Perbandingan Metode Data Mining pada Prediksi Kelulusan Mahasiswa Fakultas Teknologi Informasi di Perguruan Tinggi dengan Algoritma Naive Bayes dan K-Nearest Neighbor

Indra
Fakultas Teknologi Informasi, Program Studi Teknik Informatika
Universitas Budi Luhur
Email: indra@budiluhur.ac.id

Dwi Agustinawati
Fakultas Teknologi Informasi, Program Studi Teknik Informatika
Universitas Budi Luhur
Email: 1911511887@student.budiluhur.ac.id

ABSTRAK

Kelulusan merupakan salah satu unsur penting bagi pihak fakultas untuk menentukan parameter keberhasilan dan akan sangat berpengaruh terhadap akreditasi yang nantinya menjadi salah satu indikator dari kualitas suatu perguruan tinggi. Saat ini masih banyak mahasiswa yang menempuh lama studi tidak sesuai target yang dijadwalkan. Setiap memasuki tahun ajaran baru, mahasiswa yang diterima makin bertambah sedangkan tidak semua mahasiswa dapat lulus tepat waktu sesuai dengan masa studi yang seharusnya. Metode yang digunakan untuk penelitian ini yaitu metode *Naïve Bayes* dan metode *K-Nearest Neighbor* untuk mengetahui hasil akurasi dari prediksi kelulusan mahasiswa. Hasil dari penelitian klasifikasi prediksi kelulusan mahasiswa yang menggunakan 500 data, dengan rincian data *training* sebanyak 400 data dan data *testing* sebanyak 100 yang diambil dari perbandingan 80% data *training* dan 20% data *testing*. Setelah dilakukan perhitungan dan pengujian pada penelitian ini hasil yang diperoleh pada algoritma *Naïve Bayes* mendapatkan nilai akurasi sebesar 74% dan hasil yang diperoleh pada algoritma *K-Nearest Neighbor* mendapatkan nilai akurasi sebesar 71%. Berdasarkan hasil akurasi yang diperoleh dapat dinyatakan bahwa nilai akurasi pada algoritma *Naïve Bayes* memiliki *performance* lebih baik dari algoritma *K-Nearest Neighbor*.

Kata kunci: klasifikasi, naïve bayes, kelulusan, k-nearest neighbor

ABSTRACT

Graduation is one of the important elements for the faculty to determine the parameters of success and will greatly affect the accreditation which will later become one of the indicators of the quality of a university. Currently, there are still many students who take the length of study not according to the scheduled target. Every time we enter a new academic year, the number of students accepted increases while not all students can graduate on time according to the supposed study period. The methods used for this research are the Naïve Bayes method and the K-Nearest Neighbor method to determine the accuracy of predicting student graduation. The results of the student graduation prediction classification research using 500 data, with details of 400 training data and 100 testing data taken from a ratio of 80% training data and 20% testing data. After calculating and testing this research, the results obtained in the Naïve Bayes algorithm get an accuracy value of 74% and the results obtained in the K-Nearest Neighbor algorithm get an accuracy value of 71%. Based on the accuracy results obtained, it can be stated that the accuracy value of the Naïve Bayes algorithm has better performance than the K-Nearest Neighbor algorithm.

Keywords: *classification, naïve bayes, graduation, k-nearest neighbor*

1. PENDAHULUAN

Kelulusan merupakan salah satu unsur penting bagi pihak fakultas untuk menentukan parameter keberhasilan dan akan sangat berpengaruh terhadap akreditasi yang nantinya menjadi salah satu indikator dari kualitas suatu perguruan tinggi. Syarat kelulusan mahasiswa biasanya dapat dilihat apabila mahasiswa telah menyelesaikan semua mata kuliah yang diwajibkan, memperoleh skor minimum pada tugas ataupun ujian, dan menyelesaikan tugas akhir atau skripsi. Saat ini masih banyak mahasiswa yang menempuh lama studi tidak sesuai target yang dijadwalkan. Setiap memasuki tahun ajaran baru, mahasiswa yang diterima makin bertambah sedangkan tidak semua mahasiswa dapat lulus tepat waktu sesuai dengan masa studi yang seharusnya. Banyak faktor yang menjadi pengaruh kelulusan mahasiswa terlambat, seperti status perkawinan mahasiswa, status mahasiswa (bekerja/tidak bekerja), tingkat pemahaman mahasiswa terhadap materi kuliah yang dapat dilihat dari IPK mahasiswa. Selain itu, tingkat kelulusan juga berpengaruh pada akreditasi suatu perguruan tinggi sehingga perguran tinggi berusaha untuk membantu mahasiswa agar lulus tepat waktu [1].

Dalam suatu perguruan tinggi memiliki lulusan tepat waktu setiap semesternya maka dapat membantu suatu perguruan tinggi tersebut dalam proses penilaian akreditasi, pada penerepannya tingkat kelulusan mahasiswa tidak bisa kita pastikan mereka lulus dengan tepat waktu [2]. Semakin banyak mahasiswa yang lulus tepat waktu maka semakin baik pula kinerja perguruan tinggi tersebut, sehingga tingkat kelulusan mahasiswa tepat waktu menjadi salah satu kriteria penilaian akreditasi bagi suatu perguruan tinggi atau program studi [3]. Evaluasi tingkat kelulusan yang dilakukan selama ini hanya berpedoman terhadap data pendaftaran wisuda, tanpa memperhatikan mahasiswa yang masih mengalami permasalahan akademis maupun administratif. Sedangkan tindak lanjut universitas terhadap mahasiswa yang tidak lulus tepat waktu dimungkinkan melalui cara membujuk, mengarahkan, dan membimbing mahasiswa untuk segera menyelesaikan studinya [4].

Pada penelitian sebelumnya melakukan klasifikasi menggunakan metode *Naïve Bayes* untuk memprediksi kelulusan tepat waktu pada mahasiswa dengan menggunakan 546 data mahasiswa prodi informatika fakultas teknik UHAMKA yang terdiri dari jenis kelamin dan index prestasi semester 1 sampai semester 4. Evaluasi model menggunakan *K- fold Cross Validation* dan hasil prediksi akan digunakan oleh dosen pembimbing akademik untuk mengevaluasi mahasiswa yang hasil prediksinya kurang memuaskan. Model dengan hasil terbaik yaitu model ke-3 dengan tingkat akurasi sebesar 80.19%, *recall* 80.26%, *precision* 92.75% dan *F-Measure* 86.05% [5]. Pada penelitian yang berbeda melakukan perbandingan *data mining* klasifikasi dengan empat algoritma yaitu C4.5, *Support vector machine* (SVM), *k-nearest neigbor* (KNN,) dan *Naïve Bayes* menggunakan data mahasiswa Teknik Informatika pada tahun 2008-2013. Hasil akhir dari keempat algoritma tersebut diperoleh bahwa algoritma *Naïve Bayes* merupakan algoritma terbaik untuk memprediksi kelulusan mahasiswa yang tepat waktu dan IPK ≥ 3 dengan nilai *accuracy* (76.79%), *error* (23.17%) , dan AUC (0.850) [6].

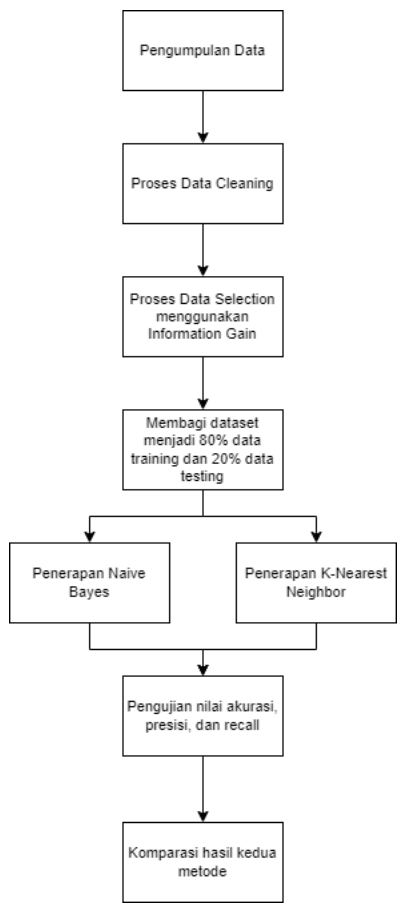
Pada penelitian sebelumnya juga melakukan perbandingan *data mining* dengan menggunakan algoritma kNN, NBC, dan SVM dalam prediksi kelulusan mahasiswa tingkat akhir dengan menggunakan data *public* sebanyak 379 data, diperoleh kesimpulan bahwa algoritma *K-Nearest Nighbor* (KNN) memiliki rata-rata yang lebih tinggi berdasarkan splitting data 90:10, 80:20, 70:30, 60:40, dan 50:50 yaitu dengan tingkat akurasi 87.8%, presisi 87.8%, dan *recall* 84%. Hasil pengujian menunjukkan bahwa algoritma *K-Nearest Nighbor* (KNN) dianggap memiliki rata-rata lebih tinggi (baik) dibandingkan *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM) terkait prediksi tingkat kelulusan mahasiswa tahap akhir (pascasarjana) [7]. Pada penelitian yang berbeda juga melakukan perbandingan dengan menggunakan metode *Naive Bayes* dan *K-Nearest Neighbor* dalam prediksi kelulusan mahasiswa dengan menggunakan data mahasiswa sebanyak 881 berdasarkan masuk pada tahun 2013 sampai 2016 dan data mahasiswa sebanyak 652 berdasarkan lulus pada angkatan 2013 sampai 2016. Berdasarkan perhitungan akurasi menggunakan *RapidMiner* Metode KNN yaitu sebesar 96.18% dan metode *Naive Bayes* sebesar 91.94%. Dari kedua nilai tersebut dapat dikatakan bahwa baik metode *Naive Bayes* ataupun metode *K-Nearest Neighbor* memiliki peluang yang kecil untuk melakukan kesalahan dalam proses prediksi [8].

Pada penelitian sebelumnya juga melakukan perbandingan model algoritma *Naïve Bayes, Decision Tree, Artificial Neural Network, K-Nearest Neighbor* (KNN), *Support Vector Machine (SVM)* dalam

melakukan prediksi tingkat kelulusan mahasiswa. Dataset yang digunakan dalam penelitian ini berjumlah 807 data mahasiswa fakultas teknik Universitas Hamzanwadi. Hasil temuan kami menunjukkan bahwa algoritma *Naïve Bayes* memiliki akurasi sebesar 92.37%, *Decision Tree* 91.60%, KNN 96.95%, SVM 93.13% dan ANN 90.84%. pengujian kelima algoritma tersebut, algoritma KNN memiliki tingkat akurasi terbaik sebesar 96.95%. hasil yang paling akurat dalam melakukan prediksi yaitu dengan model algoritma KNN dengan melakukan optimasi pada data tingkat akurasi yang didapatkan sebesar 96.95% termasuk kedalam kategori sangat baik [9]. Pada penelitian berbeda juga melakukan perbandingan hasil Analisa dua metode dalam algoritma klasifikasi untuk memprediksi kelulusan mahasiswa. Algoritma yang digunakan ialah Algoritma *K-Nearest Neighbour* dan *Naïve Bayes*. Penelitian ini juga bertujuan mengidentifikasi algoritma terbaik di antara dua pilihan algoritma klasifikasi tersebut. Penelitian ini menghasilkan kesimpulan bahwa algoritma *Naïve Bayes* memiliki tingkat akurasi yang sama dengan algoritma KNN dalam memprediksi kelulusan mahasiswa program studi Pendidikan Kedokteran yaitu sebesar 90% [10].

Keterkaitan penelitian ini dengan penelitian sebelumnya yaitu melakukan perbandingan metode *Naïve Bayes* dan *K-Nearest Neighbor* dalam prediksi kelulusan mahasiswa tepat waktu dengan menggunakan data mahasiswa sebanyak 500 data yang bersumber dari Direktorat Teknologi Informasi (DTI), data yang digunakan menggunakan data program studi Teknik Informatika dan Sistem Informasi pada tahun 2015 sampai 2018.

2. METODOLOGI PENELITIAN



Gambar 1. Metode Penelitian

Berdasarkan pada Gambar 1, langkah pertama yang dilakukan pada penelitian ini yaitu pengumpulan data, selanjutnya melakukan *pre-processing* data dengan menggunakan *data cleaning* dan *data selection* menggunakan *information gain*. Setelah persiapan data selesai dilakukan berikutnya melakukan perhitungan dengan metode *Naïve Bayes* dan *K-Nearest Neighbor*, kemudian melakukan pengujian dengan menghitung nilai akurasi, presisi, dan *recall*. Tahap terakhir setelah mengetahui hasil pengujian yaitu melakukan komparasi terharap kedua metode.

2.1. *Data Penelitian*

Tahapan awal yang dilakukan pada penelitian ini adalah menyiapkan data. Data yang digunakan adalah data mahasiswa program studi Teknik Informatika dan Sistem Informasi jenjang strata 1 yang bersumber dari Direktorat Teknologi Informasi (DTI) Universitas Budi Luhur, mulai dari angkatan 2015 sampai dengan angkatan 2018 yang berjumlah 500 data mahasiswa. Atribut yang digunakan adalah IPS semester 2 sampai semester 7 berdasarkan seleksi data yang dilakukan oleh *Information Gain* dan status kelulusan (tepat waktu/terlambat).

2.2. *Metode Pengumpulan Data*

Pada penelitian ini penulis merancang beberapa metode yang digunakan untuk mencari data, sebagai berikut:

a. Observasi
   Metode ini dilakukan dengan cara pengamatan di Direktorat Teknologi Informasi (DTI), serta terlibat secara langsung objek penelitian untuk memperoleh data penelitian.
b. Studi Pustaka
   Metode ini dilakukan karena berguna untuk memperoleh kajian pustaka, pembelajaran dari berbagai sumber dan dokumen seperti *e-book*, *journal*, dan berupa teori-teori pendukung lainnya yang berhubungan dengan penelitian ini.
c. Pengumpulan Data
   Pada tahap ini penulis mengumpulkan data mahasiswa berupa nilai *indeks* prestasi dari semester 1 sampai dengan semester 8 mulai dari tahun 2015-2018 yang di dapatkan dari Direktorat Teknologi Informasi (DTI) dan data tersebut akan dipelajari agar dapat dilakukan pemrosesan data.

2.3. *Pre-processing Data*

Pada tahap ini penulis melakukan *pre-processing* untuk memastikan data mahasiswa dapat dijadikan dataset. Berikut tahapan *pre-processing* yang dilakukan:

2.3.1. *Data Cleaning*

Tahap ini dilakukan untuk membuang data yang tidak konsisten dan *noise*. Dalam proses ini, beberapa data yang *noise* atau kosong sebagian data di hapus dan untuk beberapa yang masih kosong di isi dengan nilai *mean* yang diperoleh untuk melengkapi data tersebut. Tujuan perhitungan nilai *mean* yaitu untuk mencari nilai rata-rata dari jumlah total keseluruhan data yang digunakan dan yang tersusun dalam distribusi data. Rumus *mean* dapat dilihat pada persamaan (1).

$$\mu = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} \quad (1)$$

Keterangan:
x_i : Jumlah semua nilai
n : Banyaknya data

2.3.2. *Data Selection*

Tahap ini dilakukan untuk menyeleksi atribut yang akan digunakan. Teknik ini menghitung *information gain* atau *entropy* dari masing-masing atribut berdasarkan variabel *output*. Atribut-atribut yang memberikan lebih banyak informasi akan memiliki nilai *information gain* yang lebih tinggi dan dapat dipilih, sedangkan atribut yang kurang memberikan informasi akan mempunyai nilai *information gain* yang rendah dan dapat dibuang [11]. Rumus *feature selection information gain* dapat dilihat pada persamaan (2) dan (3).

$$Entropy(S) = -\sum_{i=1}^{n} p(x_i) * log_2\ p(x_i) \quad (2)$$

Keterangan:
n : Jumlah nilai yang mungkin untuk nilai *class* x
$p(x_i)$: Probabilitas *feature* ke-i

$$Gain(x,t) = Entropy(S) + \sum_{i=1}^{n} \frac{|x_i|}{|x|} Entropy(S_i) \quad (3)$$

Keterangan:
x : Jumlah *sample* untuk seluruh *class*
t : Atribut
S : *Entropy* seluruh *feature* (sebelum pemisahan)
n : Jumlah nilai yang mungkin untuk nilai *class* x
x_i : Jumlah *sample class* dengan nilai = i
S_i : *Entropy feature* untuk *class* i (setelah pemisahan)

2.4. *Penerapan Metode*

Tahap ini dilakukan untuk menerapkan metode dari data mining untuk mengolah dataset yang digunakan. Metode yang digunakan untuk penelitian ini yaitu metode *Naïve Bayes* dan metode *K-Nearest Neighbor*.

2.4.1. *Algoritma Naïve Bayes*

Naïve Bayes Classifier merupakan sebuah metode klasifikasi yang berakar pada *Theorema Bayes* [12]. *Theorema Bayes* adalah algoritma *Naïve Bayes* yang menggunakan pengalaman masa lalu untuk memprediksi apa yang akan terjadi di masa depan [13]. Metode ini memiliki kelebihan dalam menangani data kuantitatif dan data diskrit. Pada tahap ini akan dilakukan penerapan metode *Naïve Bayes* pada *dataset* yang akan dilakukan proses perhitungan pada data *training* dan data *testing*. Berikut tahapan yang dilakukan:

a. Membagi *dataset* menjadi 2 data yaitu data *training* dan data *testing*. Proses pembagian data dilakukan dengan membagi *dataset* menjadi 80% pada data *training* dan 20% pada data *testing*.
b. Melakukan pelabelan yang diambil dari atribut lama studi pada *dataset*. Pelabelan terdiri dari 2 yaitu tepat dan terlambat, dimana label tepat diambil dari lama studi kurang dari 4 tahun sedangkan untuk label terlambat diambil dari lama studi lebih dari 4 tahun.
c. Perhitungan data *training*

   Langkah pertama dalam perhitungan data numerik yaitu mencari nilai mean dari masing-masing atribut dan untuk rumus *mean* dapat dilihat pada persamaan (1). Langkah selanjutnya yaitu menghitung nilai *standar deviasi*. Tujuan perhitungan *standar deviasi* untuk menentukan seberapa dekat data dari sampel statistik dengan data rata-rata tersebut. adapun rumus perhitungan *standar deviasi* dapat dilihat pada persamaan (4).

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}} \quad (4)$$

Keterangan:
x_i : Jumlah semua nilai
n : Banyaknya data
μ : Nilai *mean*

Langkah selanjutnya, menghitung nilai probabilitas pada label tepat dan terlambat. Berikut rumus probabilitas dapat dilihat pada persamaan (5).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (5)$$

Keterangan:
$P(A|B)$: Probabilitas A terjadi dengan bukti bahwa B telah terjadi (probabilitas *superior*)
$P(B|A)$: Probabilitas B terjadi dengan bukti bahwa A telah terjadi
$P(A)$: Peluang terjadinya A
$P(B)$: Peluang terjadinya B

d. Perhitungan Nilai Distribusi *Gaussian*

Setelah mengetahui nilai *mean*, *standar deviasi* dan nilai probabilitas pada label, langkah berikutnya yaitu menghitung nilai probabilitas dengan menggunakan perhitungan distribusi *gaussian* dalam data numerik. Hasil perhitungan nilai *mean* dan nilai *standar deviasi* digunakan untuk perhitungan *distribusi gaussian*. Rumus perhitungan distribusi *gaussian* dapat dilihat pada persamaan (6).

$$P = n(X_i;\ \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i - \mu\sigma)^2}{2\sigma^2}} \quad (6)$$

Keterangan:
P : Peluang
X_i : Nilai atribut i
π : Nilai *phi* (3.14)
σ : *Standar deviasi*
μ : *Mean*, nilai rata-rata dari seluruh atribut

Setelah mengetahui nilai distribusi *gaussian*, langkah selanjutnya yaitu mengkalikan semua nilai *gaussian* tiap atibut dan dikalikan dengan nilai probabilitas yang memiliki label tepat atau terlambat seperti pada persamaan (7).

$$(X|Kelas) = P(X) \times P(Kelas) \quad (7)$$

Keterangan:
$P(X)$: Nilai probabilitas atribut
$P(Kelas)$: Nilai probabilitas label

2.4.2. *Perhitungan K-Nearest Neighbor*

K-Nearest Neighbor merupakan salah satu algoritma *machine learning*, algoritma *K-Nearest Neighbor* bekerja dengan cara melakukan pencarian terhadap nilai k objek atau pola pada data training yang tersedia yang paling mendekati dengan pola masukan dan memilih kelas dengan jumlah pola terbanyak diantara nilai k pola tersebut [14]. Secara singkat algoritma *K-Nearest Neighbor* menghitung jarak antara data uji dan data latih dengan menggunakan pengkuruan jarak tertentu [15]. Rumus perhitungan *K-Nearest Neighbor* dapat dilihat pada persamaan (8).

$$Euclidien\ distance = \sqrt{\sum_{i=1}^{p} (a_k - b_k)^2} \quad (8)$$

Keterangan:
a_k : Data sampel
b_k : Data uji *testing*
p : Dimensi data
i : Variabel data

2.5. *Pengujian*

Pada tahap ini akan dilakukan pengujian dengan menggunakan data testing yang nantikan akan dilakukan perhitungan akurasi, presisi, dan *recall*.

Confusion Matrix adalah pengukuran performa untuk masalah klasifikasi pada *machine learning* dimana keluarannya dapat berupa dua kelas atau lebih. Ada empat istilah yang merupakan representasi hasil proses klasifikasi pada *confusion matrix* yaitu *True Positif, True Negatif, False Positif,* dan *False Negatif.*

Tabel 1. *Confusion Matrix*

Clasification		*Actual*	
		1 *(Positive)*	0 *(Negative)*
Prediction	1 (*Positive*)	TP *(True Positive)*	FP *(False Positive)*
	0 (*Negative*)	FN *(False Negative)*	TN *(True Negative)*

Dari Tabel 1 dapat dilakukan perhitunga nilai akurasi, presisi, dan *recall*. Untuk menghitungnya dapat menggunakan rumus pada persamaan (9), (10), dan (11).

$$Acuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

3. HASIL DAN PEMBAHASAN

Data yang digunakan adalah data mahasiswa program studi Teknik Informatika dan Sistem Informasi jenjang strata 1 yang berjumlah sebanyak 500 data mahasiswa pada tahun 2015 sampai dengan tahun 2018. Atribut yang digunakan adalah indeks prestasi semester 2 sampai semester 7 berdasarkan seleksi data yang dilakukan oleh *Information Gain* dan label status kelulusan (tepat waktu/terlambat). Dari data tersebut akan dilakukan pembagian data terlebih dahulu dengan membagi *dataset* menjadi 80% pada data *training* dan 20% pada data *testing*. Data yang digunakan untuk perhitungan nilai probabilitas sebanyak 400 data yang diambil dari 80% data keseluruhan. Berikut adalah dataset yang digunakan:

Table 2. Dataset

No	*Nama*	*IPS S2*	*IPS S3*	*IPS S4*	*IPS S5*	*IPS S6*	*IPS S7*	*Status Kelulusan*
1	*Mahasiswa R1*	*3.58*	*3.85*	*3.83*	*3.86*	*3.93*	*4.00*	*Tepat*
2	*Mahasiswa R100*	*3.81*	*3.58*	*3.48*	*3.39*	*3.53*	*4.00*	*Tepat*
...								
499	*Mahasiswa R2822*	*2.90*	*3.19*	*3.72*	*3.77*	*3.83*	*3.84*	*Terlambat*
500	*Mahasiswa R2824*	*3.76*	*3.73*	*3.86*	*3.91*	*3.79*	*4.00*	*Tepat*

3.1. *Persiapan Data*

Sebelum data dilakukan perhitungan pada algoritma *Naïve Bayes* dan *K-Nearest Neighbor*, tahap pertama yang dilakukan yaitu *pre-processing* data. Berikut tahapan *pre-processing* data yang dilakukan:

3.1.1. *Data Cleaning*

Dalam *dataset* sebanyak 500 data terdapat data yang kosong pada kolom IPS 7 dan IPS 8, salah satunya pada baris ke 70 terdapat nilai yang kosong pada kolom IPS 8, berikut perhitungan nilai *mean* untuk mengisi nilai kosong pada kolom tersebut.

$$X = \frac{1508{,}99}{427} = 3{,}533934$$

Dari perhitungan di atas, diketahui untuk nilai *mean* yang dapat digunakan untuk mengisi data pada kolom IPS 8 yang kosong di baris ke 70 adalah 3,53. Tahap ini dilakukan sampai data pada kolom IPS 7 dan IPS 8 yang kosong terisi sesuai dengan tahun lulus lebih dari sama dengan 4 tahun. Untuk lebih jelas dapat dilihat pada Gambar 2.

1	NAMA	JENKEL	KELAS	IPS_SMT1	IPS_SMT2	IPS_SMT3	IPS_SMT4	IPS_SMT5	IPS_SMT6	IPS_SMT7	IPS_SMT8	IPK	Tahun Masuk	Tahun Lulus	Lama Studi
67	Mahasiswa R1093	Pria	Reguler	3.55	3.41	3.52	3.56	3.72	3.17	3.45	4.00	3.56	2016	2020	4.0
68	Mahasiswa R1094	Pria	Karyawan	3.82	3.60	3.19	4.00	3.94	3.64	3.97	4.00	3.84	2016	2020	4.0
69	Mahasiswa R1096	Pria	Reguler	3.09	3.31	3.54	3.46	3.19	3.52	3.90	4.00	3.47	2016	2020	4.0
70	Mahasiswa R1097	Pria	Reguler	3.83	3.93	3.84	3.75	3.55	3.73	4.00		3.77	2016	2021	4.5

1	NAMA	JENKEL	KELAS	IPS_SMT1	IPS_SMT2	IPS_SMT3	IPS_SMT4	IPS_SMT5	IPS_SMT6	IPS_SMT7	IPS_SMT8	IPK	Tahun Masuk	Tahun Lulus	Lama Studi
67	Mahasiswa R1093	Pria	Reguler	3.55	3.41	3.52	3.56	3.72	3.17	3.45	4.00	3.56	2016	2020	4.0
68	Mahasiswa R1094	Pria	Karyawan	3.82	3.60	3.19	4.00	3.94	3.64	3.97	4.00	3.84	2016	2020	4.0
69	Mahasiswa R1096	Pria	Reguler	3.09	3.31	3.54	3.46	3.19	3.52	3.90	4.00	3.47	2016	2020	4.0
70	Mahasiswa R1097	Pria	Reguler	3.83	3.93	3.84	3.75	3.55	3.73	4.00	3.53	3.77	2016	2021	4.5

Gambar 2. Proses Data *Cleaning*

3.1.2. *Data Selection*

Tahap ini dilakukan untuk menyeleksi atribut yang akan digunakan. Dalam penyeleksian data menggunakan *information gain* dengan mencari nilai *Entropy* terlebih dahulu pada seluruh atribut yang ada dalam *dataset* kemudian mencari nilai *information gain*. Berikut atribut yang ada sebelum dilakukan data *selection* menggunakan *information gain* dapat dilihat pada Tabel 3.

Tabel 3. Atribut Awal

Nama	*Jenkel*	*Kelas*	*IPS S1*	*IPS S2*	*IPS S3*	*IPS S4*	*IPS S5*	*IPS S6*	*IPS S7*	*IPS S8*	*IPK*	*Tahun Masuk*	*Tahun Lulus*
Mahasiswa R1	*Pria*	*Reguler*	*3.81*	*3.58*	*3.85*	*3.83*	*3.86*	*3.93*	*4.00*	*3.90*	*3.83*	*2015*	*2019*
Mahasiswa R100	*Pria*	*Reguler*	*3.53*	*3.81*	*3.58*	*3.48*	*3.39*	*3.53*	*4.00*	*4.00*	*3.62*	*2015*	*2019*
…													
Mahasiswa R2836	*Pria*	*Reguler*	*3.78*	*3.90*	*3.90*	*3.89*	*3.96*	*3.86*	*4.00*	*4.00*	*3.89*	*2018*	*2022*

Diketahui jumlah data pada label tepat sebanyak 359, jumlah data pada label terlambat sebanyak 141, dan jumlah total data sebanyak 500. Berikut merupakan perhitungan nilai *entropy* sesuai dengan persamaan (2).

$$Entropy(total) = \left(-\frac{359}{500} * log_2\left(\frac{359}{500}\right)\right) + \left(-\frac{141}{500} * log_2\left(\frac{141}{500}\right)\right)$$
$$= 0.343163 + 0.514998 = 0.858162$$

Tahap selanjutnya menghitung nilai *entropy* pada setiap atribut, untuk cara perhitungan sama seperti perhitungan *entropy* total. Dan setelah nilai *entropy* pada setiap atribut sudah diketahui, selanjutnya dilakukan untuk menghitung nilai *gain*. Berikut merupakan perhitungan nilai *gain* sesuai dengan persamaan (3).

Diketahui jumlah total data sebanyak 500, nilai *Entropy* (total) sejumlah 0.85816, jumlah data jenis kelamin pria sebanyak 433, nilai *Entropy* (jenis kelamin|pria) sejumlah 0.90042, jumlah data jenis kelamin wanita sebanyak 67, dan nilai *Entropy* (jenis kelamin|wanita) sejumlah 0.32626.

$$Gain(x,t) = 0.858162 - \left(\left(\frac{433}{500} * 0.90042\right) + \left(\frac{67}{500} * 0.32626\right)\right) = 0.03468$$

Tahap selanjutnya melanjutkan perhitungan *gain* pada atribut lainnya. Setelah mendapat nilai *gain* pada seluruh atribut dapat disimpulkan bahwa dapat dipilih 6 atribut yang akan digunakan dalam implementasi *Naïve Bayes* berdasarkan pada nilai *gain* tertinggi. Atribut yang digunakan yaitu, IPS 2 sampai dengan IPS 7. Berikut data dengan 6 atribut yang digunakan setelah dilakukan data *selection* dapat dilihat pada Tabel 4.

Table 4. Atribut Setelah Data *Selection*

Nama	*IPS S2*	*IPS S3*	*IPS S4*	*IPS S5*	*IPS S6*	*IPS S7*
Mahasiswa R1	*3.58*	*3.85*	*3.83*	*3.86*	*3.93*	*4.00*
Mahasiswa R100	*3.81*	*3.58*	*3.48*	*3.39*	*3.53*	*4.00*
…						
Mahasiswa R2836	*3.90*	*3.90*	*3.89*	*3.96*	*3.86*	*4.00*

3.2. *Implementasi Metode*

Data yang digunakan untuk perhitungan nilai *mean*, nilai *standar deviasi*, dan nilai probabilitas sebanyak 400 data yang diambil dari 80% data keseluruhan.

3.2.1. *Perhitungan Nilai Mean*

Pada tahap ini, dilakukan perhitungan untuk mencari nilai *mean* pada masing-masing atribut sesuai dengan label tepat dan label terlambat. Diketahui pada 400 data pada label status kelulusan terdapat 286 data tepat dan 114 data terlambat. Pada data IPS S2 dengan label tepat jika seluruh

data tersebut dijumlahkan maka mendapatkan hasil dengan nilai 955.19 dan pada data IPS S2 dengan label terlambat jika seluruh data tersebut dijumlahkan maka mendapatkan hasil dengan nilai 318.93. Berikut merupakan perhitungan nilai *mean* pada masing-masing atribut sesuai dengan persamaan (1).

$$Mean(IPS\ S2|Tepat) = \frac{955{,}19}{286} = 3.340$$

$$Mean(IPS\ S2|Terlambat) = \frac{318{,}93}{114} = 2.798$$

Tahap selanjutnya, menghitung nilai *mean* pada setiap atribut, untuk cara perhitungan sama seperti perhitungan *mean* pada atribut IPS S2.

3.2.2. *Perhitungan Nilai Standar Deviasi*

Pada tahap ini, dilakukan perhitungan nilai *standar deviasi* pada masing-masing atribut sesuai dengan label tepat dan label terlambat. Diketahui pada 400 data pada label status kelulusan terdapat 286 data tepat dan 114 data terlambat. Pada atribut IPS S2 dengan label tepat terdapat nilai $(x_i - \mu)^2$ dengan cara menghitung nilai tersebut menggunakan *excel* mendapatkan hasil 42.3106912587413 dan pada atribut IPS S2 dengan label terlambat terdapat nilai $(x_i - \mu)^2$ dengan cara menghitung nilai tersebut menggunakan *excel* dan mendapatkan hasil 47.1450605263158. Berikut merupakan perhitungan *standar deviasi* pada masing-masing atribut sesuai dengan persamaan (4).

$$\sigma(IPS\ S2|Tepat) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}} = \sqrt{\frac{42{,}3106912587413}{286-1}} = 0{,}385$$

$$\sigma(IPS\ S2|Terlambat) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}} = \sqrt{\frac{47{,}1450605263158}{114-1}} = 0{,}646$$

Tahap selanjutnya, menghitung nilai *standar deviasi* pada setiap atribut, untuk cara perhitungan sama seperti perhitungan *standar deviasi* pada atribut IPS S2.

3.2.3. *Perhitungan Nilai Probabilitas*

Pada tahap ini dilakukan perhitungan probabilitas pada label status kelulusan. Diketahui dari 400 data status kelulusan terdapat 286 data Tepat dan 114 data Terlambat. Data tersebut didapat dari lama studi mahasiswa, apabila mahasiswa tersebut melakukan kuliah <= 4 tahun maka ditanyakan lulus tepat waktu dan untuk mahasiswa melakukan kuliah > 4 tahun maka ditanyakan lulus dengan terlambat. Berikut merupakan perhitungan probabilitas pada label sesuai dengan persamaan (5).

$$P(Tepat) = \frac{Jumlah\ Data\ Tepat}{Jumlah\ Atribut\ Status\ Kelulusan} = \frac{286}{400} = 0{,}715$$

$$P(Terlambat) = \frac{Jumlah\ Data\ Terlambat}{Jumlah\ Atribut\ Status\ Kelulusan} = \frac{114}{400} = 0{,}285$$

Dari perhitungan diatas, dapat diketahui nilai probabilitas dari label tepat sejumlah 0.715 dan nilai probabilitas dari label terlambat sejumlah 0.285.

3.2.4. *Perhitungan Nilai Gaussian*

Pada tahap ini dilakukan perhitungan distribusi *gaussian* untuk memprediksi apakah hasil perhitungan yang dilakukan sesuai dengan data pada label status kelulusan. Berikut perhitungan nilai distribusi *gaussian* sesuai dengan persamaan (6).
Perhitungan nilai distribusi *gaussian* pada label tepat.

$$P(IPS\ S2|Tepat) = \frac{1}{\sqrt{2 \times 3{,}14 \times 0{,}385}} \times \left(-\frac{(2{,}94 - 3{,}340)^2}{2 \times 0{,}385^2}\right) = 0{,}375227563455295$$

Lakukan perhitungan yang sama pada atribut IPS S3 sampai IPS S7 pada label tepat dengan perhitungan seperti diatas. Berikut hasil perhitungan nilai distribusi *gaussian* pada label tepat.

Tabel 5. Hasil Perhitungan Nilai *Gaussian* Label Tepat

Atribut	*Nilai Gaussian*
IPS S2	*0.375227563455295*
IPS S3	*0.655735614598274*
IPS S4	*0.154252233494357*
IPS S5	*0.646523503993543*
IPS S6	*0.518652410057091*
IPS S7	*0.489756942573882*

Dari hasil perhitungan pada Tabel 5, selanjutnya dilakukan perkalian pada hasil nilai *gaussian* tiap atibut dan dikalikan dengan nilai probabilitas yang memiliki label tepat sesuai dengan persamaan (7).

$$Tepat = (0.375227563 \times 0.655735615 \times 0.154252233 \times 0.646523504 \times 0.51865241 \times 0.489756943) \times 0.715 = 0.004456586$$

Perhitungan nilai distribusi *gaussian* pada label terlambat.

$$P(IPS\ S2|Terlambat) = \frac{1}{\sqrt{2 \times 3{,}14 \times 0{,}646}} \times \left(-\frac{(2{,}94 - 2{,}798)^2}{2 \times 0{,}646^2}\right) = 0.484597794656475$$

Lakukan perhitungan yang sama pada atribut IPS S3 sampai IPS S7 pada label terlambat dengan perhitungan seperti diatas. Berikut hasil perhitungan nilai distribusi *gaussian* pada label terlambat.

Tabel 6. Hasil Perhitungan Nilai *Gaussian* Label Terlambat

Atribut	Nilai Gaussian
IPS S2	0.484597794656475
IPS S3	0.380833933132288
IPS S4	0.0779689620395021
IPS S5	0.278392878997492
IPS S6	0.229874014100321
IPS S7	0.203440542636132

Dari hasil perhitungan pada Tabel 6, selanjutnya dilakukan perkalian pada hasil nilai *gaussian* tiap atibut dan dikalikan dengan nilai probabilitas yang memiliki label terlambat sesuai dengan persamaan (7).

$$Terlambat = (0.484597795 \times 0.380833933 \times 0.077968962 \times 0.278392879 \times 0.229874014 \times 0.203440543) \times 0.285 = 0.000053391$$

3.3. *Pengujian Data*

Data yang digunakan pada *testing* untuk perhitungan nilai akurasi, presisi, dan *recall* sebanyak 100 data yang diambil dari 20% dari keseluruhan data.

3.3.1. *Implementasi Algoritma Naïve Bayes*

Pada tahap implementasi algoritma ini melakukan perhitungan pada persamaan (7) yaitu distribusi *gaussian*. Pada implementasi ini digunakan dataset sebanyak 500 data dengan perbandingan 80% untuk perhitungan data *training* sesuai dengan perhitungan pada poin 3.3 implementasi metode dan perbandingan 20% untuk perhitungan data *training* yang akan dilakukan menggunakan *tools Rapid Miner*. Berikut merupakan data hasil perhitungan *Naïve Bayes*.

Tabel 7. Hasil Prediksi *Naïve Bayes*

No	*Nama*	*IPS S2*	*IPS S3*	*IPS S4*	*IPS S5*	*IPS S6*	*IPS S7*	*Status Kelulusan*	*Predisi*
1	*Mahasiswa R2634*	*2.94*	*3.31*	*3.95*	*3.58*	*3.78*	*4.00*	*Tepat*	*Tepat*
2	*Mahasiswa R2636*	*2.55*	*3.15*	*3.74*	*3.82*	*3.92*	*3.88*	*Tepat*	*Tepat*
3	*Mahasiswa R2642*	*2.86*	*2.80*	*3.44*	*3.43*	*3.39*	*3.75*	*Terlambat*	*Tepat*
...									
100	*Mahasiswa R2836*	*3.90*	*3.90*	*3.89*	*3.96*	*3.86*	*4.00*	*Tepat*	*Tepat*

Setelah mendapatkan hasil prediksi yang dapat dilihat pada Tabel 7 menggunakan perhitungan probabilitas pada metode *Naïve Bayes*, diketahui prediksi tepat sebanyak 89 dan prediksi terlambat sebanyak 11. Kemudian akan dilakukan evaluasi menggunakan *confusion matrix* dengan menggunakan 100 data *testing*, didapatkan hasil *confussion matrix* berupa 68 data *True Positive*, 21 data *True Negative*, 5 data *False Posifitive*, dan 6 data *False Negative*. Untuk hasil evaluasi *confusion matrix* dapat dilihat pada Tabel 8.

Tabel 8. Evaluasi Confusion Matrix Naïve Bayes

Prediction	***Class***	
	Tepat	***Terlambat***
Tepat	*68*	*21*
Terlambat	*5*	*6*

Dari hasil evaluasi *confusion matrix* pada metode *Naïve Bayes* yang dapat dilihat pada Tabel 8 dapat dilakukan perhitungan nilai akurasi, presisi, dan *recall* sesuai pada persamaan (9), (10), dan (11).

$$Acuracy = \frac{68 + 6}{68 + 21 + 5 + 6} \times 100\% = \frac{74}{100} \times 100\% = 74\%$$

$$Precision = \frac{68}{68 + 21} \times 100\% = \frac{68}{89} \times 100\% = 76{,}40\%$$

$$Recall = \frac{68}{68 + 5} \times 100\% = \frac{68}{73} \times 100\% = 93{,}15\%$$

3.3.2. *Implementasi K-Nearest Neighbor*

Pada tahap implementasi algoritma ini melakukan perhitungan pada persamaan (8) yaitu menghitung jarak untuk menentukan nilai k. Pada implementasi ini digunakan dataset sebanyak 500 data dengan perbandingan 80% untuk perhitungan data *training* sesuai dengan perhitungan pada *point* 3.2 *implementasi* metode dan perbandingan 20% untuk perhitungan data *training* yang akan dilakukan menggunakan *tools Rapid Miner*. Berikut merupakan data hasil perhitungan *K-Nearest Neighbor*.

Tabel 9. Hasil Perhitungan K-Nearest Neighbor

No	*Nama*	*IPS S2*	*IPS S3*	*IPS S4*	*IPS S5*	*IPS S6*	*IPS S7*	*Status Kelulusan*	*Predisi*
1	*Mahasiswa R2634*	*2.94*	*3.31*	*3.95*	*3.58*	*3.78*	*4.00*	*Tepat*	*Tepat*
2	*Mahasiswa R2636*	*2.55*	*3.15*	*3.74*	*3.82*	*3.92*	*3.88*	*Tepat*	*Tepat*
3	*Mahasiswa R2642*	*2.86*	*2.80*	*3.44*	*3.43*	*3.39*	*3.75*	*Terlambat*	*Tepat*
…									
100	*Mahasiswa R2836*	*3.90*	*3.90*	*3.89*	*3.96*	*3.86*	*4.00*	*Tepat*	*Tepat*

Setelah mendapatkan hasil prediksi menggunakan perhitungan jarak pada metode *K-Nearest Neighbor* yang dapat dilihat pada Tabel 9, diketahui prediksi tepat sebanyak 92 dan prediksi terlambat sebanyak 8. Kemudian akan dilakukan evaluasi menggunakan *confusion matrix* dengan menggunakan 100 data *testing*, didapatkan hasil *confussion matrix* berupa 68 data *True Positive*, 24 data *True Negative*, 5 data *False Posifitive*, dan 3 data *False Negative*. Untuk hasil evaluasi *confusion matrix* dapat dilihat pada tabel dibawah ini.

Tabel 10. Evaluasi Confusion Matrix K-Nearest Neighbor

Prediction	*Class*	
	Tepat	*Terlambat*
Tepat	*68*	*24*
Terlambat	*5*	*3*

Dari hasil evaluasi *confusion matrix* pada metode *K-Nearest Neighbor* yang dapat dilihat pada Tabel 10 dapat dilakukan perhitungan nilai akurasi, presisi, dan *recall* sesuai pada persamaan (9), (10), dan (11).

$$Acuracy = \frac{68 + 3}{68 + 24 + \ 5 + 3} \times 100\% = \frac{71}{100} \times 100\% = 71\%$$

$$Precision = \frac{68}{68 + 24} \times 100\% = \frac{68}{92} \times 100\% = 73{,}91\%$$

$$Recall = \frac{68}{68 + 5} \times 100\% = \frac{68}{73} \times 100\% = 93{,}15\%$$

3.3.3. *Perbandingan Metode*

Setelah dilakukan pengujian pada kedua metode, selanjutkan akan dilakukan perbandingan dari hasil kedua metode yang digunakan. Berikut merupakan hasil perbandingan dari metode *Naïve Bayes* dan *K-Nearest Neighbor*.

Tabel 11. Hasil Perbandingan Metode

Metode	*Akurasi*	*Presisi*	*Recall*
Naïve Bayes	*74%*	*76,40%*	*93,15%*
K-Nearest Neighbor	*71%*	*73,91%*	*93,15%*

Pada penelitian ini menggunakan data mahasiswa program studi Teknik Informatika dan Sistem Informasi jenjang strata 1 pada tahun 2015 sampai dengan tahun 2018 sebanyak 500 data dengan proporsi kelas regular sebanyak 414 dan kelas karyawan sebanyak 86 sebagai dataset. Dilakukan tahapan preprocessing yang pertama yaitu *data cleaning*, kedua melakukan *data selection* dengan perhitungan *information gain* untuk menentukan atribut yang akan digunakan.

Dalam penelitian ini, dapat dilihat pada Tabel 11 bahwa algoritma *Naïve Bayes* menghasilkan nilai akurasi yang tinggi, menunjukkan bahwa perhitungan probabilitas pada algoritma *Naïve Bayes* lebih baik dibandingkan dengan perhitungan jarak k terdekat pada algoritma *K-Nearest Neighbor*. Pada penelitian ini mendukung dalam studi yang dilakukan oleh [Sri Widaningsih (2019) [ISSN (p): 1907-4964 | ISSN (e): 2655-089X]], melakukan perbandingan data *mining* klasifikasi dengan empat algoritma yaitu C4.5, *Support Vector Machine* (SVM), *K-Nearest Neigbor* (KNN,) dan *Naïve Bayes* menggunakan data mahasiswa Teknik Informatika pada tahun 2008-2013 dengan hasil akhir dari keempat algoritma diperoleh bahwa algoritma *Naïve Bayes* merupakan algoritma terbaik untuk memprediksi kelulusan mahasiswa yang tepat waktu.

Dari hasil penelitian ini dapat diketahui kelebihan dari algoritma *Naïve Bayes* yaitu perhitungan probabilitas mudah untuk diimplementasikan, sedangkan kelebihan dari algoritma *K-Nearest Neighbor* yaitu perhitungan jarak k terdekat mudah untuk diimplementasikan. Dan diketahui kekurangan dari algoritma *Naïve Bayes* yaitu jika nilai probabilitas sama maka tidak dapat menentukan nilai prediksi, sedangkan kekurangan dari algoritma *K-Nearest Neighbor* yaitu jika jarak k terdekat sama maka tidak bisa menentukan label hasil prediksi.

4. KESIMPULAN

Hasil dari penerapan metode *Naïve Bayes* dan metode *K-Nearest Neighbor* pada klasifikasi prediksi kelulusan mahasiswa. Setelah dilakukan perhitungan dan pengujian pada penelitian ini hasil yang diperoleh pada algoritma *Naïve Bayes* mendapatkan nilai akurasi sebesar 74% dan hasil yang diperoleh pada algoritma *K-Nearest Neighbor* mendapatkan nilai akurasi sebesar 71%. Berdasarkan hasil akurasi yang diperoleh dapat dinyatakan bahwa nilai akurasi pada algoritma *Naïve Bayes* memiliki *performance* lebih baik dari algoritma *K-Nearest Neighbor*. Berdasarkan hasil penelitian ini, terdapat beberapa saran yang dapat membantu dalam pengembangan topik atau sistem yang sudah dibuat pada kemudian hari yaitu dapat dilakukan perbandingan dengan algoritma lain supaya dapat mendukung pengujian yang lebih baik lagi kedepannya dan dapat mengetahui hasil perbandingan dengan algoritma apa yang mendapatkan nilai akurasi terbaik.

DAFTAR PUSTAKA

[1] N. Khasanah, A. Salim, N. Afni, R. Komarudin, and Y. I. Maulana, "Prediksi Kelulusan Mahasiswa Dengan Metode Naive Bayes," *Technol. J. Ilm.*, vol. 13, no. 3, p. 207, 2022, doi: 10.31602/tji.v13i3.7312.

[2] Y. Apridiansyah, N. D. M. Veronika, and E. D. Putra, "Prediksi Kelulusan Mahasiswa Fakultas Teknik Informatika Universitas Muhammadiyah Bengkulu Menggunakan Metode Naive Bayes," *JSAI (Journal Sci. Appl. Informatics)*, vol. 4, no. 2, pp. 236–247, 2021, doi: 10.36085/jsai.v4i2.1701.

[3] L. Setiyani, M. Wahidin, D. Awaludin, and S. Purwani, "Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Data Mining Naïve Bayes : Systematic Review," *Fakt. Exacta*, vol. 13, no. 1, p. 35, 2020, doi: 10.30998/faktorexacta.v13i1.5548.

[4] N. M. A. Mahar, Vihi Atina, and Nugroho Arif Sudibyo, "Pemodelan Prediksi Kelulusan Mahasiswa Dengan Metode Naïve Bayes Di Uniba," *J. Manaj. Inform. dan Sist. Inf.*, vol. 6, no. 2, pp. 148–158, 2023, doi: 10.36595/misi.v6i2.875.

[5] D. Anugrah Putra and M. Kamayani, "Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Naive Bayes di Program Studi Teknik Informatika UHAMKA," *Pros. Semin. Nas. Teknoka*, vol. 5, no. 2502, pp. 34–40, 2020, doi: 10.22236/teknoka.v5i.331.

[6] D. A. C, N. Bayes, and D. A. N. Svm, "PERBANDINGAN METODE DATA MINING

UNTUK PREDIKSI NILAI DAN WAKTU KELULUSAN MAHASISWA PRODI TEKNIK INFORMATIKA," vol. 13, no. 1, pp. 16–25, 2019.

[7] A. Putri, C. S. Hardiana, E. Novfuja, and ..., "Komparasi Algoritma K-NN, Naive Bayes dan SVM untuk Prediksi Kelulusan Mahasiswa Tingkat Akhir: Comparison of K-NN, Naive Bayes and SVM Algorithms for …," *... Indones. J. ...*, vol. 3, no. April, pp. 20–26, 2023, [Online]. Available: https://journal.irpi.or.id/index.php/malcom/article/view/610

[8] K. Kartarina, N. K. Sriwinarti, and N. luh P. Juniarti, "Analisis Metode K-Nearest Neighbors (K-NN) Dan Naive Bayes Dalam Memprediksi Kelulusan Mahasiswa," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 3, no. 2, pp. 107–113, 2021, doi: 10.35746/jtim.v3i2.159.

[9] Z. Amri, K. Kusrini, and K. Kusnawi, "Prediksi Tingkat Kelulusan Mahasiswa menggunakan Algoritma Naïve Bayes, Decision Tree, ANN, KNN, dan SVM," *Edumatic J. Pendidik. Inform.*, vol. 7, no. 2, pp. 187–196, 2023, doi: 10.29408/edumatic.v7i2.18620.

[10] M. Gunawan, M. Zarlis, and R. Roslina, "Analisis Komparasi Algoritma Naïve Bayes dan K-Nearest Neighbor Untuk Memprediksi Kelulusan Mahasiswa Tepat Waktu," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 513, 2021, doi: 10.30865/mib.v5i2.2925.

[11] I. G. Based, "Penerapan Feature Selection untuk Prediksi Lama Studi Mahasiswa," pp. 72–76.

[12] A. Basit, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Hasil Panen Padi," *J. Tek. Inform. Kaputama*, vol. 4, no. 2, pp. 208–213, 2020.

[13] F. Sholekhah, A. D. Putri, and L. Efrizoni, "Comparison of Naive Bayes and K-Nearest Neighbors Algorithms for Metabolic Syndrome Classification," *J. Homepage https//journal.irpi.or.id/index.php/malcom*, vol. 4, no. April, pp. 507–514, 2024.

[14] M. Norhalimi and T. A. Y. Siswa, "Optimasi Seleksi Fitur Information Gain pada Algoritma Naïve Bayes dan K-Nearest Neighbor," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 7, no. 3, pp. 237–255, 2022, doi: 10.14421/jiska.2022.7.3.237-255.

[15] M. Riyyan and H. Firdaus, "PERBANDINGAN ALGORITME NAÃ• VE BAYES DAN KNN TERHADAP DATA PENERIMAAN BEASISWA (Studi Kasus Lembaga Beasiswa Baznas Jabar)," *J. Inform. dan Rekayasa Elektron.*, vol. 5, no. 1, pp. 1–10, 2022, doi: 10.36595/jire.v5i1.547.