JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 7, No. 3, SEPTEMBER, 2022, Pp. 237 – 255
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Optimasi Seleksi Fitur *Information Gain* pada Algoritma Naïve Bayes dan K-Nearest Neighbor

Muhammad Norhalimi [(1)*], Taghfirul Azhima Yoga Siswa [(2)]
Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Muhammadiyah Kalimantan Timur, Samarinda
e-mail : muhammadnorhalimi1999@gmail.com, tay758@umkt.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 25 Juli 2022, direvisi 23 September 2022, diterima 23 September 2022, dan dipublikasikan 25 September 2022.

Abstract

There was an increase in the number of late payments of tuition fees by 3,018 from a total of 5,535 students at the end of 2020. This study uses the Python library which requires data to be of numeric type, so it requires data transformation according to the type of data in the study, data that has a scale is transformed using an ordinal encoder, and data that does not have a scale is transformed using one-hot encoding. The purpose of this study was to evaluate the performance of the Naïve Bayes algorithm and K-Nearest Neighbor with a confusion matrix in predicting late payment of tuition fees at UMKT. The dataset used in this study was sourced from the financial administration bureau as many as 12,408 data with a distribution of 90:10. Based on the results of the calculation of the selection of information gain features, the best 4 attributes that influence the research are obtained, namely faculty, study program, class, and gender. The results of the evaluation of the confusion matrix that have the best performance using the Naïve Bayes with information gain algorithm obtain an accuracy of 55.19%, while the K-Nearest Neighbor with information gain only obtains an accuracy of 50.76%. Based on the accuracy results obtained in the prediction of late payment of tuition fees by using attributes derived from information gain, it influences increasing the accuracy of Naïve Bayes, but the use of the information gain attribute on the K-Nearest Neighbor algorithm makes the accuracy obtained decrease.

Keywords: Prediction, Naïve Bayes, K-Nearest Neighbor, Information Gain, Confusion Matrix

Abstrak

Terjadi kenaikan angka keterlambatan pembayaran biaya kuliah sebanyak 3.018 dari total 5.535 mahasiswa pada periode akhir 2020. Penelitian ini menggunakan *library* Python yang mengharuskan data bertipe numerik, sehingga memerlukan transformasi data yang sesuai dengan jenis data pada penelitian, pada data yang memiliki skala dilakukan transformasi menggunakan *ordinal encoder,* pada data yang tidak memiliki skala ditransformasi menggunakan *one-hot encoding*. Tujuan penelitian ini adalah untuk mengevaluasi kinerja algoritma Naïve Bayes serta K-Nearest Neighbor dengan *confusion matrix* dalam memprediksi keterlambatan pembayaran biaya kuliah di UMKT. *Dataset* yang digunakan dalam penelitian ini bersumber dari biro administrasi keuangan sebanyak 12.408 data dengan pembagian 90:10. Berdasarkan hasil perhitungan seleksi fitur *information gain* memperoleh 4 atribut terbaik yang berpengaruh dalam penelitian yaitu fakultas, prodi, angkatan, dan gender. Hasil evaluasi *confusion matrix* yang memiliki kinerja terbaik menggunakan algoritma Naïve Bayes *with information gain* memperoleh *accuracy* sebesar 55,19%, sedangkan K-Nearest Neighbor *with information gain* hanya memperoleh *accuracy* 50,76%. Berdasarkan hasil akurasi yang diperoleh dalam prediksi keterlambatan pembayaran biaya kuliah dengan menggunakan atribut yang berasal dari *information gain* mempunyai pengaruh dalam meningkatkan akurasi Naïve Bayes, namun penggunaan atribut *information gain* terhadap algoritma K-Nearest Neighbor membuat akurasi yang diperoleh menjadi menurun.

Kata Kunci: Prediksi, Naïve Bayes, K-Nearest Neighbor, *Information Gain, Confusion Matrix*

1. PENDAHULUAN

Universitas Muhammadiyah Kalimantan Timur (UMKT) merupakan salah satu lembaga pendidikan dibawah naungan Muhammadiyah sebagai perguruan tinggi swasta yang pembiayaan kuliahnya dibebankan kepada mahasiswa melalui SPP. Berdasarkan data dari bagian Keuangan terdapat angka kenaikan dan penurunan dalam pembayaran SPP pada periode 2017 hingga 2020. Akan tetapi pada periode akhir 2020 justru mengalami kenaikan yang sangat drastis, bahkan sampai melewati jumlah batas mahasiswa yang membayar SPP tepat waktu. Jika dilihat dari jumlah mahasiswa bisa mencapai 3.018 dari total 5.535 mahasiswa Universitas Muhammadiyah Kalimantan Timur. Sehingga keterlambatan pembayaran kuliah tentu sangat berpengaruh terhadap operasional akademik dan menghambat pembangunan sarana dan prasarana penunjang pembelajaran. Untuk dapat menganalisis permasalahan tersebut, maka perlu dilakukan prediksi agar keterlambatan pembayaran biaya kuliah dapat dilakukan pencegahan dan penanganan sedini mungkin.

Penelitian tentang prediksi keterlambatan pembayaran biaya pendidikan pernah dilakukan sebelumnya seperti pada penelitian (Muqorobin et al., 2020) tentang sistem estimasi keterlambatan pembayaran SPP sekolah menggunakan algoritma Naïve Bayes, data pada penelitian ini bersumber dari dapodik 2017 hingga 2018, data yang digunakan berjumlah 236 data yang bertipe ordinal, data tersebut memiliki 6 atribut seperti, penghasilan orang tua, tanggungan keluarga, pendidikan ayah, usia ayah, pendidikan ibu, dan usia ibu. Hasil akurasi yang diperoleh sebesar 67%, kemudian pada penelitian (Rohmayani, 2020) tentang analisis prediksi keterlambatan pembayaran biaya siswa menggunakan algoritma Naïve Bayes dengan *optimasi particle swarm*, pada penelitian ini menggunakan 115 *dataset* yang bersumber dari angket yang diisi oleh siswa, *dataset* tersebut bertipe ordinal dan nominal, di dalamnya terdapat 8 atribut yang digunakan seperti, pendapatan ayah, jumlah tanggungan orang tua, uang saku/bulan, jasa keuangan, layanan akademik, program studi, cara pembayaran SPP, dan pekerjaan ibu. Hasil akurasi yang diperoleh sebesar 73,94%. Pada penelitian (Akhmad & Siswa, 2022) tentang prediksi pembayaran biaya kuliah mahasiswa di UMKT, pada penelitian ini menggunakan 12.408 data keuangan mahasiswa pada tahun 2017 sampai 2021, *dataset* tersebut bertipe ordinal, adapun atributnya seperti penghasilan ayah, penghasilan ibu, pendidikan ayah, dan pendidikan ibu, hasil akurasi yang diperoleh sebesar 52,82%. Dari beberapa penelitian sebelumnya yang menggunakan Naïve Bayes dan K-Nearest Neighbor dalam prediksi pembayaran biaya kuliah hasil akurasi yang diperoleh masih kurang maksimal sehingga pada penelitian ini menambahkah fitur seleksi information gain untuk meningkatkan kinerja dari algoritma Naïve Bayes dan K-Nearest Neighbor serta menambahkan *metode one-hot-encoding* untuk mengubah atribut kategorikal ke dalam bentuk numerik sehingga dapat bekerja lebih baik dengan algoritma klasifikasi Naïve Bayes dan *K-Nearest.* Menurut (Saputro & Sari, 2020) beberapa algoritma tidak dapat menggunakan variabel kategori sebagai masukannya, sehingga dibutuhkan perubahan terhadap variabel kategori tersebut agar dapat digunakan oleh suatu algoritma dalam proses komputasi.

Pada penelitian prediksi keterlambatan biaya kuliah ini menggunakan metode pendekatan *data mining* yaitu dengan mengkomparasi Algoritma Naïve Bayes dan K-Nearest Neighbor. Algoritma Naïve Bayes bekerja lebih baik dibanding model klasifikasi lainnya seperti *decision trees, neural networks* (Wanto et al., 2020). Algoritma Naïve Bayes juga telah banyak digunakan dalam penelitian sebelumnya pada bidang pendidikan. Seperti prediksi tingkat kelulusan mahasiswa tepat waktu pada UIN Syarif Hidayatullah Jakarta dengan hasil akurasi sebesar 80,72% (Salmu & Solichin, 2017). Prediksi tingkat kelulusan mahasiswa tepat waktu pada Fakultas Ekonomi dan Bisnis Universitas Pendidikan Nasional dengan hasil akurasi sebesar 98% (Suardika, 2019). Prediksi tingkat kelulusan peserta sertifikasi *Microsoft Office Specialist* (MOS) dengan akurasi sebesar 99.24% (Rifai et al., 2019). Prediksi masa studi mahasiswa, hasil akurasi yang diperoleh sebesar 85.17% (Amelia et al., 2017).

Algoritma K-Nearest Neighbor juga pernah dilakukan pada penelitian sebelumnya. Menurut (Suntoro, 2019) K-Nearest Neighbor memiliki kelebihan sehingga sering dipakai oleh para peneliti

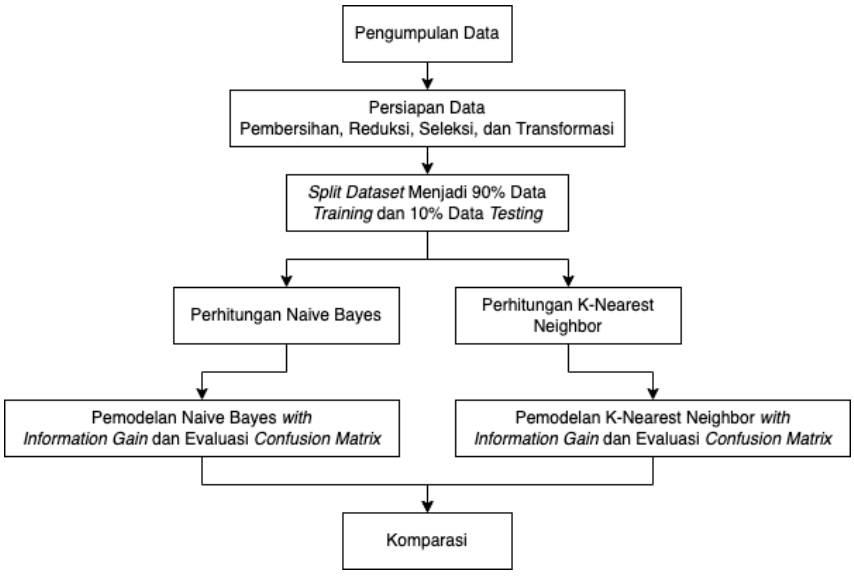
karena dapat memperoleh nilai akurasi yang tinggi dan tidak ada asumsi pada data. Penelitian yang menggunakan K-Nearest Neighbor seperti (Mustakim & Oktaviani, 2015) tentang prediksi prestasi mahasiswa, hasil akurasi yang diperoleh sebesar 82%. Prediksi tingkat kelulusan tepat waktu mendapatkan akurasi sebesar 85% pada algoritma Naïve Bayes sedangkan algoritma K-Nearest Neighbor menghasilkan 68.89% (Rahmatullah, 2019). Selanjutnya pada penelitian (Widaningsih, 2019) tentang perbandingan algoritma C4.5, Naïve Bayes, KNN, dan SVM terhadap prediksi nilai dan waktu kelulusan prodi TI menghasilkan akurasi sebesar 76,79% pada Naïve Bayes, SVM mendapatkan akurasi 74,04%, KNN dengan K=3 memperoleh akurasi 68,05%, dan C4.5 menghasilkan akurasi sebesar 75,96%.

Penelitian ini menggunakan seleksi fitur *information gain, Information gain* adalah perolehan informasi menggunakan *entropy* untuk menentukan atribut terbaik, *entropy* merupakan parameter untuk mengukur ketidakpastian yang di mana semakin tinggi *entropy*, maka semakin tinggi pula ketidakpastian (Suyanto, 2017). Seleksi fitur *information gain* telah banyak digunakan seperti pada penelitian (Muqorobin et al., 2019) tentang prediksi keterlambatan pembayaran sumbangan pembinaan pendidikan sekolah.

Penelitian ini juga menerapkan 2 metode transformasi yaitu *ordinal encoding* dan *one-hot encoding*. *Ordinal encoding* merupakan proses pemeringkatan yang dimulai dari skala terkecil 0 hingga ke n, *ordinal encoding* digunakan untuk mentrasformasi data yang memiliki tingkatan atau data yang bertipe ordinal. *One-Hot Encoding* merupakan proses untuk membuat suatu kolom baru dari variabel kategorikal, di mana setiap kategori menjadi kolom baru dengan nilai 0 atau 1 yang di mana nilai 0 menandakan tidak mewakili ada dan 1 mewakili ada (Daqiqil Id, 2021). *One-Hot Encoding* digunakan untuk mentrasformasi data yang tidak memiliki tingkatan dan datanya bertipe nominal, metode *One-Hot Encoding* pernah dilakukan pada penelitian sebelumnya seperti pada penelitian (Kinoto et al., 2020) tentang prediksi *employee churn* dengan *uplift* modeling. Perbedaan dengan penelitian ini adalah perpaduan transformasi data menggunakan *Ordinal Encoding* dan *One-Hot Encoding* serta penambahan fitur *selection information gain*. Penelitian ini bertujuan untuk mengidentifikasi atribut yang berpengaruh menggunakan fitur *selection information gain* dan mengevaluasi kinerja algoritma Naïve Bayes serta K-Nearest Neighbor dengan *confusion matrix* dalam memprediksi keterlambatan pembayaran biaya kuliah di Universitas Muhammadiyah Kalimantan Timur.

2. METODE PENELITIAN

Penelitian ini akan menggunakan 5 tahapan, adapun tahapannya seperti pada Gambar 1.



Gambar 1 Tahapan Penelitian

Berdasarkan pada Gambar 1 langkah pertama dalam penelitian adalah pengumpulan data, Langkah selanjutnya persiapan data terdiri dari pembersihan data, reduksi data, seleksi data menggunakan *information gain,* transformasi data *One-Hot Encoding*, dan *Ordinal Encoder*. Setelah persiapan data tahap berikutnya yaitu membagi dataset menjadi 90% *data training* dan 10% *data testing,* kemudian dilakukan perhitungan Naïve Bayes dan K-Nearest Neighbor secara manual. Pada tahap ini dilakukan permodelan algoritma Naïve Bayes *with Information gain* dan K-Nearest Neighbor *with Information gain,* setelah dilakukan permodelan, tahap berikutnya yaitu evaluasi menggunakan *confusion matrix* untuk melihat nilai akurasi. Tahap terakhir yaitu komparasi antara hasil *confusion matrix* dari permodelan algoritma Naïve Bayes *with Information gain* dan K-Nearest Neighbor *with Information gain.*

2.1 *Information Gain*

Information gain adalah perolehan informasi menggunakan *entropy* untuk menentukan atribut terbaik, *entropy* merupakan parameter untuk mengukur ketidakpastian yang di mana semakin tinggi *entropy*, maka semakin tinggi pula ketidakpastian (Suyanto, 2017). Rumus *feature selection information gain* dapat dilihat pada Pers. (1) dan (2) (Suntoro, 2019).

$$Entropy(S) \equiv \sum_{i=1}^{n} -pi * log_2\, p \qquad (1)$$

Di mana S merupakan himpunan kasus, n adalah jumlah partisi S, dan pi menunjukkan proporsi himpunan kasus ke-i.

$$Gain(S, A) \equiv Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \qquad (2)$$

Di mana S adalah himpunan kasus, A adalah atribut, n menunjukkan jumlah partisi atribut A, |S| adalah jumlah kasus dalam S, dan $|S_i|$ jumlah kasus pada partisi ke-i.

2.2 Algoritma Naïve Bayes

Algoritma Naïve Bayes dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas (Rajaraman & Ullman, 2011). Langkah-langkah permodelan sebagai berikut (Suntoro, 2019):

1) Membaca data *training*
2) Menghitung jumlah kelas target pada data training
3) Perhitungan menggunakan data numerik
   Langkah awal dalam perhitungan data numerik yaitu mencari nilai *mean* dan nilai standar deviasi dari masing-masing atribut yang menggambarkan data angka. Adapun rumus yang digunakan untuk menghitung nilai *mean* dapat dilihat pada Pers. (3).

$$\mu = \sum_{i}^{n} =_1\ x_i \quad atau\ \mu = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} \qquad (3)$$

   Di mana μ menyatakan nilai rata-rata hitung (*mean*), Xi merupakan nilai sampel ke -I, dan n menunjukkan jumlah sampel.

   Langkah selanjutnya menghitung nilai standar deviasi, adapun rumus perhitungan standar deviasi dapat dilihat pada Pers. (4).

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

   Di mana σ merupakan standar deviasi, Xi menyatakan nilai x ke -I, μ adalah nilai rata-rata hitung, dan n adalah jumlah sampel.

4) Nilai Distribusi *Gaussian*
Selanjutnya menghitung nilai probabilitas untuk fitur *data testing* yang memiliki data numerik. Rumus perhitungan distribusi Gaussian dapat dilihat pada Pers. (5).

$$P = (X_i = x_i \,|^{Y=yj}) = \frac{1}{\sqrt{2\pi\sigma ij}} \times e^{-\frac{(xi-\mu ij)^2}{2\sigma^2 ij}} \quad (5)$$

5) Menentukan Nilai Akhir
Setelah mendapatkan nilai distribusi *Gaussian*, langkah selanjutnya yaitu mengkalikan semua nilai *Gaussian* yang memiliki label tepat dan terlambat seperti pada Pers. (6).

$$(Kelas\,|X) = P(Kelas) \times P(X) \quad (6)$$

Kemudian setelah mendapatkan nilai akhir, langkah selanjutnya adalah membandingkan nilai antara Probabilitas keterangan "Tepat" dan Probabilitas keterangan "Terlambat".

2.3 Algoritma K-Nearest Neighbor

Algoritma 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 (Suyanto, 2017). Adapun rumus Euclidien distance dapat dilihat pada Pers. (7).

$$Euclidien\ distance = \sqrt{\sum_{i=1}^{p}(a_k - b_k)^2} \quad (7)$$

Di mana a_k merupakan sampel data, b_k merupakan data uji *testing*, P menyatakan dimensi data, dan i adalah variabel data.

2.4 *Confusion Matrix*

Confusion matrix merupakan sebuah teknik yang bisa dipakai untuk mengetahui seberapa akurat model klasifikasi menggunakan tabel *confusion matrix* (Primartha, 2021). *Confusion matrix* dapat dihitung menggunakan Pers. (8) (Id, 2021).

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

Di mana TP adalah *True Positive*, TN adalah *True Negative*, FP adalah *False Positive*, dan FN adalah *False Negative*.

3. HASIL DAN PEMBAHASAN

3.1 Data Penelitian

Data yang digunakan pada penelitian ini bersumber dari Biro Administrasi Keuangan berupa data pembayaran kuliah mahasiswa pada periode 2017 hingga 2020. Data yang diperoleh dari Bagian Administrasi Keuangan berjumlah 39.644, data tersebut terdiri dari 8.833 mahasiswa terlambat membayar biaya kuliah dan 30.811 mahasiswa tepat waktu dalam melakukan pembayaran kuliah. Data tersebut memiliki 9 atribut seperti fakultas, prodi, angkatan, gender, pendapatan ayah, pendapatan ibu, pendidikan ayah, pendidikan ibu, dan label target (tepat atau terlambat). Data biro administrasi keuangan dapat dilihat pada Tabel 1.

Tabel 1 Data Biro Administrasi Keuangan

No.	Fakultas	Prodi	Angkatan	Gender	Penghasilan Ayah	Penghasilan Ibu	Pend. Ayah	Pend. Ibu	Label
1	Ilmu Keperawatan	Keperawatan	2018	L	Rp. 2,000,000 - 4,999,999	Kurang dari Rp. 500,000		Tidak sekolah	Tepat
2	Ilmu Keperawatan	Keperawatan	2018	L	Kurang dari Rp. 500,000	Rp. 2,000,000 - 4,999,999	SMA	S1	Tepat
3	Ilmu Keperawatan	Keperawatan	2018	L	Rp. 2,000,000 - 4,999,999	Rp. 2,000,000 - 4,999,999		Tidak sekolah	Tepat
…	…	…	…	…	…	…	…	…	…
39644	Sains Dan Teknologi	Teknik Sipil	2017	L	Rp. 1,000,000 - 1,999,999	Kurang dari Rp. 500,000	SMP	SD	Terlambat

3.2 Persiapan Data

Pada tahapan ini terdapat beberapa persiapan data seperti pembersihan data, reduksi data, seleksi data, serta proses transformasi data untuk mendapatkan data yang berkualitas agar dapat mempermudah dalam proses *data mining* dan dijadikan masukan dalam tahap permodelan (*modeling*). Adapun prosesnya sebagai berikut.

3.2.1 Pembersihan Data

Pada tahap ini dilakukan pembersihan terhadap data yang tidak memiliki nilai atribut lengkap (*missing value*) dan eror. Adapun data awal berjumlah 39.644, terdapat 10.099 data yang perlu di bersihkan karena memiliki nilai atribut yang tidak lengkap dan eror, pada data ini tidak dapat dilakukan imputation terhadap data dikarenakan data yang kosong tidak dapat diterka seperti penghasilan orang tua dan pendidikan orang tua. Setelah dilakukan proses *cleaning* data menjadi 29.545. Data dapat dilihat pada Tabel 2.

Tabel 2 Data Setelah Melalui Proses Pembersihan

No	Fakultas	Prodi	Angkatan	Gender	Penghasilan Ayah	Penghasilan Ibu	Pend. Ayah	Pend. Ibu	Label
1	Ilmu Keperawatan	Keperawatan	2018	L	Kurang dari Rp. 500,000	Rp. 2,000,000 - 4,999,999	SMA	S1	Tepat
2	Ilmu Keperawatan	Keperawatan	2018	L	Rp. 500,000 - 999,999	Kurang dari Rp. 500,000	SD	SD	Tepat
3	Ilmu Keperawatan	Keperawatan	2018	L	Rp. 5,000,000 - 20,000,000	Rp. 2,000,000 - 4,999,999	S2	S1	Tepat
…	…	…	…	…	…	…	…	…	…
29545	Sains Dan Teknologi	Teknik Sipil	2017	L	Rp. 1,000,000 - 1,999,999	Kurang dari Rp. 500,000	SMP	SD	Terlambat

3.2.2 Reduksi Data

Pada tahap ini dilakukan penyeimbangan terhadap *dataset* yang ada dikarenakan data awal memiliki data tepat sebanyak 23.341 sedangkan data terlambat 6.204 sehingga data tersebut tidak seimbang. Maka perlu dilakukan penyeimbangan dengan cara mengambil secara acak data mahasiswa yang melakukan pembayaran kuliah secara tepat sebanyak 6.204 dan terlambat sebanyak 6.204 data. Proses reduksi data merujuk pada penelitian terdahulu seperti pada penelitian Ali *dkk,* (2019) di mana *dataset* yang memiliki kelas target yang paling banyak akan menyebabkan pemodelan menjadi bias terhadap kelas target yang sedikit. Sehingga dilakukan pengurangan terhadap *dataset* yang memiliki label target tepat menjadi 6.204 agar seimbang dengan label kelas target terlambat 6.204. peneliti menggunakan undersampling karena menurut (Kurniawan) teknik undersampling lebih menguntungkan dikarenakan hanya dengan sebagian data yang digunakan dalam penelitian proses komputasi bisa lebih hemat dan waktu proses menjadi lebih singkat, adapun syaratnya yaitu data yang diambil dapat mewakili dari seluruh data

populasi, dalam pengambilan data tidak ada patokan yang terkhusus, akan tetapi kita harus menghindari *sampling bias*. Adapun hasil reduksi dapat dilihat pada Tabel 3.

Tabel 3 Data Setelah Reduksi

No	Fakultas	Prodi	Angkatan	Gender	Penghasilan Ayah	Penghasilan Ibu	Pend. Ayah	Pend. Ibu	Label
1	Ekonomi Bisnis dan Politik	Manajemen	2019	P	Rp. 500,000 - 999,999	Kurang dari Rp. 500,000	SMA	SMP	Terlambat
2	Ilmu Keperawatan	Keperawatan	2019	P	Rp. 2,000,000 - 4,999,999	Kurang dari Rp. 500,000	SMA	SMA	Tepat
3	Ilmu Keperawatan	Keperawatan	2020	P	Rp. 1,000,000 - 1,999,999	Kurang dari Rp. 500,000	SMA	Tidak sekolah	Tepat
4	Sains Dan Teknologi	Teknik Informatika	2018	L	Rp. 2,000,000 - 4,999,999	Kurang dari Rp. 500,000	SMA	SMP	Tepat
…	…	…	…	…	…	…	…	…	…
12408	Ilmu Keperawatan	Keperawatan	2017	L	Rp. 2,000,000 - 4,999,999	Kurang dari Rp. 500,000	D3	SMA	Tepat

3.2.3 Seleksi Data

Pada tahap ini dilakukan penyeleksian data terhadap atribut-atribut yang akan digunakan dengan merujuk pada penelitian Muqorobin *dkk*, (2019) dengan menggunakan *feature selection information gain* agar mendapat atribut yang lebih informatif sehingga dapat meningkatkan akurasi serta efisien terhadap algoritma Naïve Bayes. Perhitungan *information gain* menggunakan *dataset* pada Tabel 3, terdapat 12.408 data yang di dalamnya memiliki 9 atribut. Adapun rumus perhitungan *information gain* dapat dilihat pada Pers. (1) dan (2).

1) Menghitung Nilai *Entropy*
Jumlah data kelas Tepat = 6204
Jumlah data kelas Terlambat = 6204
Jumlah data total = 12408

$$Entropy(total) = \left(-\frac{6204}{12408} * log_2\left(\frac{6204}{12408}\right)\right) + \left(-\frac{6204}{12408} * log_2\left(\frac{6204}{12408}\right)\right) = 1$$

Tahap selanjutnya menghitung nilai *entropy* setiap atribut, perhitungan nilai *entropy* setiap atribut sama seperti pada perhitungan *entropy* total. Berikut adalah hasil dari perhitungan *entropy* dapat dilihat pada Tabel 5.

2) Menghitung Nilai *Gain*
Tahap ini akan menghitung nilai *gain* dari masing masing atribut, adapun rumus perhitungan nilai *gain* dapat dilihat pada Pers. (2). Berikut adalah hasil dari perhitungan *gain* dapat dilihat pada Tabel 4.

Tabel 4 Perhitungan Nilai *Gain*

No.	Atribut	Gain
1	Fakultas	0,029837
2	Prodi	0,039215
3	Angkatan	0,071045
4	Gender	0,005528
5	Penghasilan Ayah	0,001513
6	Penghasilan Ibu	0,001077
7	Pendidikan Ayah	0,000974
8	Pendidikan Ibu	0,000974

Berdasarkan hasil *gain* pada Tabel 4 maka dipilih empat atribut yang memiliki nilai *gain* tertinggi untuk digunakan dalam implementasi Naïve Bayes *with information gain*, Atribut yang digunakan adalah fakultas, prodi, angkatan, dan gender.

Tabel 5 Perhitungan *Entropy*

Fakultas	**Jumlah**	**Tepat**	**Terlambat**	**Entropy**
Ekonomi Bisnis dan Politik	3362	1677	1685	0,999996
Farmasi	930	431	499	0,99614
Hukum	611	218	393	0,939988
Ilmu Keperawatan	2074	1433	641	0,892091
Keguruan Dan Ilmu Pendidikan	561	226	335	0,972595
Kesehatan Masyarakat	2021	1076	945	0,996967
Psikologi	688	302	386	0,98922
Sains Dan Teknologi	2161	841	1320	0,964263
Prodi	**Jumlah**	**Tepat**	**Terlambat**	**Entropy**
Farmasi	930	431	499	0,99614
Hubungan Internasional	356	162	194	0,994164
Hukum	611	218	393	0,939988
Keperawatan	1922	1281	641	0,918469
Kesehatan Lingkungan	447	205	242	0,995052
Kesehatan Masyarakat	1574	871	703	0,991767
Manajemen	3006	1515	1491	0,999954
Ners	152	152	0	0
Pendidikan Bahasa Inggris	287	109	178	0,957894
Pendidikan Olah Raga	274	117	157	0,984572
Psikologi	688	302	386	0,98922
Teknik Informatika	1038	461	577	0,990972
Teknik Mesin	404	128	276	0,900903
Teknik Sipil	719	252	467	0,934502
Angkatan	**Jumlah**	**Tepat**	**Terlambat**	**Entropy**
2017	1468	1320	148	0,471581
2018	2315	938	1377	0,973902
2019	5553	2524	3029	0,994026
2020	3072	1422	1650	0,996023
Gender	**Jumlah**	**Tepat**	**Terlambat**	**Entrophy**
L	5163	2314	2849	0,992241
P	7245	3890	3355	0,996063
Penghasilan Ayah	**Jumlah**	**Tepat**	**Terlambat**	**Entropy**
Kurang dari Rp. 500,000	3481	1800	1681	0,999157
Lebih dari Rp. 20,000,000	43	20	23	0,996486
Rp. 1,000,000 - Rp. 1,999,999	2871	1347	1524	0,997257
Rp. 2,000,000 - Rp. 4,999,999	3936	2016	1920	0,999571
Rp. 5,000,000 - Rp. 20,000,000	610	329	281	0,995529
Rp. 500,000 - Rp. 999,999	1467	692	775	0,99769
Penghasilan Ibu	**Jumlah**	**Tepat**	**Terlambat**	**Entropy**
Kurang dari Rp. 500,000	9086	4582	4504	0,999947
Lebih dari Rp. 20,000,000	12	7	5	0,979869
Rp. 1,000,000 - Rp. 1,999,999	966	461	505	0,998503
Rp. 2,000,000 - Rp. 4,999,999	1371	721	650	0,998065
Rp. 5,000,000 - Rp. 20,000,000	141	64	77	0,993859
Rp. 500,000 - Rp. 999,999	832	369	463	0,990773

Tabel 5 Perhitungan *Entropy* (lanjutan)

Pendidikan Ayah	Jumlah	Tepat	Terlambat	Entropy
Tidak Sekolah	2836	1429	1407	0,999957
TK	1	1	0	0
SD	1500	796	704	0,997285
SMP	1219	616	603	0,999918
SMA	4681	2302	2379	0,999805
D1	45	15	30	0,918296
D2	41	18	23	0,989245
D3	242	124	118	0,999557
S1	1512	751	761	0,999968
S2	310	142	168	0,99492
S3	21	10	11	0,998364
Pendidikan Ibu	**Jumlah**	**Tepat**	**Terlambat**	**Entropy**
Tidak Sekolah	2872	1443	1429	0,999983
TK	2	2	0	0
SD	2264	1148	1116	0,999856
SMP	1830	927	903	0,999876
SMA	3668	1797	1871	0,999706
D1	48	25	23	0,998747
D2	19	9	10	0,998001
D3	260	121	139	0,99654
S1	1352	681	671	0,999961
S2	84	47	37	0,989753
S3	9	4	5	0,991076

3.2.4 Transformasi Data

Pada tahap ini dilakukan transformasi data, yaitu merubah data yang bertipe kategori menjadi numerik. Menurut Kurniawan (2020) algoritma *machine learning* membutuhkan data numerik untuk melakukan *training dataset*, dalam melakukan tranformasi menggunakan metode *One-Hot Encoding* dan *Ordinal Encoding. One-Hot Encoding* merupakan proses untuk membuat suatu kolom baru dari variabel kategorikal, di mana setiap kategori menjadi kolom baru dengan nilai 0 atau 1 yang di mana nilai 0 menandakan tidak mewakili ada dan 1 mewakili ada (Id, 2021).

Alasan peneliti menggunakan transformasi pada penelitian ini adalah karena data yang digunakan dalam penelitian terdapat 2 jenis data yaitu kategorikal dan numerik, algoritma *Gaussian* Naïve Bayes tidak dapat digunakan jika terdapat data kategorikal (ibnu daqiqil). Sehingga perlu adanya transformasi data dari kategorikal menjadi numerik, transformasi dilakukan terhadap 2 tipe data yaitu nominal dan ordinal, sehingga pada data yang tidak memiliki tingkatan seperti fakultas, prodi dan gender dilakukan menggunakan transformasi *One-Hot Encoding*. Sedangkan pada data yang memiliki skala seperti penghasilan ayah, penghasilan ibu, pendidikan ayah, dan pendidikan ibu dilakukan menggunakan *Ordinal Encoding* dengan cara data Ordinal yang memiliki skala paling kecil diubah ke dalam bentuk numerik dengan bilangan 0 hingga n sebanyak kategori perfitur. Hasil dari transformasi *One-Hot Encoding* mendapatkan 25 atribut seperti pada tabel 6. Sedangkan hasil dari *Ordinal Encoding* dapat dilihat pada Tabel 7 dan 8. Adapun data setelah proses transformasi *One-Hot Encoding* dan *Ordinal Encoding* dapat dilihat pada Tabel 9

Tabel 6 Data Setelah Proses Transformasi *One-Hot Encoding*

No.	F. Ekonomi Bisnis dan Politik	F. Farmasi	F. Hukum	F. Ilmu Keperawatan	F. Keguruan Dan Ilmu Pendidikan	…	Label
1	1	0	0	0	0	…	Terlambat
2	0	0	0	1	0	…	Tepat
3	0	0	0	1	0	…	Tepat
4	0	0	0	0	0	…	Tepat
…	…	…	…	…	…	…	…
12408	0	0	0	1	0	…	Tepat

Tabel 7 Pembobotan Penghasilan Ayah dan Ibu

Pembobotan Penghasilan Ayah dan Ibu	
Kurang dari Rp. 500,000	0
Rp. 500,000 - Rp. 999,999	1
Rp. 1,000,000 - Rp. 1,999,999	2
Rp. 2,000,000 - Rp. 4,999,999	3
Rp. 5,000,000 - Rp. 20,000,000	4
Lebih dari Rp. 20,000,000	5

Tabel 8 Pembobotan Pendidikan Ayah Dan Ibu

Pembobotan Pendidikan Ayah Dan Ibu	
Tidak Sekolah	0
TK	1
SD	2
SMP	3
SMA	4
D1	5
D2	6
D3	7
D4 / S1	8
S2	9
S3	10

Tabel 9 Hasil Data Setelah Proses Transformasi

No.	F. Ekonomi Bisnis dan Politik	F. Farmasi	F. Hukum	F. Ilmu Keperawatan	F. Keguruan Dan Ilmu Pendidikan	…	Label
1	1	0	…	2	8	4	Terlambat
2	0	0	…	0	4	3	Terlambat
3	0	0	…	0	4	2	Terlambat
4	0	0	…	1	4	3	Terlambat
…	…	…	…	…	…	…	…
12408	0	0	…	2	8	4	Tepat

3.3 *Split* Data

Penelitian ini menggunakan *Split* data agar mendapatkan hasil estimasi akurasi yang baik dari proses *dataset*. *Dataset* yang digunakan kemudian dibagi menjadi 90% *data training* dan 10% *data testing*.

Tabel 10 Pembagian *Data Training* dan *Testing*

Hasil Pembagian Data	
Jumlah Data Training	11.167
Jumlah Data Testing	1.241

Tabel 10 merupakan hasil pembagian data 90% data *training* dan 10% data *testing*. Didapatkan hasil jumlah data *training* sebanyak 11.167 dan data *testing* 1.241.

3.4 Perhitungan Algoritma Naïve Bayes

Pada tahap ini dilakukan pemilihan model yang sesuai agar dapat mengoptimalkan hasil akurasi. Adapun algoritma yang digunakan dalam permodelan ini adalah algoritma Naïve Bayes, data yang digunakan pada permodelan adalah data *training* berjumlah 11.167 di dalamnya terdapat 30 atribut. Adapun data dapat dilihat pada Tabel 11.

Tabel 11 Data *Training*

No.	F. Ekonomi Bisnis dan Politik	F. Farmasi	F. Hukum	F. Ilmu Keperawatan	F. Keguruan Dan Ilmu Pendidikan	…	Label
1	1	0	0	0	0	…	Terlambat
2	0	1	0	0	0	…	Terlambat
3	0	0	1	0	0	…	Terlambat
…	…	…	…	…	…	…	…
11167	0	0	0	0	0	…	Tepat

Tabel 12 Data *Testing*

No.	F. Ekonomi Bisnis dan Politik	F. Farmasi	F. Hukum	F. Ilmu Keperawatan	F. Keguruan Dan Ilmu Pendidikan	…	Label
1	0	0	0	0	0	…	Terlambat
2	0	0	0	1	0	…	Tepat
3	0	0	1	0	0	…	Tepat
…	…	…	…	…	…	…	…
1241	1	0	0	0	0	…	Terlambat

Setelah membagi data, langkah selanjutnya melakukan perhitungan menggunakan algoritma Naïve Bayes adapun langkahnya sebagai berikut.

3.4.1 Menghitung nilai probabilitas

Langkah pertama adalah menghitung nilai probabilitas dari kelas label "Tepat" dan "Terlambat". Di mana jumlah data yang memiliki kelas label "Tepat" sebanyak 5.568 data dan jumlah kelas label "Terlambat" sebanyak 5.599 data kemudian dibagi dengan jumlah data label target sebanyak 11.167 data.

$$P(label = Tepat) = \frac{5.568}{11.167} = 0{,}5$$

$$P(label = Terlambat) = \frac{5.599}{11.167} = 0{,}5$$

3.4.2 Menghitung nilai *mean*

Langkah kedua adalah menghitung nilai *mean* pada setiap atribut yang ada pada data *training*. Perhitungan *mean* akan dilakukan menggunakan Pers. (3) pada Microsoft Excel, adapun hasil perhitungan dapat dilihat pada Tabel 13.

Tabel 13 Perhitungan *Mean* Tepat dan Terlambat

	F. Hukum	F. Ilmu Keperawatan	Keguruan dan Ilmu Pendidikan	Kesehatan Masyarakat	…	Pendidikan Ibu Tidak Sekolah
Tepat	0,036099	0,231142	0,36099	0,175287	…	0,230603
Terlambat	0,064666	0,102715	0,051804	0,151661	…	0,230260

3.4.3 Menghitung nilai standar deviasi

Setelah mencari nilai *mean*, langkah selanjutnya adalah menghitung nilai standar deviasi menggunakan Pers. (4) pada setiap atribut *data training* seperti berikut.

Atribut “Fakultas Ekonomi Bisnis dan Politik”

$$\sigma_{label} = Tepat \sqrt{\frac{\sum(0-0{,}270115)^2 + \ldots + (0-0{,}270115)^2 + (0-0{,}270115)^2}{5.568}} = 0{,}412470$$

$$\sigma_{label} = terlambat \sqrt{\frac{\sum(0-0{,}272419)^2 + \ldots + (1-0{,}272419)^2 + (0-0{,}272419)^2}{5.599}} = 0{,}488042$$

Perhitungan standard deviasi selanjutnya akan dilakukan menggunakan Microsoft Excel, adapun hasilnya dapat dilihat pada Tabel 14.

Tabel 14 Standard Deviasi Tepat dan Terlambat

	F. Hukum	F. Ilmu Keperawatan	Keguruan dan Ilmu Pendidikan	Kesehatan Masyarakat	…	Pendidikan Ibu Tidak Sekolah
Tepat	0,186537	0,421563	0,186537	0,380213	…	0,421221
Terlambat	0,235935	0,303587	0,221632	0,358692	…	0,420999

3.4.4 Menghitung nilai *Gaussian*

Langkah selanjutnya adalah menghitung nilai *Gaussian* menggunakan Pers. (5) pada data *testing*. Perhitungan nilai *gaussian* tepat dan terlambat akan dilakukan menggunakan Microsoft Excel, hasil perhitungan dapat dilihat pada Tabel 15 dan 16.

Tabel 15 Perhitungan Nilai *Gaussian* Tepat

No.	F. Hukum	F. Ilmu Keperawatan	Keguruan dan Ilmu Pendidikan	Kesehatan Masyarakat	…	Pendidikan Ibu Tidak Sekolah
1	0,906787	0,52882	0,906787	0,581907	…	0, 529276
2	0, 906787	0, 52882	1,47E-06	0, 581907	…	0, 529276
3	0, 906787	0, 52882	0,906787	0, 581907	…	0,529276
4	0, 906787	0, 52882	0, 906787	0, 581907	…	0,115948
…	…	…	…	…	…	…
1241	0, 906787	0,116489	0, 906787	0, 581907	…	0,529276

Tabel 16 Perhitungan Nilai *Gaussian* Terlambat

No.	F. Hukum	F. Ilmu Keperawatan	Keguruan dan Ilmu Pendidikan	Kesehatan Masyarakat	…	Pendidikan Ibu Tidak Sekolah
1	0,777314	0,683945	0,824785	0,609311	…	0,529567
2	0,777314	0,683945	8,99E-05	0,609311	…	0,529567
3	0,777314	0,683945	0,824785	0,609311	…	0,529567
4	0,777314	0,683945	0,824785	0,609311	…	0,115605
…	…	…	…	…	…	…
1241	0,777314	0,009182	0,824785	0,609311	…	0,529567

3.4.5 Menentukan nilai akhir

Pada penentuan nilai akhir, langkah yang harus dilakukan adalah mengkalikan semua nilai *Gaussian* yang memiliki label tepat dan terlambat dengan probabilitas kelas target. Selanjutnya membandingkan nilai antara Probabilitas keterangan "Tepat" dan Probabilitas keterangan "Terlambat". Maka didapatkan hasil prediksi data *testing* pertama adalah Terlambat. Perhitungan data *testing* menggunakan Pers. (6) dilakukan menggunakan *tools* tambahan yaitu Microsoft Excel seperti pada Tabel 17.

Tabel 17 Penentuan Nilai Akhir Tepat dan Terlambat

No.	Nilai Akhir Tepat	Nilai Akhir Terlambat
1	2,2E-14	0,62775E-14
2	1,67E-29	4,93229E-23
3	3,31E-13	4,8002E-12
4	1,74E-10	1,9473E-11
…	…	…
1241	3,41E-08	6,66305E-09

Berdasarkan Tabel 17 dilakukan perbandingan pada nilai akhir data *testing* terlambat dan nilai akhir data *testing* tepat, adapun datanya dapat dilihat pada Tabel 18.

Tabel 18 Aktual dan Prediksi

No.	Aktual	Prediksi
1	Tepat	Terlambat
2	Tepat	Terlambat
3	Terlambat	Terlambat
4	Tepat	Tepat
…	…	…
1241	Tepat	Tepat

Hasil dari permodelan menggunakan algoritma Naïve Bayes mendapatkan probabilitas prediksi tepat sebesar 613 dan probabilitas prediksi terlambat sebesar 628. Tahap selanjutnya yaitu proses evaluasi, evaluasi merupakan tahapan yang digunakan untuk membantu pengukuran pada algoritma. Penulis menggunakan *Confusion Matrix* untuk melakukan pengukuran evaluasi pada model. Adapun rumus perhitungan *accuracy* dapat dilihat pada Pers. (8).

Proses evalusi *confusion matrix* dengan menggunakan 1.241 data *testing*, didapatkan hasil *confusion matrix* berupa 22 data *True Positive,* 620 data *True Negative*, 591 *False Positive*, dan 8 *False Negative.* Kemudian dilakukan proses perhitungan melalui *confusion matrix* seperti pada Tabel 19.

Hasil akurasi yang diperoleh dari perhitungan *confusion matrix* menggunakan data *testing* 1.241 memperoleh *accuracy* sebesar 51,73%.

Tabel 19 Evaluasi *Confusion Matrix*

Predict	*Actual*	
	Tepat	**Terlambat**
Tepat	22	591
Terlambat	8	620

3.5 ermodelan Algoritma Naïve Bayes *with Information Gain*

Permodelan ini akan menggunakan algoritma Naïve Bayes *with information gain* pada pemrograman Python dengan pembagian data 90:10. Data yang digunakan pada permodelan adalah data setelah melalui proses seleksi fitur *Information gain* dan dilakukan telah di transformasi. Adapun data dapat dilihat pada Tabel 20, di dalamnya terdapat 26 atribut.

Tabel 20 Data *Training*

No.	F. Ekonomi Bisnis dan Politik	F. Farmasi	…	Angkatan	Gender L	Gender P	Label
1	1	0	…	2019	1	0	Tepat
2	0	1	…	2018	0	1	Terlambat
3	0	0	…	2019	0	1	Tepat
…	…	…	…	…	…	…	…
11167	0	0	…	2018	0	1	Terlambat

Tabel 21 Data *Testing*

No.	F. Ekonomi Bisnis dan Politik	F. Farmasi	…	Angkatan	Gender L		Gender P	Label
1	0	0	…	2018	0		1	Terlambat
2	0	0	…	2017	0		1	Tepat
3	0	0	…	2020	1		0	Tepat
…	…	…	…	…	…	…	…	
1241	1	0	…	2020	0		1	Terlambat

```
from sklearn.naive_bayes import  GaussianNB
classifier = GaussianNB()
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
```

Gambar 2 Permodelan Naïve Bayes *with Information Gain*

Hasil dari permodelan mendapatkan probabilitas prediksi tepat sebesar 608 dan probabilitas prediksi terlambat sebesar 633. Selanjutnya dilakukan evalusi menggunakan *confusion matrix* dengan menggunakan 1.241 data *testing*, didapatkan hasil *confusion matrix* berupa 52 data *True Positive,* 633 data *True Negative*, 556 *False Positive*, dan 0 *False Negative.* Kemudian dilakukan proses perhitungan melalui *confusion matrix* seperti pada Tabel 22.

Tabel 22 Evaluasi *Confusion Matrix*

Predict	*Actual*	
	Tepat	**Terlambat**
Tepat	52	556
Terlambat	0	633

Hasil uji coba menggunakan algoritma Naïve Bayes *with information gain* menggunakan pemrograman Python mendapatkan *accuracy* sebesar 55,19%

3.6 Perhitungan Algoritma K-Nearest Neighbor

Perhitungan Algoritma K-Nearest Neighbor dapat dilakukan dengan beberapa tahapan yaitu menentukan nilai k, menghitung jarak dengan Pers. (7), selanjutnya melakukan *sorting* dari jarak terkecil hingga terbesar, tahap terakhir yaitu klasifikasi *test* data mayoritas. Berikut contoh perhitungan manual menggunakan 10 data sampel yang diambil secara acak, di dalamnya terdapat 9 data *training* dan 1 data *testing*.

Tabel 23 Sampel Data *Training*

No.	Fakultas Ekonomi Bisnis dan Politik	Fakultas Farmasi	…	Pendidikan Ayah	Pendidikan Ibu	Label
1	1	0	…	8	4	Terlambat
2	0	0	…	4	3	Terlambat
3	0	0	…	4	2	Terlambat
4	0	0	…	4	3	Terlambat
5	0	0	…	4	0	Terlambat
6	0	1	…	4	5	Tepat
7	1	0	…	8	9	Tepat
8	0	0	…	4	4	Tepat
9	0	0	…	4	3	Tepat

Tabel 24 Sampel Data *Testing*

No.	Fakultas Ekonomi Bisnis dan Politik	Fakultas Farmasi	…	Pendidikan Ayah	Pendidikan Ibu	Label
1	1	0	…	8	4	Terlambat

Perhitungan algoritma K-Nearest Neighbor menggunakan rumus *Euclidien Distance* seperti pada Pers. (7). Setelah dilakukan perhitungan pada *data training*, selanjutnya melakukan pengurutan jarak dari yang terkecil hingga terbesar untuk menentukan nilai *k* seperti pada Tabel 25. Selanjutnya melakukan *voting* untuk memprediksi seperti yang tertera pada Tabel 26.

Tabel 25 Hasil Perhitungan Jarak

Ranking	*Euclidien Distance*	Label
1	3,162278	Terlambat
3	5,291503	Terlambat
6	5,91608	Terlambat
2	4,795832	Terlambat
9	6,557439	Terlambat
5	5,656854	Tepat
7	6,082763	Tepat
8	6,403124	Tepat
4	5,385165	Tepat

Tabel 26 Hasil Prediksi

Nilai K	*Ranking*	*Euclidien Distance*	Label
K=5	1	3,162278	Terlambat
	2	4,795832	Terlambat
	3	5,291503	Terlambat
	4	5,385165	Tepat
	5	5,656854	Tepat

Data *testing* pada percobaan ini memiliki label “Tepat” akan tetapi pada saat diprediksi dengan K=5 mendapatkan hasil label “Terlambat”. Sehingga prediksi dinyatakan salah karena data *testing* tepat saat diprediksi hasilnya terlambat.

Pada penelitian ini juga akan melakukan perhitungan algoritma K-Nearest Neighbor menggunakan K=5 dengan pembagian data 90:10. Data yang digunakan berjumlah 12.408 di dalamnya terdapat 11.167 data *training* dan 1.241 data *testing*. Adapun data yang digunakan dapat dilihat pada Tabel 11 dan 12. Proses perhitungan menggunakan pemrograman Python seperti Gambar 3.

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import confusion_matrix, accuracy_score
knn = KNeighborsClassifier (n_neighbors=5, metric='euclidean')
knn.fit(x_train, y_train)
y_pred = knn.predict(x_test)
```

Gambar 3 Permodelan Algoritma K-Nearest Neighbor

Hasil dari permodelan pada K=5 mendapatkan 316 *True Positive*, 276 *False Positif*, 304 *False Negative*, dan 345 *True Negative*. Selanjutnya dilakukan proses perhitungan menggunakan *confusion matrix* adapun rumusnya dapat dilihat pada Pers. (8).

Tabel 27 Evaluasi Confusion Matrix

Predict	***Actual***	
	Tepat	**Terlambat**
Tepat	316	276
Terlambat	304	345

Hasil uji coba menggunakan algoritma K-Nearest Neighbor menggunakan pemrograman Python mendapatkan *accuracy* sebesar 53,26%.

3.7 Permodelan Algoritma K-Nearest Neighbor *with Information Gain*

Permodelan algoritma K-Nearest Neighbor *with information gain* akan dilakukan pada pemrograman Python menggunakan K=5 dengan pembagian data 90:10. Data yang digunakan pada permodelan adalah data setelah melalui proses seleksi fitur *Information gain* dan Transformasi. Adapun data dapat dilihat pada Tabel 20 dan 21, di dalamnya terdapat 26 atribut.

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import confusion_matrix, accuracy_score
knn = KNeighborsClassifier (n_neighbors=5, metric='euclidean')
knn.fit(x_train, y_train)
y_pred = knn.predict(x_test)
```

Gambar 4 Permodelan Algoritma K-Nearest Neighbor *with Information Gain*

Hasil dari permodelan pada K=5 mendapatkan 325 *True Positive*, 292 *False Positif*, 319 *False Negative*, dan 305 *True Negative*. Selanjutnya dilakukan proses perhitungan menggunakan confusion matrix adapun rumusnya dapat dilihat pada Pers. (8).

Hasil uji coba menggunakan algoritma K-Nearest Neighbor menggunakan pemrograman Python mendapatkan *accuracy* sebesar 50,76%.

Tabel 28 Evaluasi *Confusion Matrix*

Predict	*Actual*	
	Tepat	**Terlambat**
Tepat	325	292
Terlambat	319	305

3.8 Komparasi

Tabel 29 Hasil Komparasi

Pengujian	Akurasi
Naïve Bayes	51,73%
Naïve Bayes *with Information Gain*	55,19%
K-Nearest Neighbor	53,26%
K-Nearest Neighbor *Information Gain*	50,76%

Berdasarkan pengujian yang telah dilakukan menggunakan 12.408 data seperti pada tabel 29 tentang prediksi keterlambatan biaya kuliah menggunakan fitur *selection information gain* dapat menaikkan hasil akurasi algoritma Naïve Bayes. Hasil akurasi yang diperoleh algoritma Naïve Bayes sebesar 51,73% namun ketika menggunakan fitur seleksi *information gain* akurasi yang diperoleh meningkat sebesar 55,19% lebih tinggi dari pada algoritma K-Nearest Neighbor yang mendapatkan akurasi sebesar 53,26% sedangkan dengan menambahkan fitur seleksi *information gain* akurasi yang diperoleh menurun menjadi 50,76%. Akurasi yang diperoleh cukup rendah, penemuan ini perlu dikaji lebih jauh, dikarenakan secara teori penggunaan dataset yang besar dapat meningkatkan hasil akurasi yang diperoleh dan performa komputasi juga akan lebih baik dalam hal klasifikasi (Widystuti & Darmawan, 2018).

Beberapa penelitian yang berhasil meningkatkan akurasi menggunakan fitur seleksi *information gain* seperti pada penelitian (Setiyorini & Asmono, 2019) tentang penerapan metode KNN pada klasifikasi kinerja siswa yang berhasil meningkatkan akurasi dari 74,068% menjadi 76,553%, data yang digunakan pada penelitian bertipe numerik. Penelitian (Sari, 2017) tentang penerapan algoritma klasifikasi *machine learning* untuk memprediksi performa akademik siswa pada J48 mendapatkan akurasi 90.48%, *random forest* memperoleh 90.05%, MLP mendapatkan akurasi 88.96%, SVM memperoleh 88.1% dan Naïve Bayes memperoleh 86.68%. Data yang digunakan pada penelitian bertipe numerik. Pada penelitian yang dilakukan oleh peneliti fitur seleksi *information gain* mampu menaikkan kinerja algoritma Naïve Bayes akan tetapi tidak memberi pengaruh yang signifikan terhadap kinerja algoritma Naïve Bayes, berbeda halnya dengan penambahan seleksi fitur information gain terhadap algoritma K-Nearest Neighbor yang menurunkan tingkat akurasi yang diperoleh menjadi 50,76, sebelum menggunakan fitur optimasi memperoleh akurasi sebesar 53,26%.

Menurut (Id, 2021) untuk memudahkan algoritma *machine learning* dalam melakukan prediksi salah satunya pendekatan *one-hot encoding* dengan membuat suatu kolom baru dari variabel kategorikal, di mana setiap kategori menjadi kolom baru dengan nilai 0 atau 1 yang di mana nilai 0 menandakan tidak mewakili ada dan 1 mewakili ada. Akan tetapi setelah dilakukan penelitian terdapat beberapa macam gap atau beberapa hasil yang kurang memuaskan seperti pada penelitian (Setiyorini & Asmono, 2019) tentang implementasi kinerja algoritma Naïve Bayes untuk prediksi masa studi mahasiswa memperoleh akurasi yang cukup rendah 68%, termasuk penelitian yang peneliti lakukan mendapatkan tingkat akurasi yang kurang maksimal sebesar 55,19 pada algoritma Naïve Bayes dan 50,76% pada algoritma KNN.

Perbedaan penelitian ini dengan penelitian sebelumnya adalah tipe data yang digunakan, pada penelitian sebelumnya menggunakan tipe data numerik sehingga hasil akurasi yang diperoleh cukup tinggi. Hasil akurasi yang diperoleh dari penelitian ini cukup rendah dikarenakan pada penelitian ini data yang digunakan bersifat kategorikal dan banyaknya atribut menyulitkan algoritma untuk mengambil keputusan, ketika terlalu banyak decision maka mempengaruhi

kinerja algoritma tersebut. pada algoritma Naïve Bayes keputusan yang diambil berdasarkan probabilitas namun jika jenis data banyak bertipe kategori maka tingkat kesulitan dalam mengambil keputusan meningkat.

4. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan maka dapat ditarik beberapa kesimpulan yaitu Atribut keterlambatan biaya kuliah yang terbaik dipilih berdasarkan perhitungan seleksi fitur *informain gain* seperti fakultas, prodi, angkatan dan gender. Hasil pengujian dengan pembagian data 90:10 menggunakan algoritma Naïve Bayes dengan menambahkan fitur seleksi *information gain* mendapatkan akurasi 55,19%, sedangkan pada algoritma K-Nearest Neighbor dengan *information gain* memperoleh akurasi sebesar 50,76%. Dapat disimpulkan bahwa algoritma yang memiliki kinerja terbaik yaitu algoritma Naïve Bayes dan penambahan fitur seleksi *information gain* mampu menaikkan kinerja algoritma Naïve Bayes akan tetapi tidak memberi pengaruh yang signifikan terhadap kinerja algoritma Naïve Bayes, berbeda halnya dengan penambahan seleksi fitur *information gain* terhadap algoritma K-Nearest Neighbor yang menurunkan tingkat akurasi yang diperoleh.

DAFTAR PUSTAKA

Akhmad, M. R., & Siswa, T. A. Y. (2022). Implementasi K-Nearest Neighbor Dalam Memprediksi Keterlambatan Pembayaran Biaya Kuliah Di Perguruan Tinggi. *Progresif: Jurnal Ilmiah Komputer*, *18*(2), 185. https://doi.org/10.35889/progresif.v18i2.921

Ali, H., Mohd Salleh, M. N., Saedudin, R., Hussain, K., & Mushtaq, M. F. (2019). Imbalance class problems in data mining: a review. *Indonesian Journal of Electrical Engineering and Computer Science*, *14*(3), 1552. https://doi.org/10.11591/ijeecs.v14.i3.pp1552-1563

Amelia, M. winny, Lumenta, A. S. ., & Jacobus, A. (2017). Prediksi Masa Studi Mahasiswa dengan Menggunakan Algoritma Naïve Bayes. *Jurnal Teknik Informatika*, *11*(1). https://doi.org/10.35793/jti.11.1.2017.17652

Id, I. D. (2021). *Machine Learning : Teori, Studi Kasus dan Implementasi Menggunakan Python*. UR PRESS. https://doi.org/10.5281/zenodo.5113507

Kinoto, J., Damanik, J. L., Situmorang, E. T. S., Siregar, J., & Harahap, M. (2020). Prediksi Employee Churn Dengan Uplift Modeling Menggunakan Algoritma Logistic Regression. *Jurnal Teknologi Dan Ilmu Komputer Prima (JUTIKOMP)*, *3*(2), 503–508. https://doi.org/10.34012/jutikomp.v3i2.1645

Kurniawan, D. (2020). *Pengenalan Machine Learning dengan Python*. PT Elex Media Komputindo.

Muqorobin, M., Kusrini, K., & Luthfi, E. T. (2019). Optimasi Metode Naive Bayes dengan Feature Selection Information Gain untuk Prediksi Keterlambatan Pembayaran SPP Sekolah. *Jurnal Ilmiah SINUS*, *17*(1), 1. https://doi.org/10.30646/sinus.v17i1.378

Muqorobin, M., Kusrini, K., Rokhmah, S., & Muslihah, I. (2020). Estimation System For Late Payment Of School Tuition Fees. *International Journal of Computer and Information System (IJCIS)*, *1*(1), 1–6. https://doi.org/10.29040/ijcis.v1i1.5

Mustakim, M., & Oktaviani, G. (2016). Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa. *Jurnal Sains, Teknologi, Dan Industri*, *13*(2), 195–202. https://doi.org/10.24014/sitekin.v13i2.1688

Primartha, R. (2021). *Algoritma Machine Learning*. Informatika.

Rahmatullah, S. (2019). Prediksi Tingkat Kelulusan Tepat Waktu dengan Metode Naïve Bayes dan K-Nearest Neighbor. *Jurnal Informasi Dan Komputer*, *7*(1), 7–16. https://doi.org/10.35959/jik.v7i1.118

Rajaraman, A., & Ullman, J. D. (2011). Data Mining. In *Mining of Massive Datasets* (Vol. 2, Issue January 2013, pp. 1–17). Cambridge University Press. https://doi.org/10.1017/CBO9781139058452.002

Rifai, M. F., Jatnika, H., & Valentino, B. (2019). Penerapan Algoritma Naïve Bayes Pada Sistem Prediksi Tingkat Kelulusan Peserta Sertifikasi Microsoft Office Specialist (MOS). *PETIR*, *12*(2), 131–144. https://doi.org/10.33322/petir.v12i2.471

Rohmayani, D. (2020). Analysis Of Student Tuition Fee Pay Delay Prediction Using Naive Bayes

Algorithm With Particle Swarm Optimation Optimazation (Case Study : Politeknik TEDC Bandung). *Jurnal Teknologi Informasi Dan Pendidikan*, *13*(2), 1–8. https://doi.org/10.24036/tip.v13i2.317

Salmu, S., & Solichin, A. (2017). Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naive Bayes: Studi Kasus UIN Syarif Hidayatullah Jakarta. *Seminar Nasional Multidisiplin Ilmu (SENMI)*, 701–709.

Saputro, I. W., & Sari, B. W. (2020). Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa. *Creative Information Technology Journal*, *6*(1), 1. https://doi.org/10.24076/citec.2019v6i1.178

Sari, B. N. (2016). Implementasi Teknik Seleksi Fitur Information Gain Pada Algoritma Klasifikasi Machine Learning Untuk Prediksi Performa Akademik Siswa. *Seminar Nasional Teknologi Informasi Dan Multimedia 2016*, *March*, 55–60.

Setiyorini, T., & Asmono, R. T. (2019). Penerapan Metode K-Nearest Neighbor dan Information Gain pada Klasifikasi Kinerja Siswa. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *5*(1), 7–14. https://doi.org/10.33480/jitk.v5i1.613

Suardika, I. G. I. (2019). Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naive Bayes: Studi Kasus Fakultas Ekonomi dan Bisnis Universitas Pendidikan Nasional. *Jurnal Ilmu Komputer Indonesia*, *4*(2), 37–44. https://doi.org/10.23887/jik.v4i2.2775

Suntoro, J. (2019). *Data Mining Algoritma dan Implementasi dengan Pemrograman PHP*. PT Elex Media Komputindo.

Suyanto. (2017). *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Informatika.

Wanto, A., Siregar, M. N. H., Windarto, A. P., Hartama, D., Ginantra, N. L. W. S. R., Napitupulu, D., Negara, E. S., Lubis, M. R., Dewi, S. V., & Prianto, C. (2020). *Data Mining : Algoritma Klasifikasi*. Yayasan Kita Menulis.

Widaningsih, S. (2019). Perbandingan Metode Data Mining untuk Prediksi Nilai dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika dengan Algoritma C4.5, Naïve Bayes, KNN Dan SVM. *Jurnal Tekno Insentif*, *13*(1), 16–25. https://doi.org/10.36787/jti.v13i1.78

Widystuti, W., & Darmawan, J. B. B. (2018). Pengaruh jumlah data set terhadap akurasi pengenalan dalam deep convolutional network. *Konferensi Nasional Sistem Informasi (KNSI)*, 634–636.