JURITEK: Jurnal Ilmiah Teknik Mesin, Elektro dan Komputer
Vol.4, No.1 Maret 2024
E-ISSN: 2809-0799; P- ISSN: 2809-0802, Hal 44-52
DOI: https://doi.org/10.51903/juritek.v4i1.2886

Metode Data Mining untuk Seleksi Calon Mahasiswa Baru pada Penerimaan Mahasiswa Baru di Universitas Muhammadiyah Gresik

Muhammad Zaky Al Mubarok
Teknik Informatika, Universitas Muhammadiyah Gresik
Korespondensi penulis: zaky.space@gmail.com

Umi Chotijah
Teknik Informatika, Universitas Muhammadiyah Gresik

Alamat: Jl. Sumatra No. 101 GKB Gresik 61121

Abstract. *Muhammadiyah University of Gresik is one of the educational institutions in Gresik. Every year, the ratio of new students to graduates is not the same, which can affect the accreditation of the campus. To address this issue, a prediction is made on the data of prospective new students to detect whether they can graduate on time or not. A comparison of the classification results is performed using the K-Nearest Neighbor and Naive Bayes methods. From the implementation and testing, Naive Bayes achieves an accuracy of 72%, while the K-Nearest Neighbor method achieves an accuracy of 64%. Therefore, Naive Bayes is better at classifying the data of prospective new students compared to K-Nearest Neighbor.*

Keywords: *Classification, Thesis Completion Time, Weighted Naïve Bayes, K-Nearest Neighbor, Data Mining.*

Abstrak. Universitas Muhammadiyah Gresik adalah salah satu instansi pendidikan yang ada di Gresik. Setiap tahunya rasio mahasiswa baru dan mahasiswa yang lulus tidak sama, sehingga hal tersebut dapat mempengaruhi akreditasi kampus. Untuk mengatasi hal tersebut di lakukan prediksi terhadap data calon mahasiswa baru untuk mendeteksi apakah calon mahasiswa baru dapat lulus tepat waktu atau tidak. Di lakukan perbandingan hasil klasifikasi dari metode yang di gunakan yaitu *K- Nearest Neighbor* dan *Naive Bayes*. Dari hasil implementasi dan pengujian yang di lakukan mendapatkan *Naive Bayes* mendapatkan nilai akurasi 72%, sedangkan dari metode *K-Nearest Neighbor* mendapatkan nilai akurasi 64%, sehingga *Naive Bayes* dapat mengklasifikasikan data calon mahasiswa baru lebih baik di bandingkan dengan *K-Nearest Neighbor*.

Kata kunci: Klasifikasi, Waktu Penyelesaian Skripsi, *Weighted Naïve Bayes*, *K- Nearest Neighbor*, Data Mining.

LATAR BELAKANG

Universitas Muhammadiyah Gresik atau yang biasa di sebut UMG merupakan instansi perguruan tinggi yang berlokasi di Jl. Sumatera No. 101, Gn. Malang, Randuagung, Kec. Kebomas, Kabupaten Gresik. Berdasarkan surat keputusan pimpinan daerah Muhammadiyah kabupaten Gresik majelis Pendidikan pengajaran dan kebudayaan nomor : E.1/017-V/1980 tanggal 25 Mei 1980, berdirilah Universitas Muhammadiyah Gresik yang peresmiannya di lakukan oleh Bupati Kabupaten Gresik Bapak Kolonel Wasiadji, SH yang juga sebagai pelindung. Pada periode tahun 2018 hingga 2022 di dapatkan kesimpulan bahwa jumlah mahasiswa baru di Universitas Muhammadiyah Gresik mengalami kenaikan secara signifikan pada tahun 2018 hingga 2020, namun pada tahun 2021 mengalami penurunan sebagai dampak dari pandemi *Covid 19*. *Trend* kenaikan jumlah mahasiswa baru terjadi pada tahun 2022, yang

Received Januari 13, 2024; Accepted Februari 10, 2024; Published Maret 06, 2024
*Muhammad Zaky Al Mubarok, zaky.space@gmail.com

menunjukkan bahwa Universitas Muhammadiyah Gresik tetap menjadi salah satu kampus yang terbaik di kabupaten Gresik (UNIVERSITAS MUHAMMADIYAH GRESIK 2023).

KAJIAN TEORITIS

1. Penerimaan Mahasiswa Baru

   Penerimaan mahasiswa baru (PMB) adalah salah satu agenda rutin yang sangat penting di perguruan tinggi. PMB merupakan gerbang awal proses bisnis perguruan tinggi, mencari dan menyeleksi calon mahasiswa yang selanjutnya akan dididik untuk menghasilkan sumber daya manusia yang berkualitas sebagai alumni.

2. Data Mining

   Data mining adalah suatu proses pengumpulan informasi dan data yang penting dalam jumlah yang besar atau big data (Sri Widaningsih 2019). Fungsi data mining terbagi menjadi dua, yakni deskriptif dan prediktif. Fungsi deskriptif untuk memahami lebih jauh tentang data yang diamati, dengan melakukan sebuah proses diharap bisa mengetahui perilaku dari sebuah data tersebut. Fungsi predictive ialah bagaimana sebuah proses nantinya akan menemukan pola tertentu dari suatu data, pola – pola tersebut dapat diketahui dari berbagai variabel yang ada pada data (Utomo & Mesran 2020).

3. Sistem Perbandingan Metode

   Sistem perbandingan metode adalah sistem yang dirancang untuk mengetahui hasil atau akurasi dari model suatu metode (Alim 2021). Dimana hasil dari metode terbaik akan diambil untuk membuat keputusan suatu masalah. Metode yang cocok dengan data yang digunakan dalam penelitian ini meliputi *Naïve Bayes* dan *K-Nearest Neighbor*.

4. Metode Naïve Bayes

   Model yang di gunakan untuk sistem perbandingan model yang pertama adalah model naïve bayes, model ini adalah algoritma machine learning yang di gunakan dalam berbagai klasifikasi. Untuk bisa memahami algoritma ini bisa di pahami rumus umum teorema bayes yang menjadi dasar naïve bayes sendiri (Fatmawati 2016). Langkah-langkah menentukan klasifikasi menggunakan metode *Weight Naïve Bayes*: Menghitung nilai probabilitas tiap kelas.

$$P(C_i) = \frac{\sum C_i}{n}$$

Keterangan:

$P(C_i)$: Probabilitas label kelas C_i

$\sum C_i$: Jumlah data dengan label kelas C_i

n : Jumlah total data latih

Menghitung nilai probabilitas tiap fitur.

Rumus:

$$P(x_k|C_i) = \frac{\sum x_k|C_i}{\sum C_i}$$

Keterangan:

$P(x_k|C_i)$: Probabilitas fitur x_k dengan label kelas C_i

$\sum x_k|C_i$: Jumlah data fitur x_k dengan label kelas C_i

$\sum C_i$: Jumlah data dengan label kelas C_i

Menghitung nilai probabilitas tiap kelas pada tiap data.

Rumus:

$$P(C_i|X) = P(C_i)\prod_{k=1}^{n} P(x_k|C_i)^{w_k}$$

Keterangan:

$P(C_i|X)$: Probabilitas kelas C_i pada data X

$P(C_i)$: Probabilitas label kelas C_i

$P(x_k|C_i)$: Probabilitas fitur x_k dengan label kelas C_i

w_k : Bobot atribut

Menghitung mean data numerik

$$\boldsymbol{\mu = \frac{\sum_{i=1}^{n} x_i}{N}}$$

Standar Deviasi

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}$$

Normal distribution

$$f\left(x\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

5. Metode K-Nearest Neighbor

Algoritma K-Nearest Neighbor merupakan metode klasifikasi terhadap sekumpulan data berdasarkan mayoritas, yang bertujuan untuk mengklasifikasikan objek baru berdasarkan atribut dan kategori yang sama dari sampel data training (Permana P, dkk. 2022). Metode pengukuran jarak *Euclidean* juga paling sering digunakan untuk menghitung kesamaan dari dua vektor. Kelebihan dari metode jarak *Euclidean* ini adalah tingkat kemiripan (*similarity*) lebih tinggi dibanding metode yang lain. Rumus menghitung jarak dengan Euclidean :

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

Keterangan :

d (x,y) : jarak *Euclidean*

Xtraining : *data training* ke-*i*

Ytesting : data testing

i : record (baris) ke-*i* dari tabel

n : jumlah *data training*

METODE PENELITIAN

1. Tahapan Penelitian

- Representasi Data

  Data yang digunakan dalam penelitaian ini didapatkan dari biro admisi humas dan biro administrasi akademik, data yang diambil ini adalah data penerimaan mahasiswa baru Universitas Muhammadiyah Gresik tahun 2022 dan data kelulusan dari program studi Teknik informatika, elektro, dan industri, dengan periode gelombang yaitu gelombang 1 sampai 3, dengan jalur tes CBT (*Computer Based Test*). Data yang digunakan dalam penelitian ini berupa data yang berkaitan dengan nilai tes pada saat calon mahasiswa baru melakukan tes CBT. Data ini digunakan untuk melakukan klasifikasi terhadap calon mahasiswa baru dengan jumlah data 125 yang akan dibagi menjadi 2, yaitu data latih dan data uji

- Perhitungan metode *K-Nearest Neighbor* dan *Naïve Bayes*

  Dalam Proses ini dilakukan ini pengklasifikasian untuk memprediksi ketepatan waktu lulus calon mahasiswa baru menggunakan metode naive bayes dan k-Nearest Neighbor.

- Pengujian *Confusion Matrix* pada metode *K-Nearest Neighbor* dan *Naïve Bayes* Pada tahap ini dilakukan perhitungan hasil nilai performa dari setiap metode yang terdiri dari dari nilai akurasi, nilai presisi, dan reca.

HASIL DAN PEMBAHASAN

Representasi Data

Data sebanyak 125 dibagi menjadi dua, yakni data latih dan data uji.

Table 1. Data Latih

No.	Gelombang	Jenis Kelamin	Kategorial Nilai Bhs Indonesia	Kategorial Nilai Bhs Inggris	Kategorial Nilai Matematika	Kategorial Nilai jurusan / mapel pilihan	Kategorial Nilai Cbt	Keterangan
1	Gelombang I	L	90	68	66	88	46	TIDAK TEPAT WAKTU
2	Gelombang II	L	74	66	52	72	47	TIDAK TEPAT WAKTU
3	Gelombang I	L	98	80	78	96	42	TEPAT WAKTU
...	...	...				.....	.....	.....
116	Gelombang II	L	88	73	68	83	53	TIDAK TEPAT WAKTU

Table 2. Data Uji

No.	Gelombang	Jenis Kelamin	Kategorial Nilai Bhs Indonesia	Kategorial Nilai Bhs Inggris	Kategorial Nilai Matematika	Kategorial Nilai jurusan / mapel pilihan	Kategorial Nilai Cbt	Keterangan
33	Gelombang II	L	94	84	78	88	32	TEPAT WAKTU
34	Gelombang II	L	97	85	77	89	26	TEPAT WAKTU
35	Gelombang II	L	60	40	22,5	52,5	34	TIDAK TEPAT WAKTU
.....	.....	.....	.....	.....	.....	.....	.....	
125	Gelombang II	L	74	56	46	64	46	TIDAK TEPAT WAKTU

Hasil Perhitungan Metode *K-Nearest Neighbor*

Hasil Perhitungan dengan menggunakan metode K-NN dari data uji ke 1 dengan nilai K = 3 dan 5.

No.	Jarak Euclidean	K = 3	K = 5
52	5,291502622	TIDAK TEPAT WAKTU	TIDAK TEPAT WAKTU
94	6,244997998	TIDAK TEPAT WAKTU	TIDAK TEPAT WAKTU
59	6,32455532	TIDAK TEPAT WAKTU	TIDAK TEPAT WAKTU
93	6,708203932		TIDAK TEPAT WAKTU
46	8,888194417		TEPAT WAKTU

Dari hasil perhitungan K-Nearest Neighbor dengan nilai K = 3 adalah tidak tepat waktu dari hasil mayoritas tersebut. Sedangkan untuk hasil dari K = 5 didapatkan hasil dengan mayoritas tidak tepat waktu.

Perhitungan Metode *Naïve Bayes*

Hasil Perhitungan dengan menggunakan metode Naïve Bayes dari keseluruhan data uji.

No	PREDICTED CLASS		
	PREDIKSI TEPAT WAKTU	PREDIKSI TIDAK TEPAT WAKTU	PREDIKSI CLASS
33	0,000000060904269%	0,0000001120205444%	TEPAT WAKTU
34	0,000000027482666%	0,0000000627607211%	TEPAT WAKTU
35	0,000000000354462%	0,0000000000105091%	TIDAK TEPAT WAKTU
36	0,000000112157775%	0,0000000180722389%	TIDAK TEPAT WAKTU
....	...	...	...
38	0,000000147417249%	0,0000001146746381%	TIDAK TEPAT WAKTU

Dari Hasil Perhitungan Naïve Bayes datas, didapatkan hasil keterengan pada data uji ke 33 dan 34 masuk kedalam class tepat waktu karena presentase prediksi kelas tepat waktu lebih kecil di bandingkan prediksi kelas tidak tepat waktu.

Pengujian *Confusion Matrix* pada metode *K-Nearest Neighbor* dan *Naïve Bayes*

1. Hasil Pengujian performa pada metode K-Nearest Neighbor

	Precision	Recall	F1-score	Support
Tepat Waktu	0,33	0,29	0,31	7
Tidak Tepat Waktu	0,74	0,78	0,76	18
Accuracy			0,64	25
Macro avg	0,54	0,53	0,53	25
Weighted avg	0,62	0,64	0,63	25

Dari tabel di atas di ketahui akurasi dari KNN adalah 0,64 atau 64%. Kemudian Precision dan Recall dari class “Tepat Waktu” adalah 0,33 dan 0,29. Artinya model KNN ini memiliki nilai presisi “Tepat Waktu” sebesar 33% yaitu semua prediksi tepat waktu yang benar hanya 33% yang benar benar kelas Tepat Waktu, dan memiliki nilai Recall dari class “Tepat Waktu” hanya 29% dari keseluruhan data yang benar. Selain itu Precision dan Recall dari class “ Tidak Tepat Waktu” adalah 0,74 dan 0,78. Artinya model KNN ini juga memiliki nilai presisi “Tidak Tepat Waktu” sebesar 74% yaitu semua prediksi tidak tepat waktu yang benar 74% yang benar benar kelas Tidak Tepat Waktu, dan juga memiliki nilai Recall dari class “Tidak Tepat Waktu” sebesar 78% dari keseluruhan data yang benar.

2. Hasil Pengujian performa pada metode Naïve Bayes

	Precision	Recall	F1-score	Support
Tepat Waktu	0,50	0,29	0,36	7
Tidak Tepat Waktu	0,76	0,89	0,82	18
Accuracy			0,72	25
Macro avg	0,63	0,59	0,59	25
Weighted avg	0,69	0,72	0,69	25

Dari tabel di atas di ketahui akurasi dari Naive Bayes adalah 0,72 atau 72%. Kemudian Precision dan Recall dari class “Tepat Waktu” adalah 0,50 dan 0,29. Artinya model Naive Bayes ini memiliki nilai presisi “Tepat Waktu” sebesar 50% yaitu semua prediksi tepat waktu yang benar hanya 50% yang benar benar kelas Tepat Waktu, dan memiliki nilai Recall dari class “Tepat Waktu” hanya 29% dari keseluruhan data yang benar. Selain itu Precision dan Recall dari class “ Tidak Tepat Waktu” adalah 0,76 dan 0,89. Artinya model Naive Bayes ini juga memiliki nilai presisi “Tidak Tepat Waktu” sebesar 76% yaitu semua prediksi tidak tepat waktu yang benar 76% yang benar benar kelas Tidak Tepat Waktu, dan juga memiliki nilai Recall dari class “Tidak Tepat Waktu” sebesar 89% dari keseluruhan data yang benar.

KESIMPULAN DAN SARAN

Disimpulkan bahwa model Naive Bayes memiliki performa yang lebih baik daripada model KNN dalam memprediksi kelas "Tepat Waktu" dan "Tidak Tepat Waktu" dalam dataset yang digunakan. Nilai akurasi dari Naïve bayes 72% dan KNN 64%, nilai presisi dan recal dari Naïve Bayes menunjukan angka yang lebih tinggi untuk kelas "Tepat Waktu" (50% dan 29%) maupun "Tidak Tepat Waktu" (76% dan 89%) jika dibandingkan dengan model KNN (33%

dan 29% untuk "Tepat Waktu" dan 74% dan 78% untuk "Tidak Tepat Waktu"). Hal ini menandakan bahwa model Naive Bayes mampu menghasilkan prediksi yang lebih akurat dan konsisten dalam mengenali data yang digunakan.

DAFTAR REFERENSI

Aji Prasetya Wibawa, MGAPMFAFAD 2018, 'metode metode klasifikasi',.

Alim, S 2021, *IMPLEMENTASI ORANGE DATA MINING UNTUK KLASIFIKASI KELULUSAN MAHASISWA DENGAN MODEL K-NEAREST NEIGHBOR, DECISION TREE SERTA NAIVE BAYES ORANGE DATA MINING IMPLEMENTATION FOR STUDENT GRADUATION CLASSIFICATION USING K-NEAREST NEIGHBOR, DECISION TREE AND NAIVE BAYES MODELS*,.

Amelia Lizsara, P, Oyama, S & Wardani, S 2020, *Implementasi Data Mining Menggunakan Metode*,.

Annur, H 2018, *KLASIFIKASI MASYARAKAT MISKIN MENGGUNAKAN METODE NAÏVE BAYES*,.

Fatmawati 2016, 'perbandingan algortima klasifikasi data mining model c4.5 dan naive bayes untuk prediksi penyakit diabetes',.

Isnanto, S & Widodo, S 2021, 'PENERAPAN DATA MINING PADA PENERIMAAN MAHASISWA BARU DENGAN ALGORITMA K-MEANS CLUSTERING', *Jurnal Teknik Informasi dan Komputer (Tekinkom)*, vol. 4, no. 2, p. 158.

Lestari, Y, Sunardi, S & Fadlil, A 2022, 'Sistem Pendukung Keputusan Penerimaan Peserta Didik Baru dan Pemilihan Jurusan dengan Metode AHP dan SAW', *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 3, p. 1607.

Maskuri, MN, Sukerti, K & Herdian Bhakti, RM 2021, 'Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Penyakit Stroke Stroke Desease Predict Using KNN Algorithm', *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, vol. 4, no. 1.

Mulyati, S, Maulana Husein, S & Kunci, K 2020, 'RANCANG BANGUN APLIKASI DATA MINING PREDIKSI KELULUSAN UJIAN NASIONAL MENGGUNAKAN ALGORITMA (KNN) K-NEAREST NEIGHBOR DENGAN METODE EUCLIDEAN DISTANCE PADA SMPN 2 PAGEDANGAN sistem dapat memprediksi dan mengklasifikasikan dengan baik dan cepat', , pp. 65–73.

Nikmatun, IA & Waspada, I 2019, 'IMPLEMENTASI DATA MINING UNTUK KLASIFIKASI MASA STUDI MAHASISWA MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR', *Jurnal SIMETRIS*, vol. 10, no. 2.

Pratama, AR, Rizky Aryanto, R, Taufiq, A, Pratama, M & Korespondensi, P 2022, 'MODEL KLASIFIKASI CALON MAHASISWA BARU UNTUK SISTEM REKOMENDASI PROGRAM STUDI SARJANA BERBASIS MACHINE LEARNING', , vol. 9, no. 4.

Putro, HF, Vulandari, RT & Saptomo, WLY 2020, 'Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan', *Jurnal Teknologi Informasi dan Komunikasi (TIKomSiN)*, vol. 8, no. 2.

Saifudin, A 2018, ‘METODE DATA MINING UNTUK SELEKSI CALON MAHASISWA PADA PENERIMAAN MAHASISWA BARU DI UNIVERSITAS PAMULANG’, accessed from <https://dx.doi.org/10.24853/jurtek.10.1.25-36>.

Sri Widaningsih 2019, ‘PERBANDINGAN METODE DATA MINING UNTUK PREDIKSI NILAI DAN’,.

UNIVERSITAS MUHAMMADIYAH GRESIK 2023, ‘PROFIL SINGKAT UMG’, accessed January 5, 2023, from <https://umg.ac.id/list_profil>.

Utomo, DP & Mesran, M 2020, ‘Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung’, *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 4, no. 2, p. 437.