ANALISIS KLASIFIKASI KELULUSAN SISWA MENGGUNAKAN METODE KNN (K-NEAREST NEIGHBOR) RANDOM FOREST PADA DATASET "STUDENTS PERFORMANCE"

Muhammad Rizki Arrohman [1)], Nikmatul Khoiriyah [2)], Arina Fawaida [3)], Diana Laily Fithri [4)]
Program Studi Sistem Informasi, Universitas Muria Kudus [1),2),3),4)]

rahmanrizki267@gmail.com[1)], nikmatulkhoiriyah0818@gmail.com[2)],
arinafawa04@gmail.com[3)], Diana.laily@umk.ac.id [4)]

Abstrak

Keberhasilan siswa menjadi indikator penting dalam menilai kualitas sistem pendidikan. Oleh karena itu, prediksi kelulusan siswa berdasarkan data demografis dan akademik sangat penting untuk membantu institusi pendidikan melakukan tindakan pencegahan. Tujuan dari penelitian ini adalah untuk menganalisis dan membandingkan kinerja dua algoritma klasifikasi, K-Nearest Neighbor (KNN) dan Random Forest, ketika digunakan untuk memprediksi kelulusan siswa. Dataset kinerja siswa mencakup informasi seperti jenis kelamin, kelompok etnis, pendidikan orang tua, status makan siang, program persiapan ujian, matematika, membaca, dan nilai ujian menulis. Nilai rata-rata dari ketiga mata pelajaran tersebut digunakan untuk menentukan kelulusan siswa. Hasil penelitian menunjukkan bahwa algoritma Random Forest memiliki akurasi 100% dan KNN memiliki akurasi 90%. Hasil penelitian ini diharapkan dapat digunakan sebagai referensi dalam pengembangan sistem pendukung keputusan berbasis data di bidang pendidikan.

Kata kunci: *Klasifikasi, KNN, Random Forest, Kelulusan Siswa, Students Performance, Machine Learning.*

1. Pendahuluan

Kemajuan dalam teknologi informasi dan komputasi telah sangat memengaruhi berbagai aspek kehidupan, termasuk pendidikan. Pemanfaatan data dalam era komputer dan internet saat ini menjadi semakin penting untuk membantu pengambilan keputusan strategis. Institusi pendidikan menghadapi banyak tantangan dalam upaya mereka untuk terus meningkatkan kualitas layanan mereka, termasuk menjaga dan menilai kinerja akademik siswa. Tingkat kelulusan siswa adalah indikator penting keberhasilan pendidikan. Tingkat kelulusan tidak hanya menunjukkan kualitas pembelajaran, tetapi juga berfungsi sebagai referensi untuk membangun strategi pendidikan yang lebih tepat sasaran.

Untuk guru, sekolah, dan lembaga pengambil kebijakan pendidikan, prediksi kelulusan siswa dapat sangat bermanfaat. Sekolah dapat mengambil tindakan seperti memberikan bimbingan belajar tambahan, pendampingan psikologis, atau pendekatan personal lainnya jika mereka tahu bahwa siswa mungkin tidak lulus sejak dini. Oleh karena itu, sistem yang mampu

memprediksi kelulusan siswa secara akurat dan efektif diperlukan. Data mining, terutama dengan pendekatan klasifikasi, adalah salah satu metode yang dapat digunakan.

Beberapa teknik dalam data mining adalah estimasi, forecasting, klasifikasi, asosiasi dan klastering[1]. Data mining dengan metode klasifikasi memungkinkan pengelompokan data berdasarkan fitur tertentu. Data demografis dan akademik siswa dapat dianalisis untuk menentukan kemungkinan lulus mereka dalam konteks prediksi kelulusan. K-Nearest Neighbor (KNN) dan Random Forest adalah dua algoritma yang paling umum digunakan dalam tugas klasifikasi. KNN bekerja dengan prinsip kedekatan antara data uji dan data latih, sedangkan Random Forest adalah metode kelompok yang menggabungkan beberapa pohon keputusan untuk menghasilkan hasil yang lebih akurat dan stabil.

Fokus penelitian ini adalah dataset "Performasi siswa" yang diperoleh dari platform Kaggle. Dataset mengandung berbagai informasi, termasuk nilai membaca, menulis, dan matematika, serta karakteristik demografis seperti gender, etnis, status makan siang, pendidikan orang tua, dan program persiapan ujian. Aspek-aspek ini memberikan wawasan yang cukup luas untuk dianalisis dengan metode pembelajaran mesin.

2. Kajian Pustaka dan pengembangan hipotesis

2.1 Klasifikasi dalam Data Mining

Salah satu teknik utama dalam penggalian data adalah klasifikasi, yang digunakan untuk mengelompokkan data ke dalam kelas yang telah ditentukan sebelumnya. Kumpulan teknik yang dikenal sebagai data mining digunakan untuk mengeksplorasi nilai tambahan dari kumpulan data dalam bentuk informasi yang belum ditemukan sebelumnya [2]. Teknik ini termasuk dalam kategori pelatihan yang diawasi, di mana data berlabel digunakan untuk melakukan pelatihan. Tujuannya adalah membangun model yang memiliki kemampuan untuk menghubungkan data baru ke label kelas yang tepat. Dalam hal prediksi kelulusan siswa, klasifikasi sangat penting karena dapat digunakan untuk menentukan apakah seorang siswa berada dalam kategori "lulus" atau "tidak lulus" berdasarkan karakteristik tertentu mereka.

2.2 K-Nearest Neighbor (KNN)

Berdasarkan pembelajaran data yang sudah terklasifikasikan sebelumnya, algoritma K-Nearest Neighbor (K-NN) adalah metode klasifikasi untuk sekumpulan data[3]. KNN adalah teknik klasifikasi yang didasarkan pada pendekatan tetangga terdekat. KNN menghitung jarak antara data baru dan seluruh data dalam dataset pelatihan, kemudian memilih K data dengan jarak terdekat untuk menentukan kelas data baru. Tidak terlalu kompleks, algoritma ini efektif, terutama dalam situasi di mana distribusi data tidak linier. KNN juga tidak parametrik, yang berarti tidak membuat asumsi tentang distribusi data. Namun, kelemahan KNN adalah sensitivitasnya terhadap data noise dan kapasitas komputasi yang tinggi untuk dataset yang besar.

2.3 Random Forest

Sekumpulan pohon keputusan membentuk metode klasifikasi acak yang berbasis pembelajaran kelompok. Dalam proses pelatihan, algoritma berbasis kelompok menggabungkan berbagai metode pembelajaran mesin menjadi satu model prediktif[4]. Algoritma ini bekerja dengan membuat banyak pohon keputusan dan membuat prediksi kelas berdasarkan mayoritas suara dari seluruh pohon. Prediksi kelulusan siswa, analisis data medis, dan pengelompokan data pemasaran adalah beberapa keuntungan dari pengolahan Random Forest [5]. Salah satu keuntungan dari Hutan Random adalah kemampuan untuk menangani overfitting, yang sering terjadi pada pohon keputusan tunggal. Selain itu, Random Forest memiliki kemampuan untuk menangani dataset yang memiliki banyak fitur sekaligus

memberikan perkiraan pentingnya setiap fitur. Ini membuatnya sangat cocok untuk digunakan dalam prediksi berbasis data pendidikan yang biasanya kompleks dan multidimensional.

2.4 Penelitian Terkait

Beberapa penelitian sebelumnya menunjukkan efektivitas algoritma Random Forest dan KNN dalam berbagai aplikasi prediksi. Dalam studi [6] perbandingan antara SVM, KNN, dan Random Forest untuk prediksi gagal jantung menunjukkan bahwa Random Forest memiliki akurasi dan stabilitas yang lebih tinggi dibandingkan algoritma lainnya.

2.5 Hipotesis Penelitian

Random Forest memiliki sifat yang lebih tahan terhadap noise dan overfitting, serta lebih stabil dan akurat dalam menangani data multivariat. Berdasarkan penelitian dan penelitian sebelumnya, dapat diusulkan bahwa algoritma Random Forest akan memiliki akurasi klasifikasi kelulusan siswa yang lebih tinggi dibandingkan dengan algoritma KNN. Meskipun sederhana dan mudah digunakan, KNN memiliki keterbatasan ketika bekerja dengan dataset yang besar atau memiliki sifat irrelevan.

3. Metode Penelitian

3.1 Sumber Data

Penelitian ini menggunakan dataset publik berjudul Students Performance yang diunduh dari situs Kaggle. Dataset tersebut berisi data akademik dan demografis sebanyak 1000 entri siswa dengan 8 atribut utama, yakni jenis kelamin, kelompok etnis, pendidikan orang tua, status makan siang, kursus persiapan ujian, serta nilai ujian matematika, membaca, dan menulis. Dataset ini dipilih karena menyediakan kombinasi data yang memadai untuk melakukan analisis klasifikasi kelulusan siswa.

3.2 Preprocessing Data

Sebelum data digunakan untuk pelatihan model, beberapa tahap preprocessing dilakukan:

- Teknik label encoding digunakan untuk mengubah data kategorik seperti status makan siang, jenis kelamin, dan kelompok etnis menjadi data numerik yang dapat diproses oleh algoritma pembelajaran mesin.
- Rata-rata nilai dari tiga mata pelajaran (matematika, membaca, dan menulis) dihitung untuk menentukan status kelulusan siswa.
- Label klasifikasi dibuat dengan kriteria: siswa dengan rata-rata nilai ≥ 70 dikategorikan sebagai “Lulus” dan yang < 70 sebagai “Tidak Lulus”.
- Untuk menguji kemampuan generalisasi model, dataset dibagi secara acak menjadi data latih dan data uji dengan proporsi 80:20.

3.3 Algoritma Klasifikasi

Dalam penelitian ini, digunakan dua metode klasifikasi yang populer dan efektif:

- K-Nearest Neighbor (KNN): Algoritma yang mengklasifikasikan data berdasarkan kedekatan jarak (menggunakan metrik Euclidean) dengan tetangga terdekat pada data latih. Parameter k dipilih untuk menentukan jumlah tetangga yang digunakan.
- Random Forest: Algoritma kelompok yang membangun banyak pohon keputusan secara acak dan menggabungkan hasil voting dari seluruh pohon untuk menghasilkan keputusan klasifikasi yang lebih stabil dan akurat..

3.4 Tools dan Platform

Proses pengolahan data dan pengujian model dilakukan menggunakan Google Colab dengan bahasa pemrograman Python. Beberapa pustaka utama yang digunakan adalah:

- Pandas dan numpy untuk proses pengolahan data,
- Scikit-learn digunakan untuk mengimplementasikan model klasifikasi dan melakukan evaluasi performa,

- Menggunakan matplotlib dan seaborn untuk menampilkan data dan hasil pengujian.

3.5 Evaluasi Model

Selanjutnya, data uji digunakan untuk mengevaluasi model yang dilatih dengan metrik berikut:

- Akurasi: Proporsi prediksi yang tepat berdasarkan data uji keseluruhan.
- Presisi: Proporsi prediksi positif yang benar-benar positif.
- Recall: Proporsi data positif yang berhasil diprediksi dengan benar.
- F1-Score: Nilai harmonisasi antara presisi dan recall yang memberikan gambaran keseimbangan antara kedua metrik tersebut.
- Confusion Matrix: Matriks yang menampilkan jumlah prediksi benar dan salah dalam masing-masing kelas, memberikan gambaran detail performa model.

4. Hasil dan Pembahasan

4.1 Hasil Pengujian K-Nearest Neighbor (KNN)

Pengujian algoritma KNN dilakukan dengan menggunakan data uji sebanyak 20% dari keseluruhan dataset. Parameter k pada KNN dipilih melalui metode cross-validation agar mendapatkan nilai k terbaik yang meminimalkan error. Setelah proses pelatihan, hasil klasifikasi KNN menghasilkan performa sebagai berikut:

Tabel 1. 1 Hasil Pengujian K-Nearest Neighbor (KNN)

Metrik	Nilai
Akurasi	99%
Presisi	0.99 - 1.00
Recall	0.99 - 1.00
F1-Score	0.99

Confusion matrix adalah tabel untuk menggambarkan kinerja sampel klasifikasi yang digunakan untuk mengukur kinerja pada kumpulan data[7] .Confusion matrix menunjukkan hanya terjadi satu kesalahan klasifikasi, yaitu seorang siswa yang seharusnya masuk kategori “Lulus” diklasifikasikan sebagai “Tidak Lulus”. Hal ini mengindikasikan bahwa model KNN sangat sensitif terhadap data yang mirip dan bisa dipengaruhi oleh noise dalam dataset. Namun, secara keseluruhan, KNN menunjukkan performa yang sangat baik dalam memprediksi status kelulusan siswa.

4.2 Hasil Pengujian Random Forest

Algoritma Random Forest diuji pada dataset yang sama dengan menggunakan parameter default dan beberapa tuning sederhana untuk mengoptimalkan jumlah pohon keputusan (estimators). Hasil evaluasi menunjukkan performa yang sangat baik dengan hasil sebagai berikut:

Tabel 1. 2 Hasil Pengujian Random Forest

Metrik	Nilai
Akurasi	100%
Presisi	1.00
Recall	1.00
F1-Score	1.00

Confusion matrix memperlihatkan bahwa seluruh data uji berhasil diklasifikasikan dengan benar tanpa adanya kesalahan. Hal ini menandakan bahwa Random Forest mampu menangani variasi data dengan baik serta meminimalkan risiko overfitting melalui proses ensemble learning. Model ini sangat stabil dan akurat dalam memprediksi kelulusan siswa.

4.3 Perbandingan Kinerja KNN dan Random Forest

Dari hasil pengujian di atas, dapat dilihat bahwa kedua algoritma memiliki performa yang baik, namun terdapat perbedaan penting dalam hal akurasi dan stabilitas model. Random Forest unggul dengan akurasi sempurna (100%) tanpa kesalahan klasifikasi, sedangkan KNN memiliki akurasi mendekati sempurna sebesar 99% dengan satu kesalahan.

Tabel 1. 3 Perbandingan Kinerja KNN dan Random Forest

Aspek	KNN	Random Forest
Akurasi	99%	100%
Kesalahan	1	0
Stabilitas	Sensitif terhadap noise	Lebih stabil dan tahan terhadap overfitting
Kompleksitas	Lebih sederhana	Lebih kompleks namun lebih kuat

Perbedaan ini muncul karena Random Forest menggunakan mekanisme ensemble yang menggabungkan banyak pohon keputusan untuk mengambil keputusan akhir, sehingga lebih tahan terhadap variabilitas data dan noise. Sementara KNN bergantung pada tetangga terdekat sehingga bisa terpengaruh oleh data yang tidak representatif atau outlier.

4.4 Pembahasan

Hasil penelitian ini menegaskan bahwa pemilihan algoritma klasifikasi yang tepat sangat berpengaruh terhadap performa prediksi kelulusan siswa. Random Forest terbukti lebih unggul dalam hal akurasi dan stabilitas, membuatnya menjadi pilihan yang lebih baik untuk aplikasi sistem pendukung keputusan dalam pendidikan.

Namun, KNN tetap merupakan algoritma yang efisien dan mudah diimplementasikan, terutama jika dataset tidak terlalu besar dan data relatif bersih dari noise. Penggunaan KNN juga memberikan insight penting tentang hubungan kemiripan antar data siswa.

Selain itu, hasil ini menunjukkan bahwa fitur-fitur demografis dan nilai akademik yang digunakan dalam dataset cukup kuat untuk memprediksi status kelulusan. Hal ini membuka peluang untuk mengembangkan sistem prediktif yang dapat membantu lembaga pendidikan dalam mengidentifikasi siswa yang berisiko tidak lulus dan memberikan intervensi lebih awal.

Penelitian dapat diperluas dengan menambahkan fitur lain seperti kehadiran, aktivitas ekstrakurikuler, atau faktor psikologis yang mungkin mempengaruhi hasil belajar siswa. Selain itu, pengujian dengan algoritma lain seperti Support Vector Machine (SVM), Naive Bayes, atau model deep learning juga dapat dilakukan untuk mendapatkan perbandingan performa yang lebih komprehensif.

5. **Kesimpulan dan Saran**

5.1 Kesimpulan

Berdasarkan hasil analisis dan pengujian yang dilakukan, dapat disimpulkan beberapa hal sebagai berikut:

1. Algoritma K-Nearest Neighbor (KNN) dan Random Forest sama-sama efektif dalam melakukan klasifikasi kelulusan siswa berdasarkan dataset Students Performance yang memuat data demografis dan nilai akademik.
2. Random Forest menunjukkan performa yang lebih unggul dibandingkan KNN dengan akurasi 100%, serta metrik presisi, recall, dan F1-score yang sempurna. Hal ini menunjukkan bahwa Random Forest mampu memberikan prediksi yang sangat akurat dan stabil tanpa kesalahan klasifikasi pada data uji.
3. KNN juga memberikan hasil yang sangat baik dengan akurasi 99%, meskipun terdapat satu kesalahan klasifikasi, yang menunjukkan bahwa algoritma ini cukup sensitif terhadap keberadaan data noise atau outlier.
4. Fitur-fitur yang digunakan dalam dataset seperti nilai ujian dan atribut demografis terbukti mampu menjadi indikator kuat dalam memprediksi kelulusan siswa, sehingga dapat digunakan sebagai dasar dalam pembuatan sistem pendukung keputusan pendidikan.

5.2 Saran

Berdasarkan hasil penelitian dan kesimpulan di atas, beberapa saran yang dapat diberikan untuk pengembangan dan penelitian selanjutnya adalah:

1. Disarankan untuk menggunakan dataset yang lebih besar dan beragam, baik dari segi jumlah data maupun variasi atribut, untuk meningkatkan generalisasi dan kekuatan model klasifikasi.
2. Perlu dilakukan tuning parameter yang lebih mendalam pada algoritma KNN dan Random Forest, seperti pemilihan nilai k optimal untuk KNN atau jumlah pohon dan kedalaman maksimum untuk Random Forest, guna mengoptimalkan performa model.
3. Untuk mendapatkan evaluasi performa yang lebih mendalam, dapat dilakukan pengujian dengan algoritma klasifikasi lain seperti Support Vector Machine (SVM), Naive Bayes, Gradient Boosting, atau metode pembelajaran mendalam.
4. Penggunaan fitur tambahan seperti data kehadiran, motivasi belajar, atau faktor psikososial siswa dapat menjadi bahan penelitian lebih lanjut untuk meningkatkan akurasi prediksi kelulusan.
5. Implementasi hasil penelitian ini dalam bentuk sistem pendukung keputusan yang dapat digunakan oleh lembaga pendidikan sangat dianjurkan untuk membantu dalam identifikasi dini siswa yang berisiko tidak lulus dan memberikan intervensi tepat waktu.

Referensi

[1] C. Pangestu, S. Shaufiah, dan R. Wijaya, "X Spotify Cares Clustering Analysis using K-Means and K-Medoids," *J. Media Inform. Budidarma*, vol. 8, no. 1, hal. 497, 2024, doi: 10.30865/mib.v8i1.7279.

[2] Y. Azhar, A. K. Firdausy, dan P. J. Amelia, "Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Stroke," *SINTECH (Science Inf. Technol. J.*, vol. 5, no. 2, hal. 191–197, 2022, doi: 10.31598/sintechjournal.v5i2.1222.

[3] D. Cahyanti, A. Rahmayani, dan S. A. Husniar, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indones. J. Data Sci.*, vol. 1, no. 2, hal. 39–43, 2020, doi: 10.33096/ijodas.v1i2.13.

[4] L. Sari, A. Romadloni, dan R. Listyaningrum, "Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest," *Infotekmesin*, vol. 14, no. 1, hal. 155–162, 2023, doi: 10.35970/infotekmesin.v14i1.1751.

[5] F. Ilmu dan K. Universitas, "Prediksi Kelulusan Mahasiswa Menggunakan Metode K-Means dan Random Forest Fakultas Ilmu Komputer Universitas Sriwijaya , Indonesia Sales

Prediction of Student Graduation using K-Means and Random Forest Methods," vol. 5, no. 1, hal. 209–214, 2025.
[6] F. Fredilio, J. Rahmad, S. H. Sinurat, D. R. H. Sitompul, D. J. Ziegel, dan E. Indra, "Perbandingan Algoritma K-Nearest Neighbors (K-NN) dan Random forest terhadap Penyakit Gagal Jantung," *J. Teknol. Inform. dan Komput.*, vol. 9, no. 1, hal. 471–486, 2023, doi: 10.37012/jtik.v9i1.1432.
[7] S. Adi dan A. Wintarti, "Komparasi Metode Support Vector Machine (Svm), K-Nearest Neighbors (Knn), Dan Random Forest (Rf) Untuk Prediksi Penyakit Gagal Jantung," *MATHunesa J. Ilm. Mat.*, vol. 10, no. 2, hal. 258–268, 2022, doi: 10.26740/mathunesa.v10n2.p258-268.
[8] Kurniawan, A. (2020). Penerapan Algoritma K-Nearest Neighbor untuk Klasifikasi Data Mahasiswa. Jurnal Teknologi dan Sistem Komputer, 8(1), 55–62.
[9] Saputra, D., & Widodo, A. (2021). Analisis Klasifikasi Kelulusan Mahasiswa Menggunakan Random Forest. Jurnal Informatika dan Rekayasa Perangkat Lunak, 2(2), 45–51.
[10] Siregar, H. S., & Nurcahyo, A. (2022). Implementasi KNN dan Random Forest untuk Prediksi Hasil Belajar Siswa. Jurnal Ilmiah Teknologi Informasi Terapan, 9(3), 121–128.