e-ISSN: 2964-2663
p-ISSN: 2964-9692

Penanganan Missing Value dan Perbandingan Performa Algoritma *Naïve Bayes* serta Algoritma *Decision Tree* dalam Kelulusan Mahasiswa

Sulistyaningrum Sulistyaningrum[*1], Hasbi Firmansyah[2], Eko Budi Raharjo[3], Wildani Eko Nugroho[4]
[*1,2,3,4]*Informatika, Universitas Pancasakti Tegal, Kota Tegal*
e-mail: [*1]sulistyaningrum7868@gmail.com, [2]hasbyfirmansyah@upstegal.ac.id, [3]ekobudiraharjo@yahoo.com, [4]wild4n1@gmail.com

Abstrak

Penelitian ini membahas penanganan missing value serta perbandingan performa algoritma Naïve Bayes dan Decision Tree dalam memprediksi kelulusan mahasiswa. Dataset yang digunakan mencakup data akademik mahasiswa yang dimanipulasi untuk mensimulasikan missing value. Metode imputasi, seperti Mean Imputation, K-Nearest Neighbors, dan Iterative Imputation, diterapkan untuk menangani nilai yang hilang. Evaluasi dilakukan dengan menggunakan metrik akurasi, presisi, recall, dan F1-score. Hasil penelitian menunjukkan bahwa algoritma Decision Tree memiliki performa lebih unggul dibandingkan Naïve Bayes, dengan akurasi mencapai 92,1% dibandingkan 85,3% pada Naïve Bayes. Keunggulan ini menunjukkan bahwa Decision Tree lebih efektif dalam menangkap pola data dengan hubungan antar fitur yang kompleks. Studi ini memberikan kontribusi terhadap pengembangan metode prediksi berbasis data untuk mendukung kebijakan akademik, termasuk penanganan missing value yang optimal dan pemilihan algoritma yang tepat.
***Kata Kunci**: Algoritma Decision Tree, Imputasi, Kelulusan Mahasiswa, Naïve Bayes, Missing Value*

Abstract

This study examines the handling of missing values and compares the performance of the Naïve Bayes and Decision Tree algorithms in predicting student graduation. The dataset includes academic records that were manipulated to simulate missing values. Imputation methods such as Mean Imputation, K-Nearest Neighbors, and Iterative Imputation were applied to address missing data. The evaluation utilized metrics such as accuracy, precision, recall, and F1-score. The results indicate that the Decision Tree algorithm outperforms Naïve Bayes, achieving an accuracy of 92.1% compared to 85.3% for Naïve Bayes. This superiority highlights that Decision Tree is more effective in capturing data patterns with complex inter-feature relationships. This study contributes to the development of data-driven prediction methods to support academic policies, including optimal missing value handling and the selection of appropriate algorithms.
***Keywords**: Decision Tree Algorithm, Graduation Prediction, Imputation, Missing Value, Naïve Bayes*

I. PENDAHULUAN

Dalam era pendidikan berbasis data, penggunaan data untuk memprediksi keberhasilan akademik menjadi semakin penting. Salah satu tantangan utama yang dihadapi adalah adanya missing value dalam dataset pendidikan, yang sering muncul akibat proses pengumpulan data yang tidak sempurna atau kerusakan data (Emmanuel et al., 2021). Keberadaan nilai yang hilang ini dapat mengganggu kinerja model pembelajaran mesin, sehingga penting untuk mengidentifikasi dan menerapkan metode penanganan yang efektif (Ahmad et al., 2024).

Namun, meskipun berbagai metode imputasi telah dikembangkan, masih terdapat perdebatan mengenai pendekatan mana yang paling sesuai untuk meningkatkan akurasi model pembelajaran mesin, khususnya dalam konteks pendidikan (Palanivinayagam & Damaševičius, 2023). Selain itu, algoritma seperti *Naïve Bayes* dan *Decision Tree* memiliki karakteristik yang berbeda dalam menangani data dengan *missing value*, sehingga penting untuk membandingkan performa kedua algoritma tersebut secara sistematis (Gond et al., 2021).

e-ISSN: 2964-2663
p-ISSN: 2964-9692

Untuk menjawab permasalahan ini, penelitian ini mengusulkan pendekatan sistematis untuk menangani missing value menggunakan metode imputasi terkini. Penelitian ini juga akan membandingkan kinerja algoritma *Naïve Bayes* dan *Decision Tree* dalam memprediksi kelulusan mahasiswa, sehingga dapat memberikan wawasan yang lebih baik terkait algoritma yang lebih efektif dalam konteks data pendidikan (Emmanuel et al., 2021).

Tujuan penelitian ini adalah untuk mengidentifikasi metode penanganan missing value yang paling efektif dan membandingkan performa algoritma *Naïve Bayes* dan *Decision Tree* dalam memprediksi kelulusan mahasiswa.

Studi sebelumnya telah menunjukkan bahwa metode imputasi seperti mean imputation dan algoritma berbasis *k-nearest neighbors* memiliki keunggulan tertentu dalam meningkatkan akurasi model (Hanyf & Silkan, 2024). Selain itu, algoritma *Decision Tree* telah terbukti lebih tahan terhadap missing value dibandingkan algoritma lain seperti Naïve Bayes dalam situasi tertentu (Chen & McCoy, 2024).

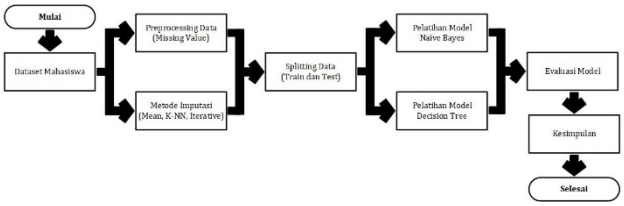
Penelitian terbaru telah berfokus pada pengembangan metode imputasi berbasis pembelajaran mesin, seperti iterative imputation dan pendekatan berbasis jaringan saraf, untuk meningkatkan akurasi prediksi (Alabadla et al., 2022). Selain itu, teknik seperti SMOTE *(Synthetic Minority Oversampling Technique)* telah diterapkan untuk mengatasi ketidakseimbangan kelas dan meningkatkan kinerja model dalam data pendidikan (Dharmasaputro et al., 2022).

Penelitian ini memberikan kontribusi dalam menawarkan pendekatan baru untuk penanganan missing value pada dataset pendidikan dan menyediakan analisis komparatif mendalam antara algoritma *Naïve Bayes* dan *Decision Tree* dalam kasus prediksi kelulusan mahasiswa.

II. METODE PENELITIAN

Desain Penelitian

Penelitian ini menggunakan desain eksperimen kuantitatif untuk mengevaluasi performa algoritma *Naïve Bayes* dan *Decision Tree* dalam menangani missing value pada dataset kelulusan mahasiswa. Pendekatan ini melibatkan tiga langkah utama: penanganan *missing value*, pelatihan algoritma, dan evaluasi performa.



Gambar 1. Desain Penelitian

Gambar 1 menunjukkan alur metode penelitian, dimulai dari pengumpulan dataset mahasiswa, penanganan *missing value*, pembagian data, pelatihan model menggunakan algoritma *Naïve Bayes* dan *Decision Tree*, hingga evaluasi serta perbandingan hasil untuk menarik kesimpulan

Dataset

Data yang digunakan dalam penelitian ini berasal dari dataset kelulusan mahasiswa yang mencakup data historis akademik seperti nilai rata-rata kumulatif (IPK), kehadiran, dan jumlah kredit yang diselesaikan. Dataset ini akan mengalami manipulasi dengan sengaja menghapus nilai untuk menciptakan missing value dengan persentase tertentu (5%, 10%, 20%, dan 30%) untuk mengevaluasi performa algoritma dalam skenario realistis (Hanyf & Silkan, 2024).

Tabel 1. Dataset Mahasiswa

No.	NIM	Jenis Kelamin	Status Mhs	Umur	Status Nikah	IPS1	IPS2	IPS3	IPS4	IPS5	IPS6	IPS7	IPS8	IPK	Status Lulus
1	619001	P	Eks	28	Belum	2,76	2,8	3,2	3,17	2,98	3	3,03	0	3,07	Terlambat
2	619002	P	Reg	32	Belum	3	3,3	3,14	3,14	2,84	3,13	3,25	0	3,17	Terlambat
3	619003	P	Eks	29	Belum	3,5	3,3	3,7	3,29	3,53	3,72	3,73	0	3,54	Terlambat
4	619004	P	Reg	27	Belum	3,17	3,41	3,61	3,36	3,48	3,63	3,46	0	3,41	Terlambat
5	619005	P	Eks	29	Belum	2,9	2,89	3,3	2,85	2,98	3	3,08	0	3,09	Terlambat
6	619006	LK	Eks	27	Belum	2,95	2,82	3,09	3,1	2,78	3,16	3,23	0	3,07	Terlambat
7	619007	P	Reg	27	Belum	2,76	3,14	2,6	2,95	3,23	3,33	3,3	3,3	3,06	Tepat
8	619008	P	Reg	26	Belum	2,62	2,89	2,32	2,5	2,5	2,86	3,05	2,5	2,91	Tepat

e-ISSN: 2964-2663
p-ISSN: 2964-9692

9	619009	P	Reg	25	Nikah	3,6	3,54	3,52	3,39	3,52	3,68	3,15	0	3,4	Terlambat
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
377	620149	P	Reg	23	Belum	2,74	2,75	2,55	3	2,98	2,8	3,14	3	3,03	Tepat
378	620150	P	Reg	23	Belum	3,02	2,94	3,25	2,87	3	2,94	3,09	3	3,16	Tepat
379	620151	LK	Reg	23	Belum	3,1	3,06	3	3,23	2,79	3	2,41	3	2,16	Tepat

Tabel 1, menunjukkan data dikumpulkan dengan menggabungkan seluruh dataset menjadi satu dengan atribut yang seragam. Data kelulusan mahasiswa terdiri dari 379 siswa yang lulus baik tepat waktu maupun terlambat, berasal dari jurusan IPS dan mencakup kelas IPS 1 hingga kelas IPS 8. Dalam dataset siswa, terdapat 14 atribut.

Tabel 1 dapat dijelaskan Dimana NIM adalah Nomor Induk Mahasiswa, IPS1-IPS8 adalah Indeks Prestasi Semester 1 hingga 8, IPK adalah Indeks Prestasi Komulatif, Eks adalah kepanjangan dari Ekstensi dimana mahasiswa kuliah sambil kerja, Reg kepanjangan dari Reguler yaitu mahasiswa yang hanya kuliah.

Penanganan *Missing Value*

Penanganan *missing value* dilakukan menggunakan beberapa metode, yaitu *Mean Imputation,* dimana metode ini menggantikan nilai yang hilang dengan rata-rata dari nilai-nilai lainnya dalam kolom yang sama. Rumusnya adalah:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

Dimana $\bar{x}$ adalah rata-rata, x_i adalah elemen data, dan n adalah jumlah elemen data yang tersedia. Meskipun sederhana, metode ini dapat memperkenalkan bias karena mengabaikan hubungan antar atribut (Ahmad et al., 2024).

K-Nearest *Neighbors (KNN) Imputation.* Metode ini menggantikan nilai yang hilang dengan rata-rata atau median dari nilai-nilai tetangga terdekat. Tetangga ditentukan berdasarkan jarak seperti jarak *Euclidean*:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2)$$

KNN mempertimbangkan hubungan antar atribut, sehingga lebih akurat dibandingkan imputasi rata-rata dalam dataset yang kompleks (Ribeiro & Freitas, 2021).

Iterative Imputation (MissForest). Metode ini menggunakan algoritma seperti *Random Forest* untuk memperkirakan nilai yang hilang secara iteratif. Setiap iterasi memperkirakan nilai hilang dalam satu fitur berdasarkan fitur lainnya. Rumus dasar prediksi dengan *Random Forest*:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x) \quad (3)$$

Dimana T adalah jumlah pohon keputusan, $f_t(x)$ adalah prediksi dari pohon t. Metode ini dikenal sangat akurat untuk dataset dengan banyak fitur dan nilai yang hilang secara acak (Hanyf & Silkan, 2024).

Algoritma *Machine Learning*

Dua algoritma yang digunakan dalam penelitian ini.

Algoritma *Naïve Bayes* adalah pendekatan probabilistik berbasis Teorema *Bayes* yang mengasumsikan bahwa fitur-fitur dalam dataset saling independen. Rumus dasar Teorema *Bayes* adalah sebagai berikut:

$$P\left(\frac{C}{X}\right) = \frac{P(C) * P(C)}{P(X)} \quad (4)$$

Dimana $P(\frac{C}{X})$ adalah probabilitas suatu kelas C diberi fitur X. $P(\frac{X}{C})$ adalah probabilitas fitur X diberikan kelas C. $P(C)$ adalah probabilitas awal dari kelas C dan $P(X)$ adalah probabilitas awal dari fitur X.

Setiap fitur dianggap memberikan kontribusi independen terhadap probabilitas akhir, yang dikenal sebagai asumsi independensi bersyarat. *Naïve Bayes* memiliki kelebihan dalam menangani dataset besar dan bekerja baik pada fitur diskrit maupun kontinu (Garba et al., 2024).

e-ISSN: 2964-2663
p-ISSN: 2964-9692

Algoritma *Decision Tree* membangun pohon keputusan berdasarkan pemisahan data secara rekursif menggunakan metrik seperti *Information Gain* atau *Gini Index*. Proses pemisahan ini dinyatakan dengan rumus *Entropy*:

$$H(S) = -\sum_{i=1}^{n} p_i * \log_2(p_i) \quad (5)$$

Dimana H(S) adalah entropi dari himpunan data S, sedangkan p_i adalah proporsi data dalam kelas i.

Untuk memilih atribut terbaik pada setiap pemisahan, algoritma menghitung *Information Gain*, Adapun rumus yang digunakan adalah:

$$IG(S,A) = H(S) - \sum_{v \epsilon V} \frac{|S_v|}{|S|} H(S_v) \quad (6)$$

Dimana $IG(S,A)$ adalah keuntungan informasi dari atribut A. V adalah himpunan nilai yang mungkin dari atribut A, sedangkan S_v adalah subset data untuk nilai v.

Evaluasi Model

Kinerja model dievaluasi berdasarkan metrik akurasi, presisi, *recall*, dan F1-score (Ahmad et al., 2024). Evaluasi kedua algoritma dilakukan menggunakan metrik seperti akurasi, presisi, *recall*, dan F1-Score:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (7)$$

Dengan hasil komparasi menunjukkan bahwa algoritma *Decision Tree* seringkali lebih akurat dibandingkan *Naïve Bayes*, terutama dalam konteks dataset pendidikan yang kompleks (Aldossari et al., 2020).

Kinerja Model

Hasil evaluasi dari kedua algoritma akan dibandingkan menggunakan analisis statistik, seperti uji-t berpasangan untuk melihat signifikansi perbedaan performa antar algoritma pada skenario yang berbeda (Chen & McCoy, 2024).

III. HASIL DAN PEMBAHASAN

Berdasarkan data penelitian yang dilakukan, hasil dapat disajikan dalam tabel dan grafik untuk menunjukkan perbandingan efektivitas algoritma Naïve Bayes dan Decision Tree pada prediksi kelulusan mahasiswa setelah penanganan missing value. Berikut adalah rincian hasil dan pembahasannya.

Pada tabel 2 disajikan hasil evaluasi performa algoritma Naïve Bayes dan Decision Tree berdasarkan metrik akurasi, presisi, recall, dan F1-Score.

Tabel 2. Hasil Evaluasi Algoritma Naïve Bayes dan Decision Tree

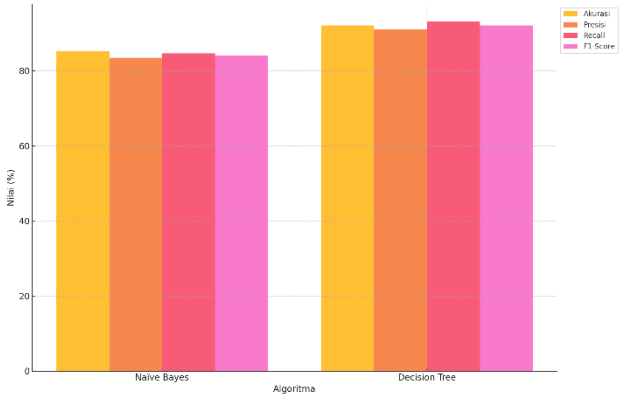
Algoritma	Akurasi	Presisi	Recal	F1-Score
Naïve Bayes	85,3%	83,5%	84,7%	84,1%
Decision Tree	92,1%	91,0%	93,2%	92,1%

Pada tabel 2 terlihat bahwa algoritma Decision Tree memiliki kinerja yang lebih tinggi dibandingkan Naïve Bayes dalam semua metrik. Keunggulan Decision Tree ini menunjukkan bahwa algoritma ini mampu menangani dataset pendidikan dengan hubungan antar atribut yang kompleks.

Tabel 2 menunjukkan hasil evaluasi algoritma Naïve Bayes dan Decision Tree berdasarkan metrik akurasi, presisi, recall, dan F1-Score. Decision Tree memiliki performa lebih tinggi dibandingkan Naïve Bayes pada semua metrik, dengan akurasi mencapai 92,1% dibandingkan 85,3%. Hal ini menunjukkan bahwa Decision Tree lebih cocok untuk dataset pendidikan yang kompleks dan memiliki variabilitas tinggi.

Untuk memberikan visualisasi yang lebih jelas, berikut adalah grafik perbandingan kinerja algoritma berdasarkan hasil evaluasi.

e-ISSN: 2964-2663
p-ISSN: 2964-9692



Gambar 2. Grafik Perbandingan Kinerja Algoritma Naive Bayes dan Decision Tree

Pada Gambar 2 menunjukkan perbandingan kinerja algoritma Naïve Bayes dan Decision Tree dalam empat metrik evaluasi: akurasi, presisi, recall, dan F1-Score. Hasil menunjukkan bahwa algoritma Decision Tree unggul dalam semua metrik, dengan performa tertinggi pada recall (93,2%) dan F1-Score (92,1%). Hal ini menunjukkan bahwa Decision Tree lebih handal dalam menangkap pola data, terutama pada dataset dengan hubungan antar fitur yang kompleks.

Gambar 2 menunjukkan performa kedua algoritma secara visual. Decision Tree unggul signifikan dalam semua metrik dibandingkan Naïve Bayes. Hal ini dapat dikaitkan dengan kemampuan Decision Tree untuk menangani fitur kontinu dan diskrit secara efisien, serta kemampuannya untuk memodelkan hubungan yang lebih kompleks antar fitur.

Hasil ini menunjukkan bahwa algoritma Decision Tree secara konsisten unggul dalam setiap aspek evaluasi dibandingkan dengan Naïve Bayes. Hal ini kemungkinan karena kemampuan Decision Tree untuk menangani atribut yang memiliki hubungan kompleks, sementara Naïve Bayes memiliki keterbatasan asumsi independensi antar fitur. Analisis ini mendukung bahwa Decision Tree lebih efektif untuk dataset pendidikan dengan karakteristik serupa.

IV. KESIMPULAN

Penelitian ini berhasil mencapai tujuan yang dinyatakan dalam pendahuluan, yaitu mengidentifikasi metode penanganan missing value yang efektif dan membandingkan performa algoritma Naïve Bayes serta Decision Tree dalam memprediksi kelulusan mahasiswa. Hasil penelitian menunjukkan bahwa Decision Tree lebih unggul dalam menangkap hubungan kompleks antar atribut, sehingga menghasilkan prediksi yang lebih akurat dibandingkan Naïve Bayes. Penanganan missing value dengan metode seperti KNN Imputation dan Iterative Imputation terbukti memberikan kontribusi signifikan terhadap peningkatan kinerja model. Prospek pengembangan penelitian ini meliputi pengujian model pada dataset yang lebih beragam, eksplorasi algoritma pembelajaran mesin lain seperti Random Forest atau Gradient Boosting, serta optimasi hyperparameter untuk meningkatkan kinerja lebih lanjut. Selain itu, hasil penelitian ini dapat diterapkan dalam pengembangan sistem prediksi kelulusan mahasiswa berbasis data untuk mendukung kebijakan akademik dan intervensi dini terhadap mahasiswa yang berisiko tidak lulus tepat waktu. Penelitian lanjutan disarankan untuk menggunakan data real-time, mengintegrasikan faktor sosial-ekonomi, dan melakukan validasi model secara lintas institusi untuk memastikan generalisasi hasil yang lebih luas..

V. REFERENSI

Ahmad, A. F., Sayeed, M. S., Alshammari, K., & Ahmed, I. (2024). *Impact of Missing Values in Machine Learning: A Comprehensive Analysis*. https://doi.org/10.48550/arXiv.2410.08295

Alabadla, M., Sidi, F., Ishak, I., Ibrahim, H., Affendey, L. S., Ani, Z. C., Jabar, M. A., Bukar, U. A., Devaraj, N. K., Muda, A. S., Tharek, A., Omar, N., & Jaya, M. I. M. (2022). Systematic Review of Using Machine Learning in Imputing Missing Values. *IEEE Access*, *10*, 44483–44502. https://doi.org/10.1109/ACCESS.2022.3160841

Aldossari, B. S., Alqahtani, F. M., Alshahrani, N. S., Alhammam, M. M., Alzamanan, R. M., Aslam, N., & Irfanullah. (2020). A Comparative Study of Decision Tree and Naive Bayes Machine Learning Model for Crime Category Prediction in Chicago. *Proceedings of 2020 6th International Conference on Computing and Data Engineering*, 34–38. https://doi.org/10.1145/3379247.3379279

e-ISSN: 2964-2663
p-ISSN: 2964-9692

Chen, A. Y., & McCoy, J. (2024). Missing values handling for machine learning portfolios. *Journal of Financial Economics*, *155*, 103815. https://doi.org/10.1016/j.jfineco.2024.103815

Dharmasaputro, A. A., Fauzan, N. M., Kallista, M., Wibawa, I. P. D., & Kusuma, P. D. (2022). Handling Missing and Imbalanced Data to Improve Generalization Performance of Machine Learning Classifier. *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, 140–145. https://doi.org/10.1109/ISMODE53584.2022.9743022

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*, 1–37. https://doi.org/10.1186/s40537-021-00516-9

Garba, M., Usman, M. A., & Gulumbe, A. M. (2024). Improving breast cancer detection with naive bayes: a predictive analytics approach. *CS & IT Conference Proceedings*, *14*(11).

Gond, V. K., Dubey, A., & Rasool, A. (2021). A Survey of Machine Learning-Based Approaches for Missing Value Imputation. *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 1–8. https://doi.org/10.1109/ICIRCA51532.2021.9544957

Hanyf, Y., & Silkan, H. (2024). A method for missing values imputation of machine learning datasets. *Int J Artif Intell ISSN*, *2252*(8938), 8938. https://doi.org/doi.org/10.11591/ijai.v13.i1.pp888-898

Palanivinayagam, A., & Damaševičius, R. (2023). Effective Handling of Missing Values in Datasets for Classification Using Machine Learning Methods. In *Information* (Vol. 14, Issue 2). https://doi.org/10.3390/info14020092

Ribeiro, C., & Freitas, A. A. (2021). A data-driven missing value imputation approach for longitudinal datasets. *Artificial Intelligence Review*, *54*(8), 6277–6307. https://doi.org/10.1007/s10462-021-09963-5