International Journal of Applied Sciences and Smart Technologies

Volume 5, Issue 2, pages 1–12

p-ISSN 2655-8564, e-ISSN 2685-9432

IJASST

Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbors terhadap Klasifikasi Data Pendaftar Beasiswa

Resiana Kinanti Jati[1, *], Loadtriana Oktavia S[2]

[1,2] *Program Studi Informatika, Fakultas Sains dan Teknologi, Universitas Sanata Dharma, Yogyakarta, Indonesia*

[*]*Corresponding Author: 215314159@student.usd.ac.id*

(Received DD-MM-YYYY; Revised DD-MM-YYYY; Accepted DD-MM-YYYY)

Abstrak

Penelitian ini membandingkan kinerja algoritma Naive Bayes dan K-Nearest Neighbors (KNN) dalam mengklasifikasikan pendaftar beasiswa di Universitas Merdeka. Data yang digunakan terdiri dari 1107 data dengan berbagai atribut seperti latar belakang akademis, kondisi keuangan, dan status keluarga. Data tersebut dipreproses untuk menangani nilai yang hilang, menghapus atribut yang tidak perlu, menangani outlier dan menormalisasikan data. Algoritma dievaluasi dengan menggunakan berbagai metrik seperti akurasi, presisi, recall, dan F1-score. Hasilnya menunjukkan bahwa Naive Bayes mengungguli KNN dalam hal akurasi, presisi, recall, dan F1-score, dengan akurasi rata-rata 73,36% dibandingkan dengan 69,18% dari KNN. Penelitian ini menyoroti pentingnya preprocessing data dan efektivitas Naive Bayes dalam mengklasifikasikan pelamar beasiswa.

Kata kunci: Beasiswa, Klasifikasi, K-Nearest Neigbor, Naïve Bayes, Penambangan Data.

Abstract

This study compares the performance of Naive Bayes and K-Nearest Neighbors (KNN) algorithms in classifying scholarship applicants at Universitas Merdeka. The data used consisted of 1107 data with various attributes such as academic background, financial condition, and family status. The data was preprocessed to handle missing values, remove unnecessary attributes, handling outliers and normalize the data. The algorithms were evaluated using various metrics such as accuracy, precision, recall, and F1-score. The results show that Naive Bayes outperformed KNN in terms of accuracy, precision, recall, and F1-score, with an average accuracy of 73.36% compared to KNN's 69.18%. The study highlights the importance of data preprocessing and the effectiveness of Naive Bayes in classifying scholarship applicants.

Keywords: Classification, Data Mining, K-Nearest Neigbor, Naïve Bayes, Scholarship.

1 PENDAHULUAN

Beasiswa merupakan bentuk penghargaan yang diberikan kepada seseorang untuk membantu dalam melanjutkan pendidikan mereka ke jenjang yang lebih tinggi. Penghargaan tersebut dapat berupa akses tertentu pada suatu institusi atau bantuan

Volume 5, Issue 2, pages 1–12
p-ISSN 2655-8564, e-ISSN 2685-9432
IJASST

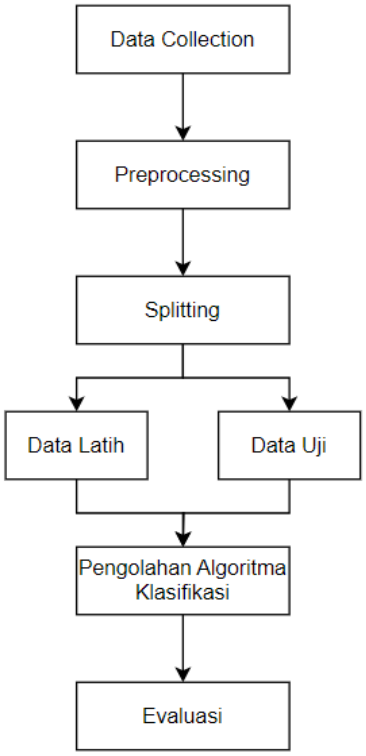
dalam bentuk keuangan [1]. Beasiswa dapat diberikan oleh lembaga pemerintah, perusahaan ataupun yayasan. Pemberian beasiswa dikategorikan sesuai dengan tujuan diberikan beasiswa tersebut, syarat sebagai penerima beasiswa, dan jenis-jenis beasiswa.

Universitas Merdeka membuka program beasiswa bagi mahasiswa berprestasi sebagai bentuk penghargaan dan dukungan dana pendidikan sebagaimana turut disertakan syarat beasiswa yaitu kondisi keuangan seperti kepemilikan surat keterangan tidak mampu, tunggakan dana studi, dan keadaan orang tua. Seleksi penerima beasiswa harus dilakukan dengan cermat agar bantuan tepat sasaran dan adil. Dalam proses seleksi ini, klasifikasi data pendaftar menjadi salah satu langkah yang dapat digunakan untuk menentukan siapa yang berhak menerima beasiswa.

2 METODE PENELITIAN

Dalam penelitian ini dilakukan dengan beberapa tahapan. Setelah data didapatkan, peneliti merancang langkah-langkah yang bertujuan untuk mendapatkan hasil perbandingan yang maksimal antara dua algoritma terhadap data pendaftar beasiswa mahasiswa Universitas Merdeka. Langkah-langkah tersebut tergambar dalam gambar 1.

Volume 5, Issue 2, pages 1–12
p-ISSN 2655-8564, e-ISSN 2685-9432
IJASST



Gambar 1. Metode Penelitian

2.1 Data Collection

Data collection merupakan suatu langkah untuk mengumpulkan atau memperoleh data [2]. Pada penelitian ini, Kumpulan data atau *dataset* didapatkan dari dosen pengampu yang mana *dataset* dibagikan melalui laman belajar mahasiswa. *Dataset* yang digunakan dalam penelitian ini adalah data pendaftar beasiswa Universitas Merdeka.

2.2 Data Preprocessing

Data preprocessing merupakan langkah untuk mempersiapkan data sebelum digunakan [2] seperti mengatasi data dengan nilai kosong atau *null* (*missing value*), menghapus atribut yang tidak diperlukan, pengecekan duplikasi data, mengatasi nilai *outliers*, dan normalisasi data.

Volume 5, Issue 2, pages 1–12
p-ISSN 2655-8564, e-ISSN 2685-9432
IJASST

2.3 Splitting Data

Splitting atau pembagian data merupakan langkah untuk membagi data menjadi dua bagian yaitu data latih (*training*) dan data uji (*testing*). Data latih digunakan untuk melatih algoritma dan data uji digunakan untuk mengukur evaluasi kinerja pada algoritma.

2.4 Naive Bayes

Naive Bayes adalah metode klasifikasi statistik yang mengikuti prinsip probabilitas yang diperkenalkan oleh Thomas Bayes. Istilah "naive" digunakan karena asumsi bahwa setiap atribut dalam dataset adalah independen satu sama lain [3].

$$P(C|X) = \frac{P(X|C).P(C)}{P(X)} \quad (1)$$

dimana :

X : Sampel daya yang memiliki class yang tidak diketahui

C : Hipotesis bahwa X adalah data class

P(C) : Probabilitas hipotesis C

P(X) : Peluang dari data sampel yang diamati (probabilitas C)

P$(X|C)$: Probabilitas berdasarkan kondisi pada hipotesis

2.5 K-Nearest Neighbor

Algoritma K-Nearest Neighbor (KNN) adalah metode yang melakukan klasifikasi pada objek dengan menggunakan data latih yang memiliki jarak terdekat dengan objek yang akan diklasifikasikan. Ini merupakan salah satu metode yang digunakan untuk analisis klasifikasi dan juga untuk prediksi [4]. Jarak antara dua titik pada data training dan titik pada data testing dapat didefinisikan dengan rumus Euclidean.

$$d = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2} \quad (2)$$

dimana :

d : jarak Euclidean

x2i : nilai pada data testing ke -i

Volume 5, Issue 2, pages 1–12
p-ISSN 2655-8564, e-ISSN 2685-9432
IJASST

x1i : nilai pada data training ke -i

p : banyaknya atribut

2.6 Evaluasi

Evaluasi merupakan langkah yang dilakukan untuk mengukur performa algoritma dengan mendapatkan nilai akurasi, ditampilkan melalui *Confusion Matrix*. *Confusion Matrix* adalah alat yang mengevaluasi keakuratan model klasifikasi dengan membedakan prediksi yang benar dan salah. *Confusion matrix* adalah tabel yang menunjukkan berapa banyak data yang diklasifikasikan dengan benar dan berapa banyak yang salah [5].

Tabel 1. Confusion Matrix

Confusion Matrix		Prediksi	
		Positif	Negatif
Aktual	Positif	TP	FN
	Negatif	FP	TN

dimana :

True Positif (TP) : Jumlah prediksi yang benar dari data yang relevant.

False Positive (FP) : Jumlah prediksi yang salah dari data yang tidak relevant.

False Negative (FN) : Jumlah prediksi yang salah dari data yang tidak relevant.

True Negative (TN) : Jumlah prediksi yang benar dari data yang relevant.

Selain itu, dalam penilaian model, menghitung akurasi, presisi, recall, dan skor f-1 score dengan rumus-rumus seperti berikut [6]:

Accuracy atau akurasi adalah cara untuk melihat seberapa baik sistem kita dalam menanggapi data.

$$acc = \frac{TN+TP}{FN+FP+TN+TP} \quad (3)$$

Precision dan presisi adalah ukuran untuk menilai seberapa tepat suatu algoritma.

$$pre = \frac{TP}{FP+TP} \quad (4)$$

International Journal of Applied Sciences and Smart Technologies
Volume 5, Issue 2, pages 1–12
p-ISSN 2655-8564, e-ISSN 2685-9432
IJASST

Recall adalah seberapa baik sistem dalam mengingat informasi yang sebelumnya disampaikan.

$$recall = \frac{TP}{FN+TP} \quad (5)$$

F-Measure adalah nilai rata-rata yang menggabungkan *precision* dan *recall*.

$$F_1 = 2\ x \frac{precision\ x}{precision+recall} \quad (6)$$

3 HASIL DAN ANALISIS

3.1 Dataset

Data yang digunakan berjumlah 1107. Atribut-atribut data yang digunakan antara lain: Angkatan, Semester, IPK, Surat Keterangan Tidak Mampu, Penghasilan bapak, Penghasilan ibu, Tagihan Listrik, Jumlah Tanggungan Masih Studi, Tunggakan, Jumlah adik, Jumlah kakak, dan Jumlah Point.

No.	Periode	Angkatan	Kabupaten	Propinsi	Semester	IPK	Srt Ket. Tidak Mampu	Penghasilan Bapak	Penghasilan Ibu	Tgh. Listrik	Jml. Tanggungan masih studi	Tunggakan	Keadaan Orang Tua	Jumlah Adik	Jumlah Kakak	Jml. Point	Status beasiswa
1	1012019	2017	NaN	NaN	3	3.87	NaN	0.0	1259300.0	75000.0	2.0	0.0	2	0.0	2.0	0.0	Y
2	1012019	2017	Magelang	Jawa Tengah	3	2.79	ada	0.0	1128100.0	298000.0	1.0	0.0	2	0.0	0.0	0.0	T
3	1012019	2017	NaN	NaN	3	3.88	NaN	2000000.0	0.0	118180.0	0.0	0.0	3	0.0	1.0	0.0	T
4	1012019	2017	NaN	NaN	3	3.95	NaN	2374288.0	5750000.0	433469.0	2.0	0.0	3	1.0	0.0	0.0	T
5	1012019	2015	Sleman	Daerah Istimewa YogYakarta	7	3.35	NaN	0.0	2921800.0	100000.0	2.0	0.0	11	1.0	0.0	0.0	Y
6	1012019	2017	NaN	NaN	3	3.31	ada	1500000.0	0.0	132953.0	1.0	1900000.0	3	0.0	5.0	0.0	T
7	1012019	2015	YogYakarta	Daerah Istimewa YogYakarta	7	3.49	ada	2464852.0	0.0	300000.0	2.0	2235000.0	9	1.0	0.0	0.0	Y
8	1012019	2017	NaN	NaN	3	3.62	ada	1500000.0	0.0	50641.0	1.0	2135000.0	9	1.0	0.0	0.0	Y
9	1012019	2015	Pemalang	Jawa Tengah	7	3.48	NaN	5094300.0	0.0	34000.0	1.0	0.0	3	0.0	0.0	0.0	T
10	1012019	2016	NaN	NaN	5	3.12	NaN	0.0	629700.0	259466.0	1.0	2890000.0	8	0.0	1.0	0.0	Y
11	1012019	2015	Ketapang	Kalimantan Barat	7	3.40	NaN	1200000.0	3293000.0	72823.0	3.0	0.0	3	2.0	1.0	2.0	T

Gambar 2. Tabel Dataset Mentah

3.2 Data Preprocessing

Langkah pertama setelah mendapatkan data adalah mempersiapkan data. Gambar 2 menunjukkan tampilan data mentah, di mana masih terdapat nilai kosong atau *null* (*missing value*) dan atribut yang tidak diperlukan. Selain itu juga diperlukan pengecekan terhadap duplikasi data, nilai *outliers*, dan normalisasi data. Hasil dari langkah preprocessing ditunjukkan pada gambar 3.

International Journal of Applied Sciences and Smart Technologies
Volume 5, Issue 2, pages 1–12
p-ISSN 2655-8564, e-ISSN 2685-9432
IJASST

	Angkatan	Semester	IPK	Srt Ket. Tidak Mampu	Penghasilan Bapak	Penghasilan Ibu	Tgh. Listrik	Jml. Tanggungan masih studi	Tunggakan	Keadaan Orang Tua	Jumlah Adik	Jumlah Kakak	Jml. Point
0	0.333333	0.0	0.900000	0.0	0.000000	0.100864	0.002644	1.333333e-07	0.000000	0.1	0.000000	0.285714	0.000000
1	0.333333	0.0	0.069231	1.0	0.000000	0.090355	0.010507	6.666667e-08	0.000000	0.1	0.000000	0.000000	0.000000
2	0.333333	0.0	0.907692	0.0	0.168705	0.000000	0.004167	0.000000e+00	0.000000	0.2	0.000000	0.142857	0.000000
3	0.333333	0.0	0.961538	0.0	0.200277	0.460548	0.015283	1.333333e-07	0.000000	0.2	0.166667	0.000000	0.000000
4	0.000000	1.0	0.500000	0.0	0.000000	0.234022	0.003526	1.333333e-07	0.000000	1.0	0.166667	0.000000	0.000000
5	0.333333	0.0	0.469231	1.0	0.126529	0.000000	0.004688	6.666667e-08	0.033015	0.2	0.000000	0.714286	0.000000
6	0.000000	1.0	0.607692	1.0	0.207916	0.000000	0.010578	1.333333e-07	0.038836	0.8	0.166667	0.000000	0.000000
7	0.333333	0.0	0.707692	1.0	0.126529	0.000000	0.001786	6.666667e-08	0.037098	0.8	0.166667	0.000000	0.000000
8	0.000000	1.0	0.600000	0.0	0.429716	0.000000	0.001199	6.666667e-08	0.000000	0.2	0.000000	0.000000	0.000000
9	0.166667	0.5	0.323077	0.0	0.000000	0.050436	0.009148	6.666667e-08	0.050217	0.7	0.000000	0.142857	0.000000
10	0.000000	1.0	0.538462	0.0	0.101223	0.263754	0.002568	2.000000e-07	0.000000	0.2	0.333333	0.142857	0.333333

Gambar 3. Hasil Preprocessing Data

3.3 Hasil Pengolahan Data

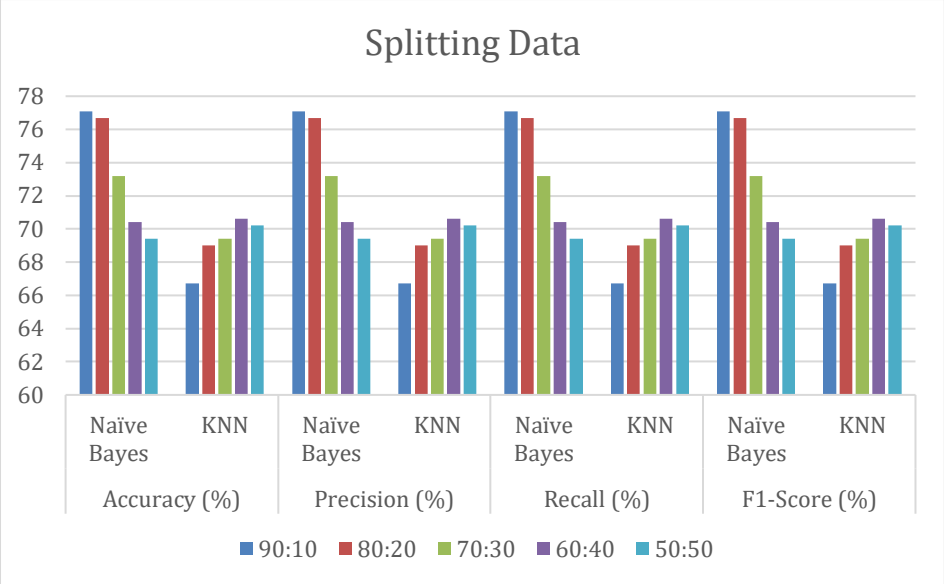
Pada proses pengolahan kedua algoritma, dilakukan beberapa kali uji coba pergantian besar rasio untuk *splitting data*. Hal ini bertujuan untuk mengetahui pembanding cara kerja algoritma Naïve Bayes dalam hal ini *Gaussian* dan KNN [7]. Besar rasio antara lain 90:10, 80:20, 70:30, 60:40, dan 50:50. Pada algoritma KNN, parameter *n_neighbor* diatur sebanyak 3, sedangkan Naïve Bayes tidak memerlukan pengaturan parameter.

Tabel 2. Hasil Splitting Data

Data Training (%)	Data Testing (%)	Accuracy (%)		Precision (%)		Recall (%)		F1-Score (%)	
		Naïve Bayes	KNN	Naïve Bayes	KNN	Naïve Bayes	KNN	Naïve Bayes	KNN
90	10	77,1	66,7	77,1	66,7	77,1	66,7	77,1	66,7
80	20	76,7	69	76,7	69	76,7	69	76,7	69
70	30	73,2	69,4	73,2	69,4	73,2	69,4	73,2	69,4
60	40	70,4	70,6	70,4	70,6	70,4	70,6	70,4	70,6
50	50	69,4	70,2	69,4	70,2	69,4	70,2	69,4	70,2
Rata-rata		73,36	69,18	73,36	69,18	73,36	69,18	73,36	69,18

Berdasarkan tabel 2, nilai akurasi, *precision*, *recall,* dan *f1-score* kedua algoritma berkisar antara 60% - 70%. Algoritma Naïve Bayes memiliki rata-rata akurasi lebih tinggi (73,36%) dibandingkan dengan KNN (69,18%). Naïve Bayes juga unggul dalam nilai rata-rata *precision*, *recall*, dan *f1 score* dibandingkan dengan KNN.

International Journal of Applied Sciences and Smart Technologies
Volume 5, Issue 2, pages 1–12
p-ISSN 2655-8564, e-ISSN 2685-9432
IJASST



Gambar 4. Komparasi algoritma Naïve Bayes dan KNN

4 KESIMPULAN

Berdasarkan hasil pengolahan data yang dilakukan untuk melakukan perbandingan algoritma Naïve Bayes (*Gaussian*) dan KNN dalam klasifikasi data pendaftar beasiswa mahasiswa Universitas Merdeka dengan menggunakan data sebanyak 1107 data, disimpulkan bahwa kinerja dari algoritma Naïve Bayes mempunyai kinerja yang lebih baik dibanding kinerja algoritma KNN. Untuk penelitian selanjutnya, dapat dilakukan eksplorasi *hyperparameter tuning* untuk lebih mengetahui perbandingan kinerja algoritma Naïve Bayes dan KNN [8].

REFERENSI

[1] E. Murniasih, *Buku Pintar Beasiswa: Panduan Komplet Meraih Beasiswa di Dalam maupun Luar Negeri*. Yogyakarta: Gagas Media, 2009.

[2] B. Delvika, S. Nurhidayarnis, P. D. Rinada, N. Abror, and A. Hidayat, "Comparison of Classification Between Naive Bayes and K-Nearest Neighbor on Diabetes Risk in Pregnant Women Perbandingan Klasifikasi Antara Naive Bayes dan K-Nearest Neighbor Terhadap Resiko Diabetes Pada Ibu Hamil," *MALCOM: Indonesian Journal of Machine Learning and Computer Science* , vol. 2, no. 2, pp. 68–75, 2022.

[3] A. Rachmat and Y. Lukito, "Klasifikasi Sentimen Komentar Politik dari Facebook Page Menggunakan Naive Bayes," *26 JUISI*, vol. 02, no. 02, 2016, [Online]. Available: https://github.com/sastrawi/sastrawi

Volume 5, Issue 2, pages 1–12
p-ISSN 2655-8564, e-ISSN 2685-9432
IJASST

[4] K. Alkhatib *et al.*, "Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm," 2013. [Online]. Available: https://www.researchgate.net/publication/262456253

[5] M. Christianto, J. Andjarwirawan, and A. Tjondrowiguno, "Aplikasi Analisa Sentimen Pada Komentar Berbahasa Indonesia Dalam Objek Video di Website YouTube Menggunakan Metode Naïve Bayes Classifier," 2020.

[6] I. Iwandini, A. Triayudi, and G. Soepriyono, "Analisa Sentimen Pengguna Transportasi Jakarta Terhadap Transjakarta Menggunakan Metode Naives Bayes dan K-Nearest Neighbor," *Journal of Information System Research (JOSH)*, vol. 4, no. 2, pp. 543–550, Jan. 2023, doi: 10.47065/josh.v4i2.2937.

[7] A. Putri, C. Syaficha Hardiana, E. Novfuja, F. Try Puspa Siregar, Y. Fatma, and R. Wahyuni, "Comparison of K-NN, Naive Bayes and SVM Algorithms for Final-Year Student Graduation Prediction Komparasi Algoritma K-NN, Naive Bayes dan SVM untuk Prediksi Kelulusan Mahasiswa Tingkat Akhir," *Institut Riset dan Publikasi Indonesia (IRPI) MALCOM: Indonesian Journal of Machine Learning and Computer Science Journal Homepage*, vol. 3, no. 1, pp. 20–26, 2023.

[8] W. Nugraha and A. Sasongko, "Hyperparameter Tuning pada Algoritma Klasifikasi dengan Grid Search Hyperparameter Tuning on Classification Algorithm with Grid Search," *SISTEMASI: Jurnal Sistem Informasi* , vol. 11, no. 2, pp. 391–401, 2022, [Online]. Available: http://sistemasi.ftik.unisi.ac.id