**Implementasi Algoritma KNN untuk Memprediksi Performa Siswa Sekolah**

I Made Ryan Prana Dhitaa1, Gst. Ayu Vida Mastrika Giria2

aProgram Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,

Universitas Udayana

Jalan Raya Kampus Udayana, Bukit Jimbaran, Kuta Selatan, Badung, Bali Indonesia

1ryanprana555@gmail.com

2vida@unud.ac.id

***Abstract***

*One of the factors that influences students graduation rates is their performance in learning. Predicting graduation rates based on student performance has the benefit of analyzing academically underperforming students and providing support to students who face difficulties in the learning process. There are several factors to consider in predicting students' graduation rates, such as academic grades, attitudes, and social factors. However, these factors alone are not sufficient to effectively predict students' performance, and educators also struggle to identify which factors affect students' performance.To predict the performance of school students, the K-Nearest Neighbor (KNN) method is utilized. The K-Nearest Neighbor method is often used in classifying students' performance due to its simplicity and ability to produce significant and competitive results. In this research, the prediction of students' graduation rates is carried out using the KNN method.The results of implementing the prediction of students' performance using the KNN method can serve as a reference for students to improve their achievements and assist educators in considering future teaching materials.*

***Keywords:*** *KNN, K-Nearest Neighbor, Students Performance, Student*

1. **Pendahuluan**

Saat ini, pendidikan adalah salah satu bagian terpenting dan esensial bagi kehidupan masyarakat. Ini digunakan untuk meningkatkan dan meningkatkan pertumbuhan individu secara akademis dan finansial. Seorang individu yang berpendidikan harus berkontribusi tidak hanya untuk keluarganya tetapi juga untuk masyarakat dan komunitas. Ini semua dicapai melalui pembelajaran yang tepat [1].

Untuk dapat menghasilkan SDM (sumber daya manusia) yang cakap, berwawasan, kompetitf dan kreatif, lembaga pendidikan diminta untuk menyelenggarakan pendidikan yang bermutu dan berkualitas bagi peserta didiknya [2]. Sehingga untuk mengimplementasikannya dilakukan beberapa perubahan kurikulum guna meningkatkan kualitas pada tingkatan pendidikan tentunya untuk mengetahui berhasil atau tidaknya kurikulum yang telah dibuat diperlukan sebuah evaluasi berupa prediksi hasil belajar siswa di sekolah serta faktor-faktor yang mempengaruhinya.

Maka untuk mewujudkan hal tersebut dibutuhkan suatu alat analisis berupa komputasi cerdas untuk menganalisis bagaimana kinerja siswa, faktor mana yang akan mempengaruhi kinerja mereka, dengan cara apa siswa dapat membuat kemajuan, dan apakah siswa memiliki potensi untuk tampil lebih baik [2].

Pada penelitian ini akan dilakukan klasifikasi performa siswa berdasarkan lingkungan serta hasil akademik siswa dengan K-Nearest Neighbor (KNN) algoritma ini digunakan karena kemampuannya dalam menyederhanakan perhitungan algoritma dan mengoptimalkan waktu. KNN adalah metode yang bekerja dengan mengelompokkan data baru berdasarkan jarak mereka dengan data lainnya. Prediksi akan dilakukan menggunakan algoritma K-Nearest Neighbor (KNN).

1. **Metode Penelitian**

Penelitian dimulai dari tahapan pengumpulan dataset lalu pemilihan algoritma yaitu algoritma K-Nearest Neighbor, melakukan normalisasi data serta data encoding, klasifikasi lalu evaluasi hasil dari data yang telah diproses, gambar 1. Tahapan penelitian merupakan ilustrasi langkah- langkah metode yang akan dikerjakan.



**Gambar 1.** Tahapan Penelitian

* 1. **Data dan Sumber Data**

Penelitian ini menggunakan dataset Student Performance terkait dengan latar belakang serta nilai siswa sekolah yang didapat dari platform kaggle tepatnya pada halaman berikut <https://www.kaggle.com/datasets/rkiattisak/student-performance-in-mathematics>

Dataset Student Performance terkait yang telah diperoleh nantinya pada penelitian ini dibagi menjadi dua yaitu data training sebanyak 85% serta testing 15% dari keseluruhan data yang telah diperoleh

* 1. **Analisis Data**

Tahap ini bertujuan untuk memastikan integritas data sehingga nantinya tidak menimbulkan masalah pada proses data training atau pelatihan data. data latar belakang siswa sekolah yang digunakan sejumlah 1000 data yang terbagi menjadi 4 kelas yaitu Sangat Kurang, Kurang, Cukup Baik, dan Baik. pada tahap ini dilakukan penghilangan data atau seleksi data dengan menyeleksi atribut apa saja yang diperlukan dalam dataset yang telah didapat terdapat 9 atribut yaitu atribut gender, race/ethnicity, parental level of education, lunch, test preparation course, math score, reading score, dan average.



**Gambar 2.** Sampel Data Student Performance

* 1. **Pre-Processing**

Pada tahapan ini tujuan utama dari data pre-processing adalah. proses identifikasi, pembenaran, dan/atau penghapusan data yang tidak akurat, tidak lengkap, tidak konsisten, atau tidak relevan dari sebuah dataset. Tujuan utama dari data cleansing adalah memastikan bahwa data yang digunakan dalam analisis atau pemodelan adalah data yang berkualitas tinggi, dapat diandalkan, dan tepat [3].

1. **Normalisasi Data**

Normalisasi data melibatkan mengubah skala data agar sesuai dengan rentang atau standar yang ditetapkan. Normalisasi dapat melibatkan pemetaan data ke rentang yang spesifik atau transformasi data untuk menghilangkan bias atau asumsi tertentu. Metode min-max normalization digunakan untuk normalisasi data pada penelitian ini dengan mentransformasikan setiap nilai dalam rentang data ke rentang baru antara 0 dan 1 [2]. Nilai terendah dalam data akan menjadi 0, sedangkan nilai tertinggi akan menjadi 1. Berikut merupakan persamaan matematis dari min-max normalization.

|  |  |
| --- | --- |
| $$X\_{normalized}=\frac{X- X\_{min}}{X\_{max}- X\_{min}}$$ | (1) |

Keterangan:

X = nilai asli dari data yang akan dinormalisasi

Xnormalizes = nilai yang dinormalisasi dari X

Xmin = nilai terendah dalam rentang data

Xmax = nilai tertinggi dalam rentang data

1. **Data Encoding**

Pada tahapan ini proses mengubah variabel kategori atau kualitatif menjadi bentuk numerik agar dapat digunakan dalam analisis dan pemodelan data. Hal ini diperlukan karena sebagian besar algoritma dan model yang digunakan dalam analisis data memerlukan data dalam bentuk numerik. berikut merupakan perubahan data dari masing-kelas dan atribut yang digunakan.

 **Tabel 1.** Contoh Kelas dan Kategori Nilai

| **Kelas** | **Nilai Kategori** |
| --- | --- |
| Kategori 1 | 0 |
| Kategori 2 | 1 |
| Kategori 3 | 2 |
| Kategori 4 | 3 |

**Tabel 2.** Contoh Atribut dan Kategori Nilai

|  |  |
| --- | --- |
| **Atribut** | **Nilai Kategori** |
| Parental level | 0 - 4 |
| Lunch | 0 - 1 |
| Test Preparation | 0 - 1 |

Tabel 1 menunjukkan nilai kelas yang telah melalui proses data encoding pada dataset yang telah di processing dengan data encoding, sedangkan tabel 2 menunjukan nilai atribut yang telah melalui proses encoding

1. **Klasifikasi**

Pada tahap ini dataset dipisah menjadi 2 bagian secara acak yaitu: data pelatihan (data training) dan data pengujian (data testing). Data pelatihan (data Training): Data pelatihan digunakan untuk melatih model KNN. pada tahap ini dataset dipisah menjadi 2 bagian secara acak yaitu: data pelatihan (data training) dan data pengujian (data testing).Data pelatihan (data Training): Data pelatihan digunakan untuk melatih model KNN. Data pelatihan terdiri dari contoh-contoh data yang memiliki label kelas yang diketahui. Model KNN akan menggunakan data ini untuk belajar pola dan hubungan antara fitur-fitur yang ada dalam dataset dengan label kelas yang sesuai. sedangkan data pengujian (data Testing) merupakan Data pengujian digunakan untuk menguji kinerja model KNN yang telah dilatih [2][4].

Data pengujian terdiri dari contoh-contoh data yang juga memiliki label kelas yang diketahui, tetapi label kelasnya disembunyikan dari model saat proses pengujian. Model KNN akan memprediksi label kelas untuk data pengujian berdasarkan informasi yang telah dipelajari selama tahap pelatihan. Pemisahan data dilakukan secara acak untuk memastikan bahwa data pelatihan dan data pengujian mewakili dataset secara proporsional. Biasanya, pemisahan data dilakukan dengan membagi dataset secara acak menjadi dua bagian, di mana sebagian besar data digunakan untuk pelatihan (misalnya 70-80% dari dataset) dan sisanya digunakan untuk pengujian (misalnya 20-30% dari dataset). Namun, perbandingan ini dapat bervariasi tergantung pada ukuran dataset dan kebutuhan spesifik [2][5][6]

1. **Hasil dan Pembahasan**
	1. **Data Encoding**

Fitur pada dataset yang memiliki atribut data kualitatif akan diubah kedalam bentuk numerik dengan menggunakan metode one shot encoding pada setiap data sehingga dengan demikian data akan dapat diolah ke dalam perhitungan yang akan dilakukan. Berikut merupakan gambar encoding dataset student performance.



**Gambar 3.** Dataset Sebelum Melakukan Data Encoding



**Gambar 4.** Dataset Sesudah Melakukan Data Encoding

* 1. **Normalisasi data**

Nilai siswa sekolah bervariasi dapat menyulitkan sistem dalam melakukan perhitungan sehingga normalisasi data perlu dilakukan pada nilai hasil ujian siswa sekolah yang pada penelitian ini metode min-max normalization digunakan melakukan normalisasi pada data. Metode ini dilakukan apa setiap fitur dari data variabel yang memerlukan normalisasi pada data dengan mengubah skala data ke dalam rentang antara. berikut merupakan normalisasi yang dilakukan pada fitur nilai dari dataset.



**Gambar 5.** Mix-Max Normalization

* 1. **Pengujian dan Evaluasi**

Pada tahap ini, pengujian dan evaluasi dilakukan untuk mengklasifikasikan sejumlah data uji menggunakan model klasifikasi yang telah dibentuk sebelumnya. Pengujian dilakukan dengan jumlah (k-5), di mana k adalah jumlah data uji yang digunakan. Dalam kasus ini, k-5 adalah data uji yang digunakan karena dalam beberapa percobaan yang telah di lakukan k-5 memiliki precision serta recall data yang lebih baik dibandingkan k lainya dalam pengujian dalam penelitian ini.



**Gambar 6.** Hasil Prediksi dengan Nilai K-5

Total data uji yang digunakan adalah 150, yang terbagi menjadi empat kelas berdasarkan distribusinya. Kelas "baik" memiliki 25 data, kelas "cukup baik" memiliki 37 data, kelas "kurang" memiliki 68 data, dan kelas "sangat kurang" memiliki 16 data. Untuk mengukur kinerja model klasifikasi yang telah dibentuk, digunakan metode confusion matrix. Confusion matrix adalah metode yang digunakan untuk mengevaluasi performa model klasifikasi dengan membandingkan hasil prediksi model dengan nilai sebenarnya dari data uji. Gambar 7 menunjukkan hasil dari confusion matrix yang diperoleh setelah melakukan pengujian terhadap model klasifikasi yang telah dibentuk sebelumnya.



**Gambar 7.** Mix-Max Normalization

Keterangan :

0 = Sangat Kurang

1 = Kurang

2 = Cukup Baik

3 = Baik

Berdasarkan confusion matrix seperti yang terlihat pada Gambar 7, kita dapat mengetahui bahwa pada kelas "Sangat Kurang" dengan 16 sampel data, semua data terprediksi dengan benar sebagai kelas "Sangat Kurang" dan tidak ada data yang salah diprediksi sebagai kelas lain. Pada kelas "Kurang" dengan 69 sampel data, 67 data terprediksi dengan benar sebagai kelas "Kurang", namun terdapat 1 data yang salah diprediksi sebagai kelas "Sangat Kurang" dan 1 data yang diprediksi sebagai kelas “Cukup Baik” . Pada kelas "Cukup Baik" dengan 38 sampel data, 37 data terprediksi dengan benar sebagai kelas "Cukup Baik", tetapi terdapat 1 data yang salah diprediksi sebagai kelas "Baik". Pada kelas "Baik" dengan 27 sampel data, 25 data terprediksi dengan benar sebagai kelas "Baik", namun terdapat 2 data yang salah diprediksi sebagai kelas "Cukup Baik".

Dari confusion matrix tersebut, akurasi model klasifikasi dapat dihitung berdasarkan persentase nilai yang diprediksi dengan benar dibagi dengan total jumlah data uji [4].

|  |  |
| --- | --- |
| $$Average Accuracy= \frac{\sum\_{i=1}^{l}\frac{tp\_{i}+ tn\_{i}}{tp\_{i}+ ƒp\_{i}+ ƒp\_{i}+ tn\_{i}}}{l}$$ | (2) |

Setelah melakukan perhitungan tersebut akan di dapatkan data akurasi setiap kelas dalam system yang telah dibuat seperti yang terlihat pada tabel berikut

**Tabel 3.** Kinerja Model Klasifikasi

|  |  |  |  |
| --- | --- | --- | --- |
| **Kelas** | **Accuracy** | **Precision** | **Recall** |
| Sangat Kurang | 1.0000 | 1.0000 | 1.0000 |
| Kurang | 0.9933 | 0.9709 | 0.9963 |
| Cukup Baik | 0.9867 | 0.9880 | 0.9855 |
| Baik | 0.9733 | 0.9580 | 0.9734 |
| **Rata-rata** | **0.9667** | **0.9569** | **0.9677** |

1. **Kesimpulan**

Penelitian ini menyimpulkan bahwa telah berhasil dibuat sebuah aplikasi untuk melakukan seleksi data penerima beasiswa dengan menggunakan algoritma KNN. Evaluasi algoritma KNN menggunakan metode confusion matrix menunjukkan hasil rata-rata akurasi dari metode KNN ini sebesar 96%. Ini menunjukan bahwa implementasi algoritma KNN pada prediksi peforma siswa sekolah memiliki akurasi yang cukup tinggi dengan demikian klasifikasi peforma dengan KNN dapat dijadikan sebagai system pendukung untuk membantu guru serta instansi pendidikan untuk mempertimbangkan langkah-langkah yang dapat dilakukan untuk pemajukan pendidikan di Indonesia kedepannya

**Daftar Pustaka**

1. Kurniawati, F. N. A., 2022. Meninjau Permasalahan Rendahnya Kualitas Pendidikan Di Indonesia Dan Solusi. Academy of Education Journal, Volume 13, p. 13.
2. Daru Prasetyawan, ,. R. G., 2022. K-Nearest Neighbor untuk Memprediksi Prestasi Mahasiswa Berdasarkan Latar Belakang Pendidikan dan Ekonomi. Jurnal Informatika Sunan Kalijaga, Volume 7, p. 12.
3. Ni Made Rika Padeswari Kusuma, L. G. A., 2022. Implementasi Algoritma K-Nearest Neighbor (K-NN) dalam Deteksi Dini Penyakit Hepatitis C. JNATIA (Jurnal Nasional Teknologi Informasi dan Aplikasinya), Volume 1, p. 8. [3]
4. Deti Fusvita1), A. 2., 2021. Penerapan Algoritma KNN (K-Nearest Neighbour) Dalam Klasifikasi Data Pinjaman ANggota Koprasi. Jurnal Ilmiah Binary STMIK Bina Nusantara Jaya, Volume 03, p. 5.
5. Sebastian Raschka, V. M., 2019. Python Machine Learning. Third Edition ed. s.l.:Packt.
6. Mutiara Ayu Banjarsari, H. I. B. A. F., 2015. Penerapan K-Optimal Pada Algoritma Knn untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu Komputer Fmipa Unlam Berdasarkan IP Sampai Dengan Semester 4. jurnaL Ilmu Komputer (KLIK), Volume 02, p. 15.
7. Lubis, A. R., Lubis, M., & Khowarizmi, A.-. (2020). Optimization of distance formula in K-Nearest Neighbor method. Bulletin of Electrical Engineering and Informatics, 9(1), 326–338. <https://doi.org/10.11591/eei.v9i1.1464>

Halaman ini sengaja dibiarkan kosong