitur, dan pelatihan model *Random Forest* menggunakan teknik *cross-validation*. Model ini kemudian dievaluasi dengan metrik akurasi, presisi, dan recall, serta dibandingkan dengan metode prediksi lain untuk melihat keunggulannya dalam meningkatkan akurasi prediksi.



Gambar 1 Metodologi Penelitian

3.1 Tahapan Metode *Knowledge in Database (KDD)*

Tahapan-tahapan tersebut meliputi seleksi data, pra-pemrosesan, transformasi, penerapan metode data mining, hingga interpretasi hasil. Dengan menerapkan metode ini, diharapkan dapat diperoleh model prediksi yang akurat dan relevan untuk mendukung analisis kelulusan santri Tahfidz di Pondok Pesantren Al-Kautsar.

3.1.1 *Data Selection*

Penelitian ini menggunakan data santri Tahfidz dari Pondok Pesantren Al-Kautsar. Dataset ini memiliki atribut "Nama Santri", serta beberapa atribut lainnya seperti "Jumlah Hafalan", "Nilai Ujian", dan "Kehadiran",

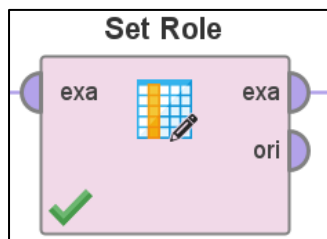
masing-masing atribut menunjukkan kinerja santri. Selain itu, atribut “Status Kelulusan”, yang menunjukkan status kelulusan santri berdasarkan kriteria yang telah ditetapkan.

#	A	B	C	D	E
1	Nama Santri	Jumlah Hafalan	Nilai Ujian	Kehadiran	Status Kelulusan
2	Dina Mardiana	12	86	82	Lulus
3	Nabila Gita Pramesti	10	78	99	Lulus
4	Sintia Dona Nurpadilah	17	56	91	Tidak Lulus
5	Lulis Puspitasari	12	67	99	Lulus
6	Nyinyi Sulistiana	22	68	74	Tidak Lulus
7	Tita Rustiana	29	97	85	Lulus
8	Sri Sawitri	19	99	92	Lulus
9	Shinta Lusiana	15	61	82	Lulus
10	Rismayanti	29	73	94	Lulus
11	Siti Nurazizah Koswara	10	61	78	Lulus
12	Beni Permana	16	87	79	Lulus
13	Anita Puspitasari	16	54	84	Tidak Lulus
14	Jujun C. Iskandar	16	50	85	Tidak Lulus
15	Fitriyadin Muhammad In	30	57	99	Tidak Lulus
16	Agung Gumelar	10	87	80	Lulus
17	Nadia Minakam	17	98	82	Lulus
18	Jatin Sawaludin	21	74	92	Lulus
19	Gita Avisa	11	95	86	Lulus
20	Feryn Febriyani	17	82	72	Tidak Lulus
21	Melan Setiawati	16	54	79	Tidak Lulus
22	Devia Ridwan	23	63	84	Lulus
23	Silva Fauziah Gina Febriya	28	96	100	Lulus

Gambar 2 Dataset

3.1.2 Preprocessing

komponen *Set Role* dalam aplikasi AI Studio, yang digunakan untuk menentukan peran (*role*) dari setiap atribut dalam dataset. Fungsi ini penting dalam tahap praproses data untuk mempersiapkan dataset agar sesuai dengan kebutuhan model.

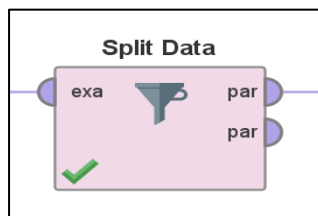


Gambar 3 Tahapan operator set role

Dalam komponen *Set Role*, pengguna dapat menetapkan kolom mana yang akan berfungsi sebagai target atau label (output) yang akan diprediksi, serta kolom-kolom mana yang berperan sebagai atribut (input) untuk membantu prediksi.

3.1.3 Transformation

komponen *Split Data* dalam aplikasi AI Studio, yang digunakan untuk membagi dataset menjadi dua bagian, yaitu data pelatihan (*training data*) dan data pengujian (*testing data*). Tahap ini sangat penting dalam proses pemodelan, karena memungkinkan model untuk dilatih dan diuji dengan data yang berbeda.

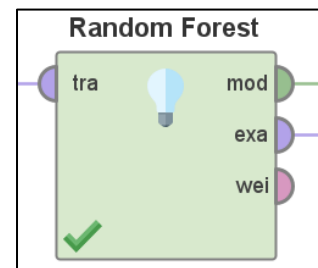


Gambar 4 Split data

Data pelatihan digunakan untuk melatih model agar mengenali pola-pola dalam dataset, sedangkan data pengujian digunakan untuk mengukur kinerja model terhadap data baru yang belum pernah dilihat selama proses pelatihan.

3.1.4 Data Mining

komponen *Random Forest* dalam aplikasi AI Studio, yang digunakan sebagai algoritma pemodelan untuk memprediksi target berdasarkan atribut-atribut yang tersedia.

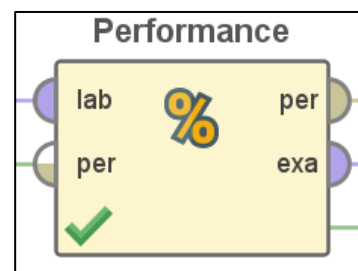


Gambar 5 Komponen Random Forest

Data latih yang telah disiapkan sebelumnya akan dimasukkan ke dalam model *Random Forest* untuk membangun model prediktif. *Random Forest* bekerja dengan membangun beberapa pohon keputusan (*decision trees*) dari sampel acak dalam data dan menggabungkan hasil dari setiap pohon untuk mendapatkan prediksi akhir yang lebih stabil dan akurat. Hal ini membantu mengurangi risiko *overfitting* yang sering terjadi pada model tunggal.

3.1.5 Evaluasi

Selanjutnya Pada tahap *performance* merupakan tahapan yang dapat menampilkan nilai akurasi



Gambar 6 Performance

Tahapan Performance, merupakan tahapan terakhir dari model ini untuk mengukur sejauh mana algoritma *Random Forest* mampu memberikan hasil prediksi yang akurat. Pada tahap ini, metrik evaluasi seperti akurasi, precision, recall, dan AUC dihitung berdasarkan perbandingan antara data aktual (label) dengan prediksi yang dihasilkan oleh model.

4. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan dataset santri Tahfidz di Pondok Pesantren Al-Kautsar yang mencakup atribut "Jumlah Hafalan", "Nilai Ujian", "Kehadiran", dan "Status Kelulusan" sebagai target prediksi. Proses analisis data menggunakan algoritma *Random Forest* melalui tahapan *Knowledge Discovery in Databases* (KDD).

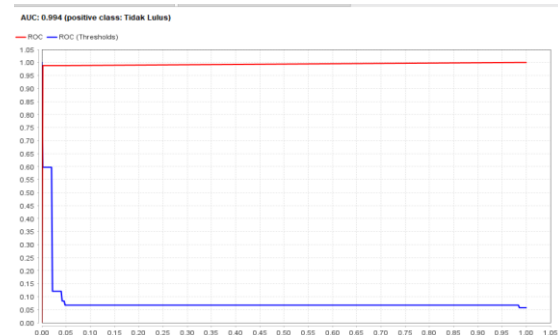
4.1 Hasil Penelitian

accuracy: 99.64%			
	true Lulus	true Tidak Lulus	class precision
pred Lulus	191	1	99.48%
pred Tidak Lulus	0	82	100.00%
class recall	100.00%	98.80%	

Gambar 7 Hasil Accurasy terbesar

Algoritma *Random Forest* menunjukkan akurasi yang sangat baik, dengan tingkat akurasi sebesar 99.64%. Model ini mampu memprediksi seluruh data "Lulus" dengan benar (recall 100%) dan hampir seluruh data "Tidak Lulus" (recall 98.80%). Dari segi presisi, prediksi untuk kelas "Lulus" mencapai 99.48%, sedangkan untuk kelas "Tidak Lulus" presisinya sempurna di angka 100%. Secara detail, model berhasil memprediksi 191 data dari kelas "Lulus" secara benar dan hanya melakukan kesalahan pada 1 data, yang diprediksi sebagai "Lulus" meskipun sebenarnya "Tidak Lulus". Selain itu, tidak ada data kelas "Lulus" yang salah diprediksi sebagai "Tidak Lulus". Hasil ini mengindikasikan bahwa algoritma *Random Forest* bekerja secara efektif dan andal dalam memprediksi kelulusan santri Tahfidz dengan tingkat kesalahan yang

sangat kecil, menjadikannya alat yang kuat untuk mendukung proses evaluasi santri di Pondok Pesantren Al-Kautsar.



Gambar 8 Hasil Kurva ROC-AUC

Hasil kurva ROC-AUC menunjukkan model prediksi dengan algoritma *Random Forest* dalam membedakan antara kelas "Lulus" dan "Tidak Lulus," dengan nilai AUC (Area Under the Curve) sebesar 0.994. Nilai AUC yang mendekati 1 mengindikasikan bahwa model memiliki kemampuan yang sangat baik dalam membedakan kedua kelas. Kurva ROC yang hampir mendekati sudut kiri atas menunjukkan bahwa model mencapai tingkat *True Positive Rate* (TPR) yang tinggi dengan *False Positive Rate* (FPR) yang sangat rendah pada hampir semua ambang batas (*threshold*). Artinya, model dapat secara efektif mengidentifikasi santri yang "Tidak Lulus" tanpa banyak melakukan kesalahan klasifikasi pada santri yang "Lulus." Dengan kinerja ini, model menunjukkan tingkat keandalan yang tinggi dan sangat sesuai untuk digunakan dalam mendukung evaluasi kelulusan santri Tahfidz di Pondok Pesantren Al-Kautsar.

4.2 Pembahasan

- *Random Forest* efektif dalam memprediksi kelulusan santri dengan akurasi sangat tinggi, membuktikan bahwa algoritma ini cocok untuk dataset kompleks seperti pendidikan di pesantren.
- Kombinasi decision tree dalam *Random Forest* meningkatkan

kemampuan model untuk menangkap pola non-linear dari atribut seperti hafalan dan kehadiran.

- Implementasi Model prediksi ini dapat digunakan oleh pesantren untuk mengidentifikasi santri yang memerlukan bimbingan tambahan, meningkatkan efisiensi program Tahfidz.

5. KESIMPULAN

1. Dengan penerapan algoritma Random Forest sangat efektif dalam menghasilkan model kelulusan santri Tahfidz di Pondok Pesantren Al-Kautsar. Dengan akurasi mencapai 99,64%, algoritma ini jauh lebih baik daripada metode lainnya seperti regresi linier yang hanya menghasilkan akurasi sebesar 77%. Hasil ini menggaris bawahi kemampuan *Random Forest* dalam menangani kompleksitas data pendidikan yang melibatkan variabel numerik dan kategorikal. Hasil ini tidak hanya mengkonfirmasi keunggulan teknis algoritma tetapi juga menunjukkan relevansi metode ini dalam konteks pendidikan keagamaan.
2. Langkah-langkah optimalisasi, seperti seleksi fitur dan pemanfaatan teknik *Knowledge Discovery in Databases* (KDD), berhasil meningkatkan model prediksi. Analisis lebih lanjut menunjukkan bahwa atribut seperti jumlah hafalan dan tingkat kehadiran memberikan kontribusi paling besar terhadap akurasi prediksi. Keberhasilan ini menekankan pentingnya kombinasi pendekatan teknis yang canggih dan pemahaman kontekstual terhadap faktor-faktor pendidikan untuk menciptakan model yang lebih akurat dan aplikatif.
3. Model evaluasi menggunakan metrik seperti presisi (100%) dan recall (98,80%) menampilkan bahwa *Random Forest* mampu memberikan prediksi yang stabil dan dapat diandalkan. Hal ini menjadikannya alat yang potensial

untuk mendukung pengambilan keputusan berbasis data di pesantren. Dengan kemampuannya ini, pengelola pesantren dapat merancang strategi pembelajaran yang lebih efektif, seperti memberikan perhatian lebih pada faktor non-akademik yang terbukti signifikan dalam menentukan kelulusan santri. Kesimpulan ini memberikan kontribusi yang signifikan terhadap pemahaman dan penerapan mesin pembelajaran dalam pendidikan pesantren, sekaligus membuka peluang untuk penelitian lanjutan dalam skala yang lebih luas.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada pihak-pihak terkait yang telah memberi dukungan terhadap penelitian ini.

DAFTAR PUSTAKA

- [1] L. Zhao, J. Ren, L. Zhang, and H. Zhao, "Quantitative Analysis and Prediction of Academic Performance of Students Using Machine Learning," *Sustain.*, vol. 15, no. 16, 2023.
- [2] V. Bansal, H. Buckchash, and B. Raman, "Computational Intelligence Enabled Student Performance Estimation in the Age of COVID-19," *SN Comput. Sci.*, vol. 3, no. 1, pp. 1–11, 2022.
- [3] D. Duei Putri, G. F. Nama, and W. E. Sulistiono, "Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier," *J. Inform. dan Tek. Elektro Terap.*, vol. 10, no. 1, pp. 34–40, 2022.
- [4] R. Hidayat, T. I. Syafii, S. Agustian, and M. Alam, "School Management Develop Students' Cognitive Ability Through The Tahfidz Program," *J. Islam. Educ. Ethics*, vol. 1, no. 2, pp. 125–130, 2023.
- [5] D. Masri, P. Prabowo Warsodirejo, and S. Aymanda Nababan, "Correlation Between the Tahfiz Al-Qur'an Program and Students' Emotional Intelligence at MAS Muallimin UNIVA Medan," *TA'DIB J.*, vol. 26, no. 1, 2023.
- [6] R. Rismayati, I. Ismarmiaty, and S. Hidayat, "Ensemble Implementation for Predicting Student Graduation with Classification Algorithm," *Int. J. Eng. Comput. Sci. Appl.*,

- vol. 1, no. 1, pp. 35–42, 2022.
- [7] X. Han *et al.*, “Random forest can accurately predict the development of end-stage renal disease in immunoglobulin a nephropathy patients,” *Ann. Transl. Med.*, vol. 7, no. 11, pp. 234–234, 2019.
 - [8] N. M. Farhan and B. Setiaji, “Sistem Prediksi Kelulusan Santri Tahfidz Qur’an Menggunakan Algoritma C4.5,” *Indones. J. Comput. Sci.*, vol. 12, no. 2, pp. 284–301, 2023.
 - [9] K. Sumwiza, C. Twizere, G. Rushingabigwi, P. Bakunzibake, and P. Bamurigire, “Enhanced cardiovascular disease prediction model using random forest algorithm,” *Informatics Med. Unlocked*, vol. 41, no. August, p. 101316, 2023.
 - [10] Z. Li, Y. Li, and Z. Xie, “Exploring the significant predictors to the quality of master’s dissertations,” *IEEE Access*, vol. 8, pp. 21152–21158, 2020.
 - [11] H. Peni, A. Tjahyaningtijas, N. Husin, and H. Al Habib, “Machine learning on Bibliometric studies academic education :,” vol. 02010, pp. 1–8, 2023.
 - [12] S. Guzmán-Castillo *et al.*, “Implementation of a Predictive Information System for University Dropout Prevention,” *Procedia Comput. Sci.*, vol. 198, no. 2020, pp. 566–571, 2022.