Analisis Komparatif Metode K-*Nearest Neighbors* Dan *Random Forest* Untuk Rekomendasi Pemilihan Pesantren Berdasarkan Profil Minat Siswa (siswa xxx tingkat SD dan SMP)

A Comparative Analysis of K-Nearest Neighbors and Random Forest Methods for Recommendations on Selecting Islamic Boarding Schools Based on Student Interest Profiles (primary and middle school students at xxx)

Mas Nurul Hamidah[1)], **Rahmawati Febrifyaning Tias**[2)] , **Rifki Fahrial Zainal**[3)]

[1), 2),3)] *Program Studi Teknik Informatika, Fakultas Teknik, Universitas Bhayangkara Surabaya*
Jl A.yani, No. 114, Surabaya
Email : nurul@ubhara.ac.id[1)], rahmawati@ubhara.ac.id[2)], rifki@ubhara.ac.id[3)]
Corresponding author. Phone : *+62 85730067739*

Abstrak

KNN dan Random Forest adalah salah satu metode klasifikasi, dalam penelitian ini akan akan membandingkan 2 metode pada machine learning yaitu KNN dan Random forest untuk merekomendasikan jenis pesantren yang berdasarkan pada minat siswa, penerapan komparasi 2 metode klasifikasi dalam sistem rekomendasi pemilihan jenis pesantren berdasarkan minat siswa di Tingkat SD dan SMP Xxx, Adapun jenis pesantrenya adalah salafi,khalafi dan campuran, dengan atribut seperti kecenderungan akademik, minat keagamaan, keterlibatan ekstrakurikuler, dan latar belakang keluarga. penerapan metode machine learning untuk mendukung pengambilan keputusan dalam pemilihan pesantren yang sesuai dengan karakter siswa, yang masih jarang dijumpai pada lembaga pendidikan Islam. Evaluasi performa dilakukan menggunakan metrik Mean Absolute Error (MAE) dan Root Mean Square Error (RMSE). Hasil pengujian menunjukkan bahwa algoritma Random Forest memberikan hasil yang lebih baik dengan MAE sebesar 0.23 dan RMSE sebesar 0.57, dibandingkan KNN yang memiliki MAE sebesar 0.6 dan RMSE sebesar 0.96. Hasil tersebut memperlihatkan Random Forest lebih efektif dalam memberikan rekomendasi pemilihan pesantren yang sesuai, dan dapat dijadikan dasar dalam pengembangan sistem pendukung keputusan bagi sekolah berbasis pesantren.
Kata kunci: *KNN, Machine Learning, Random Forest, Sekolah Berbasis Pesantren*

Abstract

KNN and Random Forest are one of the classification methods, in this study will compare 2 methods in machine learning namely KNN and Random forest to recommend the type of Islamic boarding school based on student interests, the application of a comparison of 2 classification methods in the recommendation system for selecting the type of Islamic boarding school based on student interests at the Elementary and Middle School levels of Xxx, The types of Islamic boarding schools are salafi, khalafi and mixed, with attributes such as academic tendencies, religious interests, extracurricular involvement, and family background. application of machine learning methods to support decision making in selecting Islamic boarding schools that are in accordance with student character, which is still rarely found in Islamic educational institutions. Performance evaluation is carried out using the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) metrics. The test results show that the Random Forest algorithm gives better results with an MAE of 0.23 and an RMSE of 0.57, compared to KNN which has an MAE of 0.6 and an RMSE of 0.96. Thus, Random Forest shown to be more effective in providing recommendations for selecting appropriate Islamic boarding schools, and can be used as a basis for developing a decision support system for Islamic boarding school-based schools.
Keywords: *KNN, Machine Learning, Random Forest, Islamic Boarding School*

1. PENDAHULUAN

Pondok pesantren adalah salah satu lembaga Pendidikan yang menerapkan Pendidikan karakter yang tidak hanya menerapkan nilai nilai keislaman seperti akhlak namun juga membentuk kepribadian dan kedisiplinan siswa [1]. Dengan berkembangnya generasi ke generasi

yang berdasarkan pada kebutuhan dan minat siswa, pesantren pun mengalami diversifikasi dalam model pembelajarannya, di antaranya adalah pesantren *salafiyah* yang fokus pada penguasaan kitab kuning [2], pesantren *khalafiyah* yang mengintegrasikan kurikulum umum dan agama [3], serta model *campuran* yang menggabungkan keduanya [4]. Dalam praktiknya, pemilihan jenis pesantren sering kali belum mempertimbangkan kesesuaian antara karakter siswa dan pendekatan pendidikan yang ditawarkan oleh masing-masing jenis pesantren, sehingga berpotensi menimbulkan ketidakefektifan dalam proses belajar-mengajar [5].

Di masa society 5.0 yang beriringan dengan perkembangan teknologi di bidang informatika, pendekatan *machine learning* adalah salah satu solusi untuk membantuk pemilihan pesantren yang lebih obyektif berdasarkan minat siswa, Adapaun Metode klasifikasi seperti *K-Nearest Neighbors* (KNN) dan *Random Forest* (RF) dapat digunakan untuk merekomendasikan jenis pesantren berdasarkan profil minat siswa, seperti kecenderungan akademik, minat keagamaan, latar belakang keluarga, dan aktivitas ekstrakurikuler. Pemilihan metode KNN dalam penelitian ini didasarkan pada karakteristik algoritma tersebut yang berorientasi pada instance-based learning, dimana proses pembelajaran dilakukan berdasarkan kemiripan antar data. KNN memiliki kemampuan yang unggul dalam mengklasifikasikan data baru dengan cara mengidentifikasi sejumlah tetangga terdekat (*nearest neighbors*) dari data yang telah diketahui kelasnya, kemudian menentukan label berdasarkan mayoritas kelas dari tetangga tersebut. Dalam konteks rekomendasi jenis pesantren, setiap siswa direpresentasikan sebagai suatu vektor fitur yang mencakup profil minat, bakat, preferensi akademik, serta kecenderungan sosial-religius. Dengan demikian, KNN mampu memberikan rekomendasi yang bersifat personal, karena keputusan didasarkan pada kesamaan profil antar siswa. Selain itu, KNN memiliki fleksibilitas tinggi, tidak memerlukan asumsi distribusi data tertentu, serta efektif untuk data dengan dimensi terbatas namun memiliki makna semantik yang kuat antar atribut.

Sementara itu, penggunaan metode RF didasarkan pada kekuatan algoritma tersebut dalam menghasilkan model klasifikasi yang akurat dan stabil melalui pendekatan *ensemble learning*. RF, sebagaimana diperkenalkan oleh Breiman, merupakan penggabungan dari sejumlah *decision tree* yang dilatih menggunakan subset acak dari data dan fitur, sehingga mampu mengurangi varians dan meminimalkan risiko *overfitting* yang umum terjadi pada model tunggal. Dalam konteks sistem rekomendasi pendidikan, algoritma ini sangat sesuai karena mampu menangani data dengan variabel yang heterogen, baik berupa data kategorikal maupun numerik, serta dapat menilai pentingnya setiap fitur (*feature importance*) terhadap hasil prediksi. Kemampuan ini memungkinkan analisis yang lebih mendalam terhadap faktor-faktor yang paling berpengaruh dalam menentukan jenis pesantren yang sesuai dengan minat siswa.

Dari perspektif implementatif, kedua algoritma ini juga memiliki keunggulan dalam hal efisiensi komputasi dan ketersediaan dukungan perangkat lunak. Keduanya telah banyak diimplementasikan dalam pustaka *machine learning modern*, seperti Scikit-learn, Weka, dan TensorFlow, yang memungkinkan integrasi mudah ke dalam sistem rekomendasi berbasis web atau aplikasi pendidikan. Selain itu, baik KNN maupun RF memiliki kemampuan generalisasi yang baik terhadap data baru, sehingga dapat diterapkan pada berbagai populasi siswa dengan latar belakang yang berbeda. Oleh karena itu, pemilihan metode KNN dan RF dalam penelitian ini bukan hanya didasarkan pada kinerja algoritmik, tetapi juga pada kesesuaiannya terhadap konteks permasalahan, kebutuhan interpretasi hasil, serta potensi pengembangan sistem rekomendasi yang adaptif dan berkelanjutan di masa mendatang. Namun, masih perlu dikaji lebih lanjut sejauh mana efektivitas kedua metode tersebut dalam konteks rekomendasi pendidikan berbasis pesantren.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengembangkan model klasifikasi yang mampu merekomendasikan jenis pesantren secara akurat berdasarkan profil siswa SD dan SMP XXX. Penelitian ini juga bertujuan untuk mengevaluasi dan membandingkan kinerja algoritma KNN dan RF menggunakan metrik *Mean Absolute Error* (MAE) dan *Root Mean Square Error* (RMSE), guna mengetahui metode mana yang lebih optimal dalam memberikan rekomendasi yang tepat sasaran. Dengan pendekatan ini, diharapkan dapat terbentuk sistem pendukung keputusan yang berbasis data, akurat, dan relevan bagi lembaga pendidikan Islam dalam proses penempatan santri sesuai potensi dan minatnya.

2. DASAR TEORI

2.1. Pondok Pesantren

Pondok pesantren adalah salah satu lembaga Pendidikan keislaman yang berakar kuat dalam sistem Pendidikan di Indonesia, santri adalah sebutan bagi siswa atau pelajar yang tinggal di asrama pondok, untuk mengikuti proses pembelajaran secara intensif dengan berbagai kegiatan pendukung didalamnya, dan beberapa peraturan yang diterapkan oleh masing masing lembaga pesantren [6]. Umumnya pesantren mengajarkan Pendidikan karakter untuk memperkuat karakter masing masing santri, pengajar di pesantren disebut dengan ustadz atau ustadzah sedangkan pendiri pesantren biasanya disebut dengan kyai sebagai figure sentral dalam pengajaran, Adapun ciri kahs lainya dari pesantren adalah kitab kuning sebagai materi ajar utamanya [7]. Seiring berkembangnya zaman jenis pesantren terdiri dari pesantren salafiyah yang masih mempertahankan tradisi klasik yang focus pada pembelajaran menggunakan kitab kuning, ciri khas dari pesntren ini terletak pada metode pembelajarannya seperti sorogan, bandingan ataupun metode tradisional lainnya [8], sedangkan pesantren khalafiyah Pesantren jenis ini telah mengalami modernisasi dengan mengadopsi kurikulum pendidikan nasional, seperti pelajaran Matematika, IPA, dan Bahasa Inggris. Di samping itu, pelajaran agama tetap diajarkan secara intensif, sehingga terbentuk integrasi antara ilmu agama dan ilmu umum [9]. Pesantren tidak hanya berfungsi sebagai institusi pendidikan, tetapi juga sebagai pusat dakwah, pembinaan karakter, dan penguatan budaya keislaman [10]. Menurut Undang-Undang No. 18 Tahun 2019 tentang Pesantren, pesantren memiliki tiga fungsi utama, yaitu: pendidikan, dakwah, dan pemberdayaan masyarakat. Fungsi ini menjadikan pesantren sebagai lembaga yang berperan strategis dalam mencetak generasi yang berakhlak mulia, mandiri, dan berwawasan keagamaan yang kuat [11].

2.2. Machine Learning

Machine learning adalah salah satu bagian dari *Artificial Intelligence* yang memungkinkan computer belajar dari data untuk membuat prediksi ataupun Keputusan. Dalam konteks sistem rekomendasi pendidikan, *machine learning* digunakan untuk mengidentifikasi pola dari data siswa dan memetakan kecocokan dengan jenis lembaga pendidikan yang sesuai [12]. *Machine learning* dibagi menjadi beberapa kategori utama, yaitu: *Supervised Learning* yaitu pembelajaran dengan label, *Unsupervised Learning* yaitu pembelajaran tanpa label, dan *Reinforcement Learning* yang berbasis umpan balik dari lingkungan, Penelitian ini menggunakan pendekatan supervised learning, khususnya metode klasifikasi [13]. Metode KNN adalah salah satu algoritma klasifikasi *supervised learning* yang sering digunakan dalam penelitian terdahulu. Prinsip utama dari KNN adalah bahwa suatu data akan diklasifikasikan ke dalam kelas yang paling umum di antara *k* tetangganya yang terdekat. Kedekatan atau kemiripan antar data dihitung menggunakan fungsi jarak, umumnya *Euclidean Distance* [14]. Metode RF juga salah satu algoritma pada klasifikasi algoritma *ensemble learning* yang membangun banyak pohon keputusan (*decision trees*) dan menggabungkannya untuk menghasilkan prediksi yang lebih stabil dan akurat. Setiap pohon dalam RF dilatih menggunakan subset acak dari data dan fitur, teknik ini dikenal sebagai *bagging* (*Bootstrap Aggregating*) [15]. Dalam klasifikasi, hasil akhir ditentukan melalui voting mayoritas dari hasil klasifikasi setiap pohon. RF unggul karena mampu menangani data berdimensi tinggi, mengurangi risiko *overfitting*, dan memiliki ketahanan terhadap data yang bervariasi.

2.3. Metode *K-Nearest Neighbors* (KNN)

Metode KNN merupakan salah satu algoritma dalam *supervised learning* yang berbasis pada pendekatan *instance*-based atau *lazy learning*, dimana proses pembelajaran tidak dilakukan melalui pembentukan model eksplisit, melainkan berdasarkan perbandingan langsung antara data baru dan data latih yang telah tersedia. Prinsip dasar dari KNN adalah bahwa suatu objek atau data baru akan diklasifikasikan ke dalam kelas tertentu berdasarkan mayoritas kelas dari sejumlah

k tetangga terdekat yang memiliki jarak paling kecil terhadap data tersebut. Ukuran kedekatan atau jarak ini umumnya dihitung menggunakan metrik seperti *Euclidean Distance*, *Manhattan Distance*, atau *Minkowski Distance*, bergantung pada karakteristik data yang dianalisis. Secara matematis, apabila terdapat data baru x yang ingin diklasifikasikan, maka jarak terhadap setiap data latih x_i dapat dihitung menggunakan persamaan 1:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n}(x_j - x_{ij})^2} \quad .....(1)$$

Persamaan 1 digunakan untuk mengukur perbedaan antara setiap pasangan komponen x_j dan x_{ij} , kemudian mengkuadratkannya untuk menghindari nilai negatif, menjumlahkan seluruh hasil kuadrat tersebut, dan akhirnya mengambil akar kuadrat dari jumlah totalnya. Hasilnya adalah nilai positif yang merepresentasikan jarak "lurus" atau jarak terpendek antara dua titik dalam ruang n-dimensi. Metode ini banyak digunakan dalam berbagai algoritma pembelajaran mesin dan analisis data, seperti KNN, *clustering*, serta pengukuran kemiripan antar vektor fitur. Selanjutnya, algoritma KNN akan memilih k data dengan jarak terkecil, kemudian menentukan kelas atau label dari data baru berdasarkan kategori yang paling sering muncul di antara tetangga tersebut. Nilai k yang digunakan menjadi parameter penting dalam KNN, karena nilai k yang terlalu kecil berpotensi menghasilkan klasifikasi yang sensitif terhadap noise, sedangkan nilai k yang terlalu besar dapat mengaburkan batas antar kelas. Dalam konteks rekomendasi jenis pesantren, setiap siswa direpresentasikan sebagai vektor fitur yang mencakup berbagai dimensi minat, seperti minat akademik, minat keagamaan, preferensi lingkungan belajar, hingga kecenderungan sosial dan kegiatan ekstrakurikuler. Melalui pendekatan KNN, profil seorang siswa baru dapat dibandingkan dengan profil siswa lain yang telah diketahui jenis pesantren yang sesuai dengan mereka. Dengan demikian, sistem dapat memberikan rekomendasi jenis pesantren yang memiliki karakteristik paling mirip dengan kelompok siswa yang memiliki minat serupa.

Kelebihan utama KNN adalah kesederhanaan implementasinya, fleksibilitas terhadap berbagai jenis data, serta kemampuannya memberikan hasil yang intuitif tanpa memerlukan model pelatihan yang kompleks. Namun, algoritma ini juga memiliki kelemahan, seperti sensitivitas terhadap skala data dan tingginya kebutuhan komputasi untuk dataset yang besar. Oleh karena itu, pada penelitian ini KNN digunakan terutama untuk tahap *profiling* dan *similarity matching*, sebelum hasilnya diproses lebih lanjut dengan metode klasifikasi yang lebih kompleks seperti Random Forest untuk menghasilkan prediksi akhir yang lebih akurat dan stabil.

2.4. Metode Random Forest

Metode Random Forest (RF) merupakan algoritma *ensemble learning* yang dikembangkan oleh Breiman sebagai penggabungan dari sejumlah decision tree independen yang dibangun secara acak terhadap subset data dan subset fitur. Pendekatan ini bertujuan untuk mengurangi varians model dan mengatasi permasalahan *overfitting* yang sering terjadi pada model pohon tunggal (*single decision tree*). Prinsip utama RF adalah membangun banyak pohon keputusan (*forest*) dan menggabungkan hasil prediksi dari setiap pohon dengan cara *voting* (untuk klasifikasi) atau *averaging* (untuk regresi). Setiap pohon dalam RF dilatih menggunakan proses yang disebut *bootstrap aggregating* (*bagging*), di mana data latih diambil secara acak dengan pengembalian (*sampling with replacement*). Selain itu, pada setiap node pohon, hanya sebagian acak dari fitur yang dipertimbangkan untuk pemisahan (*split*), sehingga tercipta keberagaman antar pohon. Hal ini menghasilkan model yang lebih tahan terhadap variasi data dan tidak mudah terjebak pada pola tertentu dalam dataset. Secara formal, jika terdapat *N* pohon keputusan dalam suatu forest, maka hasil prediksi akhir untuk data input x dapat dinyatakan seperti persamaan 2:

$$\hat{y} = mode\{h_1(x), h_2(x). \ldots, h_N(x)\} \quad .....(2)$$

Persamaan 2 menunjukkan proses pengambilan keputusan akhir dalam metode ensemble learning, khususnya pada pendekatan voting classifier. Nilai prediksi akhir $\hat{y}$ diperoleh dari

modus (nilai yang paling sering muncul) di antara hasil prediksi dari sejumlah model atau classifier individu, yaitu $h_1(x), h_2(x)$, hingga $h_N(x)$. Dengan kata lain, setiap model memberikan prediksi terhadap input x, kemudian hasil dari semua model dikumpulkan, dan prediksi yang paling banyak dipilih dianggap sebagai hasil akhir. Pendekatan ini meningkatkan keandalan sistem klasifikasi karena menggabungkan pendapat dari beberapa model untuk mengurangi risiko kesalahan dari satu model tunggal.

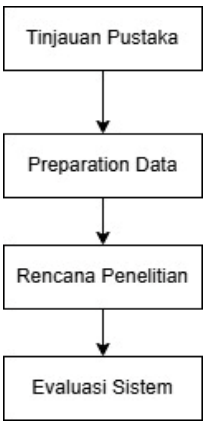
Dalam konteks sistem rekomendasi jenis pesantren, RF berperan dalam proses klasifikasi akhir dengan mempertimbangkan berbagai faktor multidimensi dari profil siswa. Algoritma ini mampu mengidentifikasi variabel yang paling berpengaruh dalam menentukan kecocokan jenis pesantren, seperti minat akademik, orientasi spiritual, lingkungan sosial, serta nilai-nilai preferensi individu. Kemampuan untuk menghitung feature importance juga memberikan nilai tambah, karena memungkinkan peneliti dan pengembang sistem untuk memahami aspek-aspek dominan yang memengaruhi hasil rekomendasi. Dibandingkan dengan algoritma tunggal lainnya, RF menawarkan beberapa keunggulan, antara lain kemampuan menangani data berskala besar, ketahanan terhadap *noise*, serta performa yang baik terhadap data yang tidak seimbang (*imbalanced data*). Selain itu, metode ini tidak memerlukan pra-pemrosesan yang kompleks dan relatif robust terhadap data dengan korelasi antar variabel. Oleh karena itu, dalam penelitian ini RF dipilih sebagai algoritma utama dalam tahap klasifikasi, karena memberikan keseimbangan antara akurasi, interpretabilitas, dan efisiensi komputasi.

3. METODOLOGI PENELITIAN

Pada tahapan ini dilakukan evaluasi komparasi 2 metode *machine learning* yaitu metode KNN dan RF dengan menggunakan dataset 327 siswa SD dan SMP di sekolah XXX serta Bahasa pemrograman phyton untuk mengetahui hasil dari perbandingan 2 metode tersebut

3.1. Tahapan Penelitian

Tahapan Penelitian terdiri dari 4 tahapan yaitu Tinjauan Pustaka, Preparation data, Rancangan penelitian, dan Evaluasi Sistem. Gambar 1 merupakan tahapan penelitian yang diusulkan.



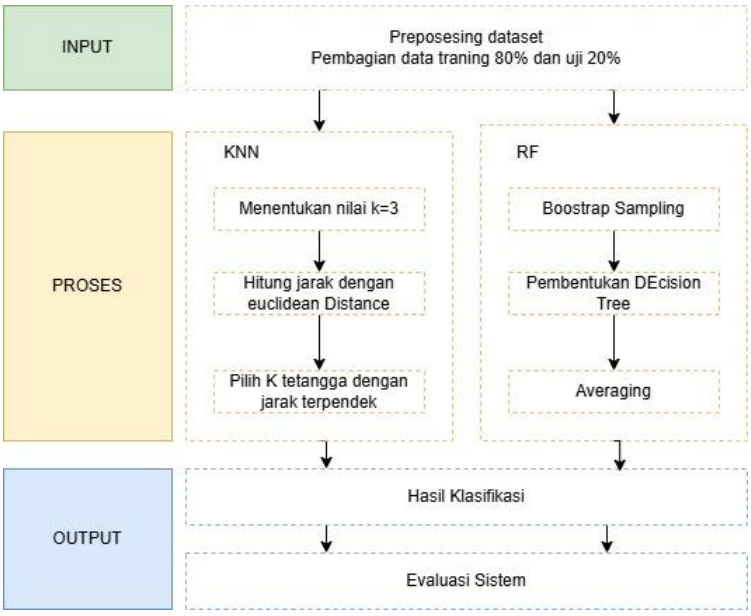
Gambar 1. Tahapan Penelitian

Gambar 1 menunjukkan alur tahapan penelitian yang terdiri dari empat langkah utama, yaitu Tinjauan Pustaka, *Preparation* Data, Rencana Penelitian, dan Evaluasi Sistem. Tahapan pertama, Tinjauan Pustaka, berfungsi untuk mengumpulkan dan menganalisis berbagai referensi ilmiah yang relevan guna membangun dasar teori penelitian. Selanjutnya, pada tahap *Preparation Data*, dilakukan proses pengumpulan, pembersihan, dan pengolahan awal data agar siap digunakan dalam analisis atau pengujian. Tahap berikutnya, Rencana Penelitian, mencakup perancangan metode, model, serta strategi eksperimen yang akan diterapkan untuk mencapai tujuan penelitian. Terakhir, tahap Evaluasi Sistem dilakukan untuk mengukur kinerja dan efektivitas sistem atau

model yang telah dikembangkan, dengan membandingkan hasil yang diperoleh terhadap kriteria atau metrik yang telah ditetapkan sebelumnya.

Diawali dengan melalukan tinjauan pustaka beberapa referensi berkaitan dengan metode klasifikasi KNN dan random forest, kemudian dilakukan peninjuan beberapa artikel tersebut guna memperoleh informasi mengenai metode-metode terkini yang digunakan dalam penelitian ini, selanjutkan di lakukan persiapan data yang akan digunakan dalam penelitian ini, *dataset* yang akan di olah dalam penelitian ini diperoleh dari lembaga sekolah XXX di Tingkat SD dan SMP yang berjumlah 327 siswa, serta dilakukan peninjauan *dataset* yang relevan, melakukan pembersihan data mentah yang tidak diperlukan sebelum digunakan dalam analisis atau pelatihan model *machine learning* yang biasa disebut dengan preprosesing data, tahap selanjutnya adalah merancang sebuah sistem komparasi dengan menggunakan metode KNN dan Random Forest yang akan di terapkan pada lembaga sekolah Xxx Tingkat SD dan SMP yang bertujuan untuk mengklasifikasi pondok pesantren berdasarkan minat dan profil dari siswa tersebut, adappun 3 kelompok klasifikasi nya yaitu antara lain : salafiya, khlafiyah, dan campuran, sedangkan data profil siswa di antaranya, usia, jenis kelamin, nilai rapor agama, nilai rapor umum, Bahasa yang di kuasai, minat hafalan,kedisplinan. Tujuan evalasuai kedua metode ini adalah agar dapat mengetahui hasil performa terbaik jika di terapkan pada *dataset* tersebut.[17]. Tahapan yang terakhir yaitu melakukan evaluasi sistem untuk menganalisa dan menguji performa kedua metode tersebut dengan menggunakan metrik evaluasi MAE dan RMSE.

3.2. Rancangan Penelitian



Gambar 2. Rancangan Sistem

Pada rancangan penelitian dilakukan untuk merancang dan mengevaluasi kedua metode yaitu KNN dan Random Forest, untuk mengetahui hasil terbaik jika di terapkan pada *dataset* siswa di sebuah lembaga sekolah SD dan SMP di Surabaya. Gambar 2 menunjukkan diagram alur proses klasifikasi menggunakan metode KNN dan RF. Proses dimulai dari tahap input, yaitu *preprocessing dataset* yang mencakup pembagian data menjadi 80% untuk pelatihan (*training*) dan 20% untuk pengujian (*testing*). Pada tahap proses, terdapat dua metode yang digunakan secara paralel. Untuk KNN, langkah-langkahnya meliputi penentuan nilai k=3, perhitungan jarak antar data menggunakan rumus *Euclidean Distance*, kemudian pemilihan tiga tetangga terdekat sebagai dasar klasifikasi. Sedangkan pada Random Forest, dilakukan *bootstrap sampling* untuk menghasilkan subset data, pembentukan sejumlah *decision tree*, dan penggabungan hasil klasifikasi melalui proses *averaging*. Tahap output menampilkan hasil klasifikasi dari kedua metode, yang kemudian dievaluasi pada tahap akhir untuk menilai kinerja sistem secara

keseluruhan. Diagram ini menggambarkan alur kerja sistem klasifikasi yang terstruktur dari input data hingga evaluasi performa model.

3.3. Evaluasi Sistem

Tujuan adanya evaluasi sistem adalah untuk dapat menganalisa performa kedua metode yaitu KNN dan RF jika di terapkan pada *dataset* untuk membanding hasil MAE dan RMSE jaika hasilnya lebih rendah maka metode tersebut yang baik jika di terapkan pada *dataset.* Rumus perhitungan MAE dan RMSE masing-masing disajikan pada Persamaan 3 dan 4.

$$MAE = \frac{1}{TN}\sum_{u\epsilon U, i\epsilon I}|p_{ui} - r_{ui}| \quad .....(3)$$

$$RMSE = \sqrt{\frac{1}{TN}}\sum_{u\epsilon U, i\epsilon I}(p_{ui} - r_{ui})^2 \quad .....(4)$$

Persamaan 3 dan persamaan 4 merupakan metrik evaluasi kinerja sistem prediksi atau rekomendasi, yaitu *MAE* dan *RMSE*. Persamaan MAE menghitung rata-rata selisih absolut antara nilai prediksi puip_{ui}pui dan nilai aktual ruir_{ui}rui, yang menggambarkan seberapa besar kesalahan prediksi tanpa memperhatikan arah (positif atau negatif) kesalahan tersebut. Sementara itu, persamaan RMSE mengukur akar dari rata-rata kuadrat selisih antara nilai prediksi dan nilai aktual, sehingga memberikan penalti yang lebih besar terhadap kesalahan yang memiliki selisih besar. Kedua metrik ini digunakan untuk menilai akurasi model prediksi, di mana nilai MAE dan RMSE yang lebih kecil menunjukkan bahwa hasil prediksi model semakin mendekati nilai sebenarnya dan memiliki performa yang lebih baik.

4. PENGUJIAN DAN PEMBAHASAN

Pada penelitian ini langkah pertama yang di lakukan adalah pengambilan data secara kolektif di lembaga sekolah XXX berjumlah 327 siswa SD dan SMP berdasarkan kuesioner yang diberikan kepada sejumlah siswa tersebut. Setelah itu akan dilakukan pemisahan data antara data training sebesar 80% dan data latih sebesar 20%. Tabel 1 menampilkan data karakteristik santri yang digunakan untuk proses analisis atau klasifikasi terhadap kecenderungan kelompok tertentu yang dilabeli sebagai Khalafi, Salafi, atau Campuran. Setiap baris mewakili satu individu dengan berbagai atribut seperti usia, jenis kelamin, asal daerah (Jawa atau luar Jawa), serta tingkat pendidikan (SD, SMP, MTs). Selain itu, terdapat variabel kemampuan bahasa yaitu Bahasa Arab dikuasai dan Bahasa Inggris dikuasai, masing-masing dinilai dalam kategori tidak, sedang, atau baik. Atribut lain meliputi minat terhadap kitab kuning dan minat hafalan, yang dinilai menggunakan skala numerik, serta cita-cita yang diklasifikasikan sebagai umum atau dakwah. Kolom terakhir, label, menunjukkan hasil klasifikasi akhir berdasarkan kombinasi atribut tersebut. Data seperti ini umumnya digunakan dalam penelitian berbasis pembelajaran mesin untuk mengidentifikasi pola, membangun model prediksi, atau memahami karakteristik kelompok santri berdasarkan faktor-faktor pendidikan dan minat mereka.

Tabel 1. *Dataset*

usia	JK	Asal daerah	Tingkat	Bahasa Arab dikuasai	Bahasa Nggris dikuasai	Minat kitab kuning	Minat hafalan	Cita Cita	label
18	L	Luar Jawa	SMP	sedang	sedang	4	1	umum	Khalafi
15	L	Jawa	SD	sedang	sedang	4	5	dakwah	Salafi
16	L	Jawa	SD	tidak	tidak	1	2	umum	Salafi
18	P	Jawa	SMP	tidak	sedang	1	4	dakwah	Salafi
14	L	Jawa	SMP	sedang	sedang	3	2	umum	Khalafi
19	L	Jawa	SMP	tidak	sedang	4	3	umum	Campuran
16	P	Luar Jawa	MTs	tidak	tidak	5	2	dakwah	Salafi
16	L	Jawa	SD	sedang	baik	2	3	umum	Khalafi
18	P	Luar Jawa	MTs	tidak	sedang	3	2	umum	Khalafi

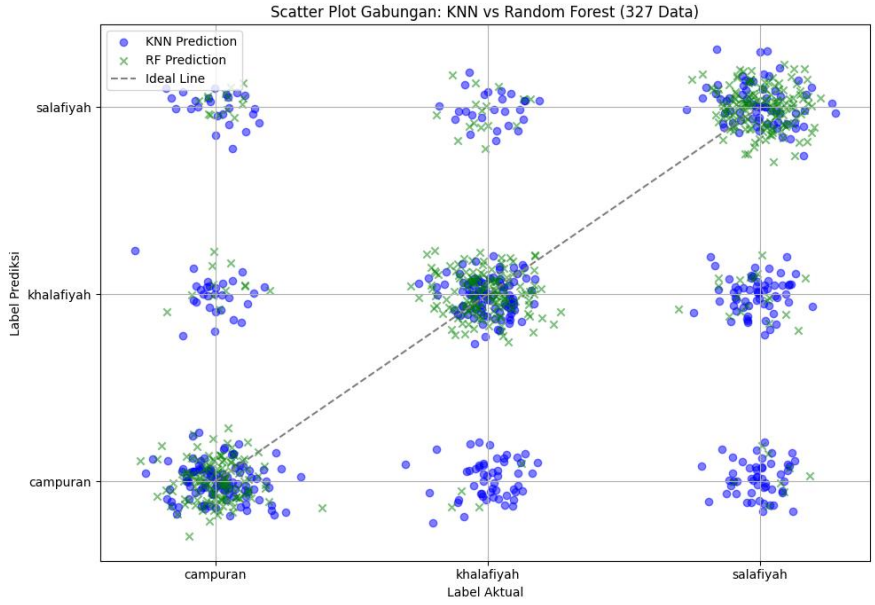
Dari dataset tersebut dilakukan proses data klasifikasi dengan menggunakan metode KNN dan RF, untuk metode KNN akan menggunakan nilai K=3 dan rumus *euclidean distance* sebagai penentu jerak terdekat antar siswa sedangkan metode RF, dari data pelatihan, diambil beberapa subset secara acak dengan teknik *bootstrap* (pengambilan dengan pengembalian). Setiap subset digunakan untuk membangun satu decision tree. Pada proses Decision tree etiap node dalam pohon, hanya sebagian fitur yang dipilih secara acak untuk digunakan dalam proses split (pembelahan). Ini berbeda dari pohon tunggal yang mempertimbangkan semua fitur, sehingga menghasilkan pohon yang lebih bervariasi dan mengurangi overfitting. Tabel 2 menunjukkan hasil nilai dari klasifikasi menggunakan KNN dan RF.

Tabel 2. Perbandingan KNN vs Random Forest

No	Data Uji	Label Sebenarnya	Prediksi KNN	Prediksi RF
1	Test_1	Campuran	Campuran	Campuran
2	Test_2	Salafi	Salafi	Salafi
3	Test_3	Salafi	Campuran	Salafi
4	Test_4	Campuran	Khalafi	Khalafi
5	Test_5	Salafi	Campuran	Salafi

Tabel 2 menampilkan tabel perbandingan hasil prediksi antara algoritma KNN dan RF terhadap data uji yang terdiri dari lima sampel, yaitu *Test_1* hingga *Test_5*. Setiap baris menunjukkan label sebenarnya (*label sebenarnya*) serta hasil prediksi dari kedua model. Dari tabel tersebut terlihat bahwa pada beberapa kasus, seperti *Test_1* dan *Test_2*, kedua model memberikan hasil prediksi yang sama dan sesuai dengan label sebenarnya. Namun, pada data seperti *Test_3* dan *Test_5*, model KNN menghasilkan prediksi yang berbeda dari label sebenarnya, sedangkan Random Forest memberikan hasil yang benar. Hal ini menunjukkan bahwa Random Forest cenderung memberikan akurasi prediksi yang lebih stabil dibandingkan KNN, karena kemampuannya dalam melakukan *ensemble learning* melalui penggabungan beberapa *decision tree*, sehingga lebih tahan terhadap variasi data dibandingkan metode berbasis jarak seperti KNN.

Dari hasil perhitungan kedua metode tersebut terdapat nilai MAE dan RMSE nya sebagai metrik evaluasi. Untuk metode KNN MAE = 0.6 sedangkan RMSE= 0.96 dan metode RF nilai MAE = 0.23 dan RMSE = 0.57 , untuk gambar sebaran data dalam perhitungan KNN dan RF dapat dilihat pada Gambar 3, dimana data sebaran dari 3 kelompok kalsifikasi yaitu salafiya, khlafiya dan campuran.



Gambar 3. *Scatter Plot* Model KNN dan RF

Gambar 3 menunjukkan scatter plot gabungan hasil prediksi antara algoritma KNN dan RF terhadap 327 data uji dengan tiga kategori label, yaitu *Campuran*, *Khalafiyah*, dan *Salafiyah*. Sumbu horizontal menggambarkan label aktual, sedangkan sumbu vertikal menunjukkan label hasil prediksi. Titik berwarna biru merepresentasikan prediksi dari KNN, sementara tanda silang hijau menunjukkan hasil prediksi dari Random Forest. Garis putus-putus diagonal (*ideal line*) menandakan posisi di mana prediksi model sepenuhnya sesuai dengan label sebenarnya. Berdasarkan distribusi titik, terlihat bahwa sebagian besar prediksi Random Forest (tanda hijau) lebih terkonsentrasi di sekitar garis ideal dibandingkan dengan prediksi KNN (titik biru), yang menunjukkan bahwa Random Forest memiliki akurasi dan konsistensi prediksi yang lebih baik. Namun, terdapat juga sejumlah kesalahan klasifikasi pada kedua model, terutama di area peralihan antar kelas, yang menandakan adanya tumpang tindih karakteristik antar kategori data.

5. KESIMPULAN

Penelitian ini menggunakan dataset 327 siswa tingkat SD dan SMP di Xxx Surabaya dengan penerapan algoritma KNN dan RF untuk sistem rekomendasi jenis pesantren. Hasilnya menunjukkan bahwa kedua metode memiliki performa yang baik dengan nilai MAE dan RMSE di bawah 1.0, namun RF menunjukkan nilai kesalahan yang lebih kecil. Hal ini membuktikan bahwa RF lebih unggul dalam menangani data kompleks dan mengurangi risiko overfitting dibandingkan KNN, yang sensitif terhadap pemilihan parameter dan skala data. Kinerja baik dari kedua metode ini tetap memiliki keterbatasan, terutama dalam kualitas dan kelengkapan data profil siswa. KNN sangat bergantung pada data yang terstandardisasi dan bebas dari noise, sedangkan RF walaupun lebih stabil, memerlukan sumber daya komputasi lebih besar karena kompleksitas model dan banyaknya decision tree. Selain itu, interpretabilitas hasil RF menjadi tantangan tersendiri karena proses pengambilan keputusannya sulit dijelaskan secara sederhana kepada pengguna non-teknis, seperti guru atau konselor pendidikan. Generalisasi model juga terbatas karena dataset yang digunakan mungkin hanya mewakili karakteristik lokal, sehingga perlu diuji ulang pada populasi siswa di wilayah lain.

Keterbatasan lain berasal dari aspek non-teknis dan psikologis, seperti motivasi, nilai religius, serta faktor afektif siswa yang sulit dikonversi menjadi data numerik. Karena itu, sistem rekomendasi ini sebaiknya digunakan sebagai alat bantu pengambilan keputusan, bukan sebagai penentu mutlak dalam penempatan siswa ke jenis pesantren tertentu. Untuk pengembangan berikutnya, penelitian disarankan memperkaya dataset dengan variabel tambahan seperti hasil asesmen psikologis, preferensi ekstrakurikuler, dan data perilaku digital. Dengan demikian, model dapat memahami minat siswa secara lebih komprehensif dan menghasilkan prediksi yang lebih akurat. Selain itu, pendekatan *hybrid* dan *explainable* AI (XAI) perlu dipertimbangkan agar sistem tidak hanya kuat secara prediktif, tetapi juga mampu memberikan alasan yang jelas di balik setiap rekomendasi. Integrasi metode seperti deep learning dan reinforcement learning dapat membantu sistem belajar secara adaptif dari umpan balik pengguna. Terakhir, pengembangan sistem di masa depan perlu memperhatikan etika, privasi data, dan keadilan algoritmik agar sistem rekomendasi ini tidak bias terhadap kelompok tertentu. Dengan demikian, sistem berbasis KNN dan RF ini diharapkan menjadi solusi cerdas, etis, dan berorientasi pada pengembangan potensi siswa secara menyeluruh.

DAFTAR PUSTAKA

[1] A. Karim et al., “Islamic spiritual leadership of kyai in fostering santris' entrepreneurial spirit and independence in boarding school,” Soc. Sci. Humanit. Open, vol. 12, p. 101817, 2025, doi: https://doi.org/10.1016/j.ssaho.2025.101817

[2] M. H. Marwiji, I. M. Nusantara, and S. Barat, “Transformasi Kurikulum dan Sistem Pembelajaran di Pondok Pesantren Salaf pada Era Disrupsi,”, Ta'alum Jurnal Pendidikan Islam, vol. 12, pp. 1–23, 2024. https://doi.org/10.21274/taalum.2024.12.1.1-23

[3] G. Asyari, K. S. Damanik, D. Avira, U. Muhammadiyah, and S. Utara, “Mengenal Jenis Pesantren Salafiyyah , Khalafiyyah dan Komprehensif di Pondok Pesantren Salafiyah

Luqmanul Hakim Marelan,", Innovative: Journal of Social Science Research, vol. 4, pp. 16099–16111, 2024. https://doi.org/10.31004/innovative.v4i3.12136

[4] A. J. Asifudin, "Manajemen Pendidikan untuk Pondok Pesantren," Manageria: Jurnal Manajemen Pendidikan Islam, vol. 1, no. November, pp. 355–366, 2016. https://doi.org/10.14421/manageria.2016.12-10

[5] I. F. Wijaya, A. Moro, Y. Belghitar, and M. A. Prabowo, "Islam and early-stage entrepreneurial activity in Indonesia: religion is not the opium of entrepreneurship," J. Small Bus. Enterp. Dev., vol. 32, no. 8, pp. 137-168, 2025, doi: https://doi.org/10.1108/JSBED-03-2024-0159.

[6] Y. Hanafi et al., "The new identity of Indonesian Islamic boarding schools in the 'new normal': the education leadership response to COVID-19," Heliyon, vol. 7, no. 3, p. e06549, 2021, doi: https://doi.org/10.1016/j.heliyon.2021.e06549

[7] M. Di, P. Pesantren, and A. L. Anshar, " Manajemen Pondok Pesantren; Studi Pengelolaan Santri Muallaf Di Pondok Pesantren Al Anshar Ambon ", Jurnal Fikratuna , Volume 9, Nomor 1 2018 ," vol. 9, pp. 54–70, 2018. https://doi.org/10.33477/fkt.v9i1.675

[8] I. Mujahidin, S. Tinggi, A. Islam, and P. Thawalib, " Problem Hasrat Belajar Mahasiswa Dan Tawaran Kontekstualisasi Konsep Ruang Publik Jurgen Habermas Di Iaida Sebagai Solusinya", Jurnal Komunikasi dan Penyiaran Islam," vol. 1, no. 1, pp. 31–44, 2021. https://doi.org/10.54150/syiar.v1i1.33

[9] A. Tadbir, I. Education, and M. Journal, "Manajemen Promosi Pesantren Berbasis Santri," At Tadbir Islamic Education Management Journal, vol. 3, no. 1, pp. 75–89, 2025. https://doi.org/10.54437/attadbir.v3i1.2114

[10] J. S. Komputer, I. Anas, S. Zakir, U. Islam, N. Sjech, and M. D. Djambek, "Artificial Intelligence : Solusi Pembelajaran Era," vol. 8, pp. 35–46, 2024. https://doi.org/10.30645/j-sakti.v8i1.764

[11] N. Tamascelli, A. Campari, T. Parhizkar, and N. Paltrinieri, "Artificial Intelligence for safety and reliability: A descriptive, bibliometric and interpretative review on machine learning," J. Loss Prev. Process Ind., vol. 90, p. 105343, 2024, doi: https://doi.org/10.1016/j.jlp.2024.105343

[12] M. M. Rofi, F. A. Setiawan, F. Riana, T. Informatika, F. Teknik, and U. I. Khaldun, "Perbandingan Metode K-NN Dan Random Forest Pada Klasifikasi," Sainteks vol. 10, no. 1, pp. 84–89, 2024. https://doi.org/10.31949/infotech.v10i1.8856

[13] C. Ma, Z. Dai, and W. Zhang, "Recognition of car horns based on principal component analysis of MEL frequency Cepstral coefficients and support vector machine," Comput. Electr. Eng., vol. 120, p. 109666, 2024, doi: https://doi.org/10.1016/j.compeleceng.2024.109666

[14] A. Ramadhan and B. Susetyo, "Penerapan Metode Klasifikasi Random Forest Dalam Mengidentifikasi Faktor Penting Penilaian Mutu Pendidikan Application Of Random Forest Classification Method In Identi- Fying Important Factors Of Educational Quality," Jurnal Pendidikan dan Kebudayaan, vol. 4, pp. 169–182, 2019. https://doi.org/10.24832/jpnk.v4i2.1327

[15] P. K. Neighbor, "Perbandingan K-Nearest Neighbor dan Random Forest dengan Seleksi Fitur Information Gain untuk Klasifikasi Lama Studi Mahasiswa Indonesian". Journal of Applied Statistics, vol. 5, no. 1, pp. 58–66, 2022. https://doi.org/10.13057/ijas.v5i1.58056