PREDIKSI KELULUSAN SISWA SD 008 SUNGAI DUSUN MENGGUNAKAN METODE K-NEAREST NEIGHBOR DENGAN OPTIMASI NILAI K DAN 5-FOLD CROSS VALIDATION

Rahmathidayahtullah[1], Usman[2], Ilyas[3]
[1,2,3] Sistem Informasi, Fakultas Ilmu dan Komputer, Universitas Islam Indragiri
Jln. Provinsi Parit 1
Email: rahmathidayahrahmat92@gmail.com, Usmannounsky13211@gmail.com, IlyasDaeng@gmail.com

ABSTRAK

Prediksi kelulusan siswa merupakan bagian penting dalam manajemen pendidikan yang dapat mendukung pengambilan keputusan berbasis data. Penelitian ini menerapkan algoritma K-Nearest Neighbor (K-NN) dengan optimasi nilai K menggunakan metode 5-Fold Cross Validation untuk memprediksi status kelulusan siswa SD 008 Sungai Dusun. Dataset yang digunakan terdiri dari 212 data siswa dengan lima atribut nilai mata pelajaran: Pendidikan Agama Islam (PAI), Pendidikan Pancasila dan Kewarganegaraan (PPKn), Bahasa Indonesia, Matematika, dan IPA. Preprocessing data meliputi penanganan missing value, transformasi kategorik ke numerik, dan normalisasi Min-Max. Pengujian dilakukan pada K = 3, K = 5, dan K = 9. Hasil menunjukkan bahwa K = 9 memberikan performa terbaik dengan akurasi 96,70%, precision 95.90%, dan recall 98.32%. Analisis confusion matrix menunjukkan False Negative hanya 2 data, mengindikasikan model sangat efektif dalam mendeteksi siswa berisiko tidak lulus. Penelitian ini membuktikan bahwa K-NN dengan optimasi parameter dan cross validation merupakan pendekatan yang andal untuk prediksi kelulusan siswa sekolah dasar.

Kata Kunci : K-Nearest Neighbor, Cross Validation, Prediksi Kelulusan, Data Mining, Machine Learning

ABSTRACT

Student graduation prediction is an important aspect of educational management that can support data-driven decision making. This study applies the K-Nearest Neighbor (K-NN) algorithm with K-value optimization using 5-Fold Cross Validation to predict the graduation status of students at SD 008 Sungai Dusun. The dataset consists of 212 student records with five subject score attributes: Islamic Religious Education (PAI), Pancasila and Civics Education (PPKn), Indonesian Language, Mathematics, and Natural Science (IPA). Data preprocessing includes missing value handling, categorical-to-numeric transformation, and Min-Max normalization. Tests were conducted for K = 3, K = 5, and K = 9. Results show that K = 9 yields the best performance with 96.70% accuracy, 97.90% precision, and 98.32% recall. Confusion matrix analysis shows only 2 False Negatives, indicating the model is highly effective in detecting students at risk of not graduating. This study demonstrates that K-NN with parameter optimization and cross validation is a reliable approach for elementary school graduation prediction.

Keywords : *K-Nearest Neighbor, Cross Validation, Graduation Prediction, Data Mining, Machine Learning*

1 PENDAHULUAN

Pendidikan merupakan fondasi utama dalam pembentukan kualitas sumber daya manusia. Setiap satuan pendidikan, termasuk sekolah dasar, memiliki tanggung jawab untuk memastikan bahwa seluruh siswa memperoleh proses pembelajaran yang optimal agar mampu mencapai kelulusan sesuai dengan standar yang ditetapkan [1]. Namun dalam praktiknya, pencapaian

kelulusan siswa tidak selalu berjalan optimal. Perbedaan kemampuan akademik, tingkat kedisiplinan, serta pengaruh lingkungan belajar menjadi faktor yang dapat memengaruhi hasil belajar siswa.

SD 008 Sungai Dusun menghadapi tantangan serupa, di mana perbedaan performa siswa dari tahun ke tahun menunjukkan perlunya pendekatan yang lebih sistematis dalam memantau dan memprediksi potensi kelulusan sejak dini [2]. Pemanfaatan data siswa kini tidak lagi terbatas pada fungsi administrasi semata, tetapi dapat diolah untuk menghasilkan informasi prediktif melalui pendekatan data mining dan machine learning [3][4].

Algoritma K-Nearest Neighbor (K-NN) merupakan salah satu algoritma klasifikasi yang banyak digunakan dalam prediksi hasil belajar siswa [6]. Algoritma ini bekerja dengan melakukan klasifikasi berdasarkan kedekatan jarak antar data dan terbukti efektif pada berbagai kasus klasifikasi. Untuk memperoleh evaluasi yang lebih objektif dan stabil, penelitian ini menerapkan metode 5-Fold Cross Validation sebagai teknik evaluasi model [8][9].

Tujuan penelitian ini adalah: (1) mengetahui akurasi dan keandalan model K-NN melalui cross validation; (2) menerapkan teknik cross validation dalam evaluasi kinerja model K-NN; dan (3) menganalisis performa model berdasarkan nilai akurasi, precision, dan recall.

2 TINJAUAN PUSTAKA

2.1 Data Mining

Data mining merupakan proses penggalian informasi dan pengetahuan yang bernilai dari sekumpulan data dalam jumlah besar. Proses ini merupakan bagian dari Knowledge Discovery in Database (KDD) yang mencakup seleksi data, pembersihan data, transformasi data, pemodelan, serta evaluasi hasil [14]. Dalam bidang pendidikan, data mining banyak digunakan untuk menganalisis data akademik siswa guna mendukung pengambilan keputusan berbasis data, termasuk prediksi kelulusan [15].

2.2 K-Nearest Neighbor (K-NN)

K-Nearest Neighbor (K-NN) adalah algoritma klasifikasi supervised learning yang mengklasifikasikan data baru berdasarkan mayoritas kelas dari K tetangga terdekatnya. Jarak antar data dihitung menggunakan Euclidean Distance. Kelebihan K-NN adalah kesederhanaan implementasi dan efektivitasnya pada data numerik [7]. Kelemahan utamanya adalah sensitivitas terhadap nilai K dan kebutuhan komputasi yang meningkat seiring bertambahnya data.

$$d_{(xy)} = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$$

Ket:

d : Nilai jarak (kedekatan)

x : Data latih

y : Data uji

n : Total atribut 1 . . . n

i : Atribut nilai antara 1 . . . n

2.3 Cross Validation

K-Fold Cross Validation adalah teknik evaluasi model yang membagi dataset menjadi K bagian (fold) secara seimbang. Pada setiap iterasi, satu fold digunakan sebagai data uji dan sisanya sebagai data latih. Proses ini diulang sebanyak K kali dan hasilnya dirata-ratakan. Teknik ini memberikan estimasi performa yang lebih stabil dan mengurangi bias akibat pembagian data yang tidak merata [8]. Penelitian ini menggunakan 5-Fold Cross Validation.

2.4 Penelitian Terdahulu

Beberapa penelitian sebelumnya telah mengkaji penggunaan K-NN dalam konteks pendidikan. Kartarina et al. (2021) memperoleh akurasi 96,18% menggunakan K-NN pada data mahasiswa. Budiyantara et al. (2020) mencapai akurasi 90,00% dalam prediksi kelulusan mahasiswa dengan perbandingan beberapa algoritma. Asriningtas et al. (2025) memperoleh akurasi 84,00% pada prediksi kelulusan tepat waktu mahasiswa D3. Etriyanti (2021) melaporkan akurasi 53,08% pada prediksi lama studi mahasiswa menggunakan K-NN. Penelitian ini berkontribusi dengan menambahkan optimasi nilai K secara sistematis dan evaluasi komprehensif menggunakan cross validation [2][5][3].

3 METODE PENELITIAN

3.1 Dataset

Dataset dalam penelitian ini diperoleh dari SD 008 Sungai Dusun, mencakup data siswa periode 2020-2025. Dataset terdiri dari 212 record dengan enam atribut, yaitu lima atribut nilai mata pelajaran (PAI, PPKn, Bahasa Indonesia, Matematika, IPA) dan satu atribut target (status kelulusan: LULUS/TIDAK LULUS).

Tabel 1. Atribut Dataset

No.	Atribut	Tipe Data
1	PAI	Numerik
2	PPKn	Numerik
3	Bahasa Indonesia	Numerik
4	Matematika	Numerik
5	IPA	Numerik
6	Kelulusan (Target)	Kategorik

3.2 Preprocessing Data

Tahapan preprocessing meliputi: (1) Penanganan Missing Value - dilakukan pemeriksaan dan pembersihan data yang tidak lengkap; (2) Transformasi Kategorik ke Numerik - atribut label kelulusan diubah ke format numerik untuk kebutuhan komputasi; (3) Normalisasi Min-Max - seluruh atribut numerik dinormalisasi ke rentang [0, 1]

3.3 Penerapan Algoritma K-NN dan Cross Validation

Model K-NN diuji dengan tiga variasi nilai K (K = 3, K = 5, K = 9) menggunakan 5-Fold Cross Validation. Pada setiap fold, model dilatih pada 4 fold dan diuji pada 1 fold. Metrik evaluasi yang digunakan meliputi akurasi, precision, recall, dan confusion matrix. Tools yang digunakan adalah AI Studio. Ai Studio merupakan platform analisis data berbasis visual yang digunakan untuk membangun, menguji, dan mengevaluasi model machine learning tanpa harus melakukan pemrograman secara kompleks Melalui antarmuka berbasis *drag-and-drop*.

4 HASIL DAN PEMBAHASAN

4.1 Hasil Preprocessing

Setelah proses preprocessing, dataset akhir terdiri dari 212 data bersih tanpa missing value. Seluruh atribut numerik telah dinormalisasi ke rentang 0-1, sedangkan atribut label tetap dalam bentuk kategorik. Dataset siap digunakan untuk proses klasifikasi dan evaluasi.

4.2 Penentuan Nilai K Optimal

Pengujian dilakukan terhadap K = 3, K = 5, dan K = 9. Hasil perbandingan performa setiap nilai K disajikan pada Tabel 2.

Tabel 2. Perbandingan Performa Nilai K

Nilai K	Akurasi	Precision	Recall
3	96,23%	95.87%	97.48%
5	95,78%	95.08%	97.48%
9*	96,70%*	97,90%*	98,32%*

**Nilai K terbaik*

Pengujian dilakukan menggunakan tiga variasi nilai tetangga terdekat, yaitu K = 3, K = 5, dan K = 9. Hasil pengujian menunjukkan bahwa K = 9 memberikan performa terbaik dengan akurasi 96,70%, precision 95,90%, dan recall 98,32%.

Secara teoritis, nilai K yang lebih kecil seperti K = 3 cenderung sangat dipengaruhi oleh tetangga terdekat sehingga lebih sensitif terhadap variasi lokal pada data. Pada dataset penelitian ini, terdapat distribusi kelas yang tidak sepenuhnya seimbang, di mana jumlah siswa **TIDAK LULUS** lebih sedikit dibandingkan kelas **LULUS**. Kondisi tersebut menyebabkan beberapa data pada kelas minoritas lebih mudah dipengaruhi oleh tetangga dari kelas mayoritas sehingga prediksi menjadi kurang stabil ketika menggunakan nilai K yang kecil.

Sementara itu, K = 5 mulai mengurangi sensitivitas terhadap variasi lokal, namun pengaruh data yang berada di sekitar batas antar kelas masih cukup besar sehingga peningkatan performanya belum optimal. Pada K = 9, keputusan klasifikasi ditentukan oleh lebih banyak tetangga sehingga pengaruh data yang menyimpang (outlier) maupun variasi individu menjadi lebih kecil. Selain itu, proses voting melibatkan lebih banyak data sehingga pola distribusi kelas pada dataset dapat direpresentasikan dengan lebih baik. Kondisi ini menghasilkan model yang lebih stabil dan memiliki kemampuan generalisasi yang lebih baik, yang tercermin dari nilai akurasi, precision, dan recall tertinggi dibandingkan nilai K lainnya.

4.3 Evaluasi Model dengan Cross Validation

Evaluasi model menggunakan 5-Fold Cross Validation dengan K = 9 menghasilkan:
Akurasi: 96,70% ± 3,52%
Precision: 95.90% ± 6.19%
Recall: 98.32% ± 2.33%

Standar deviasi yang relatif kecil pada akurasi dan recall menunjukkan model yang stabil. Standar deviasi Precision yang lebih tinggi (±6.19) mengindikasikan adanya ketidakseimbangan distribusi kelas antar fold, kemungkinan disebabkan oleh jumlah sampel kelas TIDAK LULUS yang lebih sedikit. Standar deviasi precision yang lebih tinggi (±6,19) menunjukkan bahwa nilai precision lebih bervariasi pada setiap fold. Hal ini terjadi karena jumlah data pada kelas minoritas (TIDAK LULUS) relatif lebih sedikit, sehingga perubahan kecil pada jumlah prediksi yang salah (false positive) dapat memberikan pengaruh yang cukup besar terhadap nilai precision.

4.4 Analisis Confusion Matrix

Tabel 3. Hasil Confusion Matrix (K=9)

Kategori	Label	Jumlah
True Positive (TP)	LULUS benar	117
True Negative (TN)	TIDAK LULUS benar	88
False Positive (FP)	LULUS → TIDAK LULUS	5
False Negative (FN)	TIDAK LULUS → LULUS	2

Model berhasil mengklasifikasikan 205 dari 212 data dengan benar. False Negative (FN = 2) merupakan kesalahan paling kritis dalam konteks pendidikan karena berarti ada 2 siswa yang berpotensi tidak lulus namun diprediksi lulus sehingga tidak mendapat perhatian. False Positive (FP = 5) bersifat lebih aman karena hanya mengakibatkan intervensi tambahan yang tidak diperlukan. Secara keseluruhan, model terbukti sangat andal dengan tingkat kesalahan yang sangat rendah.

4.5 Perbandingan dengan Penelitian Terdahulu

Tabel 4. Perbandingan dengan Penelitian Terdahulu

Penelitian	Algoritma	Akurasi (%)
Penelitian ini (2026)	K-NN (K=9) + 5-Fold CV	96,72
Kartarina et al. (2021)	K-NN	96,18
Meriyana et al. (2025)	K-NN (k=1)	97,90
Asriningtas et al. (2025)	K-NN	84,00
Budiyantara et al. (2020)	K-NN	90,00
Etriyanti (2021)	K-NN	53,08

Hasil penelitian ini (96,72%) kompetitif dibandingkan penelitian terdahulu dan melampaui sebagian besar studi yang dikaji. Akurasi sedikit di bawah Meriyana et al. (2025) yang mencapai 97,90%, namun perbedaan konteks data (klasifikasi stunting vs. kelulusan sekolah dasar) menjadikan perbandingan langsung perlu dilakukan secara hati-hati. Keunggulan penelitian ini terletak pada penerapan evaluasi komprehensif melalui kombinasi optimasi nilai K dengan cross validation, menghasilkan model yang tidak hanya akurat tetapi juga stabil.

4.6 Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan: (1) cakupan data terbatas pada satu sekolah sehingga generalisasi ke sekolah lain perlu dikaji lebih lanjut; (2) pengujian hanya pada K = 3, 5, dan 9, nilai K lain seperti K = 7 dan K = 11 belum dieksplorasi; (3) ketidakseimbangan kelas belum ditangani dengan teknik seperti SMOTE; (4) hanya menggunakan Euclidean Distance; dan (5) tidak ada perbandingan dengan algoritma lain pada dataset yang sama.

5 KESIMPULAN

Penelitian ini telah berhasil menerapkan algoritma K-Nearest Neighbor (K-NN) dengan optimasi nilai K menggunakan 5-Fold Cross Validation untuk memprediksi kelulusan siswa SD 008 Sungai Dusun. Kesimpulan yang diperoleh adalah:

1) Nilai K = 9 memberikan performa terbaik dengan akurasi 96,70%, precision 95.90%, dan recall 98.32%, menunjukkan keseimbangan optimal antara bias dan variansi.
2) Metode 5-Fold Cross Validation menghasilkan evaluasi yang objektif dan stabil, dengan standar deviasi akurasi ±3,52% dan recall ±2.33%.
3) Confusion matrix menunjukkan hanya 2 False Negative dari 212 data, membuktikan efektivitas model dalam mendeteksi siswa yang berpotensi tidak lulus.
4) Model ini layak diterapkan sebagai sistem pendukung keputusan dini bagi sekolah dalam mengidentifikasi siswa yang membutuhkan perhatian akademik tambahan.

Penelitian selanjutnya disarankan untuk memperluas cakupan data, menambah variasi atribut (kehadiran, kedisiplinan), menerapkan teknik penanganan class imbalance seperti SMOTE, serta membandingkan K-NN dengan algoritma lain seperti Random Forest atau SVM.

REFERENSI

[1] M. Jafar and D. Aisyah, "Muslim Jafar, Devy Aisyah, Amrina," vol. 2338, no. 02, pp. 13-34, 2022.

[2] S. R. Asriningtas, D. U. Surya, and E. R. N. Wulandari, "Pemodelan Prediksi Kelulusan Tepat Waktu Mahasiswa D3," J. Tek. Elektro dan Inform., vol. 20, pp. 63-73, 2025.

[3] E. Etriyanti, "Perbandingan Tingkat Akurasi Metode KNN dan Decision Tree dalam Memprediksi Lama Studi Mahasiswa," J. Ilm. Bina STMIK Bina Nusant. Jaya Lubuklinggau, vol. 3, no. 1, pp. 6-14, 2021.

[4] N. Maftucha, S. Salma, N. Rahmayuna, and N. Wakhidah, "Perbandingan Algoritma Machine Learning dalam Memprediksi Kelulusan Siswa," J. Tekno Kompak, vol. 19, no. 2, pp. 116-128, 2025.

[5] A. Budiyantara, I. Irwansyah, E. Prengki, P. A. Pratama, and N. Wiliani, "Komparasi Algoritma Decision Tree, Naive Bayes dan K-Nearest Neighbor untuk Memprediksi Mahasiswa Lulus Tepat Waktu," JITK, vol. 5, no. 2, pp. 265-270, 2020.

[6] A. Muni and M. Jibril, "Penerapan Metode Naive Bayes Classifier dalam Pemilihan Kualitas Bibit Kelapa," J. Perangkat Lunak, vol. 5, no. 3, pp. 313-322, 2023.

[7] M. D. Hendriyanto and B. N. Sari, "Penerapan Algoritma K-Nearest Neighbor dalam Klasifikasi Judul Berita Hoax," J. Ilm. Inform., vol. 10, no. 02, pp. 80-84, 2022.

[8] A. S. Fauzan, A. I. P. Sari, and I. Ali, "Analisis Perbandingan Algoritma Decision Tree dan Naive Bayes untuk Mengevaluasi Prestasi Belajar Siswa," JATI, vol. 8, no. 1, pp. 741-747, 2024.

[9] M. R. Arrohman et al., "Analisis Klasifikasi Kelulusan Siswa Menggunakan Metode KNN dan Random Forest," JMSI, vol. 6, no. 2, pp. 299-305, 2025.

[10] D. I. S. D. N. Balerejo, "Klasifikasi Kelulusan Siswa SD menggunakan KNN," vol. 10, pp. 1123-1129, 2025.

[11] F. Sugandi, "Prediksi Kelulusan Mahasiswa STMIK Dharmawacana menggunakan Algoritma K-Nearest Neighbors," Sienna, vol. 4, no. 1, pp. 20-26, 2023.

[12] A. E. Gumanti, Taslim, S. Handayani, and T. Dafwen, "Penerapan Algoritma K-NN untuk Klasifikasi Topik Skripsi Mahasiswa," J. ITACS, vol. 1, pp. 44-50, 2022.

[13] H. Latifah and S. Mujiyono, "Perbandingan Algoritma Naive Bayes, K-NN, ID3, dan SVM dalam Menentukan Prediksi Kelulusan Siswa," J. Mahasiswa Tek. Inform., vol. 2, no. 1, pp. 38-45, 2023.

[14] I. P. Tafonao and A. F. K. Sibero, "Teknik Klasifikasi Prediksi Kelulusan Mahasiswa Sistem Informasi menggunakan K-NN," J. Mahajana Inf., vol. 7, no. 1, pp. 83-90, 2022.

[15] W. Lastari, "Penerapan Data Mining untuk Memprediksi Prestasi Siswa SMA pada Dinas Pendidikan Provinsi Jambi," vol. 8, no. 2, pp. 310-321, 2023.

[16] D. Supriadi, Y. Yupianti, and E. Suryana, "Penerapan K-Nearest Neighbor untuk Klasifikasi Tingkat Kelulusan Siswa SMA Negeri 11 Bengkulu," J. Media Infotama, vol. 19, no. 2, pp. 472-478, 2023.