Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa

Naïve Bayes Algorithm Performance Test for Student Study Prediction

Irkham Widhi Saputro*[1], Bety Wulan Sari[2]
[1,2]Program Studi Ilmu Komputer, Jurusan Informatika, Universitas AMIKOM Yogyakarta
E-mail: ***[1]irkham.saputro@students.amikom.ac.id, [2]bety@amikom.ac.id**

Abstrak

Universitas AMIKOM Yogyakarta adalah salah satu perguruan tinggi yang memiliki ribuan mahasiswa baru khususnya pada prodi Informatika. Pada tahun 2012 tercatat ada 1009 mahasiswa baru, dan pada tahun 2013 juga tercatat ada sebanyak 859 mahasiswa baru. Namun sayangnya, dari sekian banyak mahasiswa hanya sekitar 50% saja yang dapat lulus dengan tepat waktu. Data tersebut untuk membuat sistem klasifikasi menggunakan teknik data mining dengan metode Naïve Bayes. Dataset yang akan digunakan sebanyak 300 data yang bersumber dari data alumni angkatan 2012, dan 2013 dengan masing-masing data sebanyak 150. Data yang diperoleh memiliki 144 mahasiswa dengan keterangan lulus tepat waktu, dan 156 mahasiswa dengan keterangan lulus tidak tepat waktu. Proses pengujian akan dilakukan menggunakan metode 10-Fold Cross Validation, dan Confusion Matrix. Hasil pengujian menunjukkan bahwa rata-rata performa dari model Naïve Bayes mempunyai nilai akurasi sebesar 68%, nilai precision sebesar 61.3%, nilai recall sebesar 65.3%, dan nilai f1-score sebesar 61%. Nilai performa dari model dapat dipengaruhi oleh dataset yang digunakan untuk pembuatan model.

Kata Kunci *— data mining, Naïve Bayes, K-Fold Cross Validation, Confusion Matrix*

Abstract

AMIKOM Yogyakarta University is one of the colleges that has thousands of new students, especially in the Informatics study program. In 2012 there were 1009 new students, and in 2013 there were 859 new students. But unfortunately, of the many students only around 50% can graduate on time. The data is to make the classification system using data mining techniques with the Naïve Bayes method. The dataset will be used as much as 300 data sourced from alumni data of 2012, and 2013 with each data as much as 150. The data obtained has 144 students with information passed on time, and 156 students with graduation information not on time. The testing process will be carried out using the 10-Fold Cross Validation, and Confusion Matrix method. The test results show that the average performance of the Naïve Bayes model has an accuracy value of 68%, precision value is 61.3%, recall value is 65.3%, and f1-score is 61%. The performance value of the model can be influenced by the dataset used for modeling.

Keywords *— data mining, classification, Naïve Bayes, graduation time*

1. PENDAHULUAN

1.1. Latar Belakang

Banyaknya jumlah mahasiswa tidak menjamin bagusnya akreditasi suatu prodi. Menurut Badan Akreditasi Nasional Perguruan Tinggi (BAN-PT) pada Peraturan Badan Akreditasi Nasional perguruan Tinggi Nomor 2 Tahun 2017 tentang Sistem Akreditasi Nasional Pendidikan Tinggi, salah satu instrumen akreditasi yang mempengaruhi akreditasi adalah dampak, capaian,

Dikirim: 16 November 2018; Direvisi: 03 Januari 2019, 11 Januari 2019; Diterima: 24 Januari 2019

mutu serta produktivitas luaran yang bermanfaat bagi masyarakat. Dapat diartikan bahwa kualitas lulusan merupakan salah satu aspek yang penting bagi perguruan tinggi dalam memperoleh nilai akreditasi. Universitas AMIKOM Yogyakarta merupakan salah satu perguruan tinggi yang memiliki ribuan mahasiswa baru setiap tahunnya, khususnya pada prodi Informatika. Namun dari sekian banyak mahasiswa baru yang masuk, hanya sekitar 50% yang dapat lulus dengan tepat waktu. Hal ini menjadi permasalahan yang serius bagi pihak perguruan tinggi karena selain tidak dapat memaksimalkan nilai akreditasi, perbandingan antara jumlah dosen, dan jumlah mahasiswa dalam proses belajar mengajar dikelas juga semakin tinggi. Oleh karena itu, pada penelitian ini penulis ingin menerapkan teknik data mining menggunakan metode klasifikasi dengan menerapkan algoritma Naive Bayes untuk memprediksi kelulusan mahasiswa yang telah selesai menempuh tahun ke-2 menggunakan variabel penelitian berupa data induk, dan data historis dari mahasiswa.

Penelitian terkait kelulusan mahasiswa atau penggunaan algoritma Naïve Bayes untuk klasifikasi antara lain adalah penelitian yang dilakukan oleh Arif Jananto [1]. Pada penelitian ini dilakukan terhadap 266 data, dengan 200 data digunakan sebagai data training, dan 66 digunakan sebagai data testing. Hasil yang diperoleh menggunakan metode train test split berupa tingkat kesalahan prediksi antara 20% hingga 34%. Selain penelitian tersebut, penelitian yang dilakukan oleh Supardi Salmu yang menggunakan algoritma Naïve Bayes untuk memprediksi tingkat kelulusan mahasiswa di UIN Syarif Hidayatullah Jakarta [2]. Penelitian ini menggunakan model Cross Industry Standard Process for Data Mining (CRISP-DM), dan menggunakan 12 atribut sebagai prediktornya. Data yang digunakan sebanyak 1162 data sebagai data training dan 587 data sebagai data testing. Hasil yang diperoleh melalui metode pengujian confusion matrix berupa akurasi sebesar 80.72%. Penelitian lainnya dilakukan oleh Riszki Wijayatun Pratiwi yang menggunakan sebanyak 25 variabel bebas yang memiliki hubungan dengan kru pembuatan film, jenis, dan durasi film. Proses klasifikasi dilakukan menggunakan software Rapid Miner. Metode pengujian yang digunakan adalah confusion matrix, dan menghasilkan nilai accuracy, precision, dan recall sebesar 55.80%, 32.41%, dan 46.70% [3].

1.2. Kepribadian dan Gangguan kepribadian Ambang (Borderline)

Data Mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam database, data warehouse, atau penyimpanan informasi lainnya [4]. Data Mining adalah proses yang menggunakan teknik statistic, matematik, kecerdasarn buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai basis data besar [5]. Berdasarkan beberapa definisi tersebut, dapat disimpulkan bahwa Data Mining adalah suatu proses untuk menemukan pola dari tumpukan data yang tersimpan dalam suatu penyimpanan elektronik yang akan digunakan untuk mendapatkan informasi-informasi yang belum diketahui sebelumnya.

Data Mining memiliki suatu rangkaian proses yang harus dilakukan sebelum dapat memperoleh informasi baru. Tahap-tahap dalam data mining adalah sebagai berikut [4]:

a. Data cleaning
   Pembersihan dari merupakan proses menghilangkan *noise* dan data yang tidak konsisten.
b. Data integration
   Proses dimana menggabungkan data dari berbagai macam sumber data. Proses ini dilakukan ketika menggunakan sumber data yang lebih dari satu.
c. Data selection
   Proses menyeleksi data dimana data yang akan digunakan dalam proses *data mining* diambil dan membiarkan data yang tidak digunakan.
d. Data transformation
   Proses mengubah data ke dalam bentuk yang dapat digunakan dalam perhitungan suatu algoritma
e. Data mining
   Proses menemukan pola dari dataset yang digunakan sebagai basis pengetahuan.

f. Pattern evaluation
Merupakan proses menganalisis hasil dari proses mining menggunakan suatu satuan ukur.
g. Knowledge presentation
Merupakan proses untuk menampilkan hasil dari proses *mining*.

1.3. Naïve Bayes

Naïve Bayes merupakan algoritma yang digunakan untuk klasifikasi yang menggunakan teorema bayes dan berasumsi bahwa nilai antar variabel saling bebas (independen) pada suatu nilai *output*. Dalam hal ini, diasumsikan bahwa kehadiran atau ketiadaan dari suatu variabel tertentu tidak berhubungan dengan kehadiran atau ketiadaan dari variabel lainnya. Teorema bayes dapat ditulis menggunakan persamaan 1 [4]:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad (1)$$

Dimana:
$P(A|B)$ = Probabilitas posterior dari A pada kondisi B (*posterior probability*).

$P(B|A)$ = Probabilitas posterior dari B pada kondisi A(*likelihood*).

$P(A)$ = Probabilitas prior dari A (*class prior probability*).

$P(B)$ = Probabilitas prior dari B (*predictor prior probability*).

Proses untuk menghitung probabilitas kelas suatu data dimulai dengan menentukan *likelihood* berdasarkan dataset yang digunakan, menggunakan metode yang sesuai dengan bentuk dari data yang digunakan. *Likelihood* yang diperoleh akan dikalikan dengan probabilitas dari masing-masing kelas. Hasil dari proses tersebut akan digunakan sebagai acuan untuk mengklasifikasi data baru. Pada praktiknya, seringkali P(B) dihiraukan, karena nilai P(B) selalu tetap.

1.4. One Hot Encoding

One Hot Encoding merupakan suatu metode untuk mentransformasi variabel diskret (categorical) ke dalam bentuk binary sehingga dapat bekerja lebih baik dengan algoritma klasifikasi. Beberapa algoritma tidak dapat langsung menggunakan variabel diskret sebagai masukannya, sehingga diperlukan perubahan terhadap variabel diskret tersebut agar dapat digunakan oleh suatu algoritma di dalam proses komputasi. Contoh transformasi dapat dilihat pada Tabel 1 dan Tabel 2.

Tabel 1. Contoh Data Sebelum One Hote Encoding

Sample	Category	Numerical
1	Human	1
2	Human	1
3	Penguin	2
4	Octopus	3
5	Alien	4
6	Octopus	3
7	Alien	4

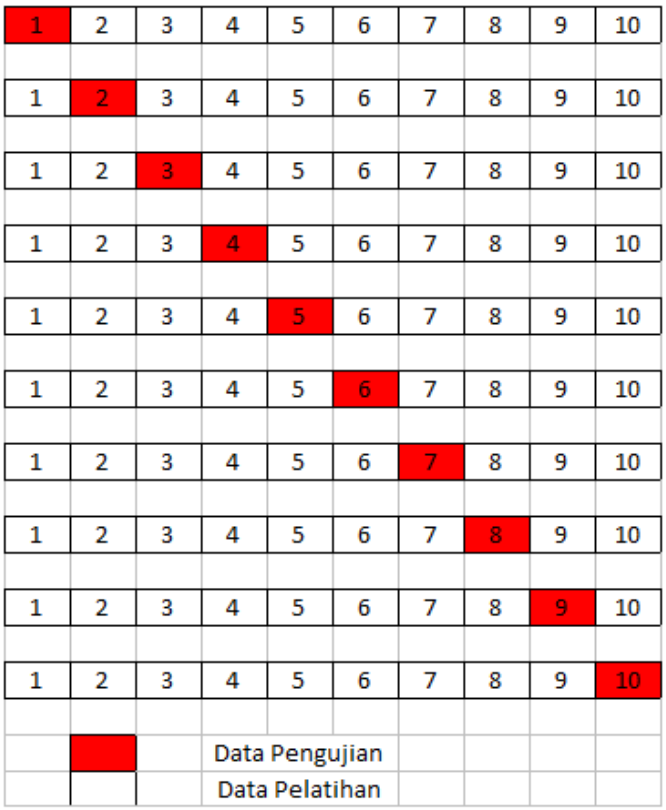
Tabel 2. Contoh Data Sesudah One Hot Encoding

Sample	Human	Penguin	Octopus	Alien
1	1	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1
6	0	0	1	0
7	0	0	0	1

Dari contoh tersebut dapat dilihat bahwa nilai yang terdapat pada variabel category diubah ke dalam bentuk kolom. Kemudian, setiap sampel yang memiliki kemunculan nilai tersebut akan diberi angka satu (1) dan angka nol (0) bagi yang tidak muncul. *One Hot Encoding* hanya dapat merubah data berbentuk diskret, dan menghiraukan data yang bertipe numerik.

1.5. K-Fold Cross Validation

K-Fold Cross Validation merupakan suatu metode untuk membagi data ke dalam beberapa bagian (fold) sebanyak k untuk menentukan data training, dan data testing. Menurut Hastie et al [6], dengan k=5 atau k=10 dapat digunakan untuk memperkirakan tingkat kesalahan yang terjadi, sebab data training pada setiap fold cukup berbeda dengan data training yang asli. Secara keseluruhan, 5 atau 10-fold cross validation sama-sama direkomendasikan dan disepakati bersama. Skema 10 Fold Cross Validation digambarkan pada Gambar 1.



Gambar 1. Skema 10 Fold Cross Validation

1.6. Confusion Matrix

Merupakan tabel yang menggambarkan performa dari sebuah model atau algoritma secara spesifik. Setiap baris dari matrix tersebut, merepresentasikan kelas aktual dari data, dan setiap kolom merepresentasikan kelas prediksi dari data (atau sebaliknya). Matrix tersebut dijelaskan pada Tabel 3.

Tabel 3. Confusion Matrix

	Predicted Negative	**Predicted Positive**
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

1. *True Positive* = Berarti seberapa banyak data yang aktual kelasnya positif, dan model juga memprediksi positif.
2. *True Negative* = Berarti seberapa banyak data yang aktual kelasnya negatif, dan model memprediksi negatif.
3. *False Positive* = Berarti seberapa banyak data yang aktual kelasnya negatif, namun model memprediksi positif.
4. *False Negative* = Berarti seberapa banyak data yang aktual kelasnya positif, namun model memprediksi negatif.

Melalui 4 data tersebut, dapat diperoleh data data lain yang sangat berguna untuk mengukur perfoma sebuah model, diantaranya:

1. *Accuracy* = Total keseluruhan seberapa sering model benar mengklasifikasi. Formula *accuracy* dapat ditulis menggunakan persamaan 2.

$$\frac{TP + TN}{Total} \tag{2}$$

2. *Precision* = Ketika model memprediksi positif, seberapa sering prediksi itu benar. Formula *precision* dapat ditulis menggunakan persamaan 3.

$$\frac{TP}{FP + TP} \tag{3}$$

3. *Recall* (*Sensitivity / True Positive Rate*) = Ketika kelas aktualnya positif, seberapa sering model memprediksi positif. Formula *recall* dapat ditulis menggunakan persamaan 4.

$$\frac{TP}{FN + TP} \tag{4}$$

4. *F1-Score* = Merupakan rata-rata harmonik dari *Precision* dan *Recall*. Formula *f1-score* dapat ditulis menggunakan persamaan 5.

$$2 * \frac{precision * recall}{precision + recall} \tag{5}$$

2. METODE PENELITIAN

Adapun metode penelitian yang digunakan pada penelitian ini adalah sebagai berikut:

1. Analisis Masalah dan Studi Literatur
   Tahap ini merupakan langkah awal yang dilakukan untuk menentukan rumusan masalah dari penelitian, serta memastikan permasalahan yang ada. Selanjutnya dari permasalahan yang ada dianalisa untuk mengetahui bagaimana cara penyelesaian terhadap masalah tersebut dan menentukan ruang lingkup terhadap masalah yang diteliti. Setelah solusi ditetapkan, langkah

selanjutnya adalah mempelajari teori-teori yang berkaitan dengan solusi yang dipilih. Teori dapat diperoleh dari berbagai macam literatur, dan buku-buku nasional, maupun internasional.

2. Mengumpulkan Data
   Prosedur yang digunakan untuk mengumpulkan data adalah dengan mengajukan surat izin kepada pihak Universitas AMIKOM Yogyakarta khususnya bagian Direktorat Innovation Center. Data yang diperoleh berbentuk file excel hasil dari *query database*. Setelah data dikumpulkan, selanjutnya data akan diolah menggunakan metode *Naïve Bayes*.
3. Implementasi dan Pengujian
   Berikut merupakan metode yang akan digunakan dalam proses *data mining*.
   a. Data Cleaning
      Langkah awal dalam proses *data mining* adalah memastikan bahwa data yang diperoleh tidak mengandung nilai kosong atau nilai yang tidak konsisten dengan nilai yang seharusnya. Proses cleaning dilakukan secara manual terhadap 300 data yang dimiliki. Dari 300 data yang dimiliki, tidak ada data yang memiliki nilai kosong (null) ataupun nilai yang tidak konsisten.
   b. Data Selection
      Data *selection* yaitu proses memilih data berdasarkan variabel penelitian. Data yang akan digunakan dalam penelitian adalah data Nomor Induk, tipe sekolah, jenis kelamin, kota sekolah, IP semester 1, IP semester 2, IP semester 3, IP semester 4, IPK, dan keterangan lulus. Nomor induk hanya digunakan sebagai identitas data, dan tidak digunakan dalam klasifikasi. Data IPK juga tidak disertakan karena data IPK akan dihitung pada sistem dengan menghitung rata-rata dari data nilai IP setiap semester. Data yang tidak termasuk dalam variabel penelitian yang digunakan akan dibiarkan tidak terpakai.
   c. Data Transformation
      Data transformation adalah proses dimana data diubah ke dalam bentuk yang dapat diproses oleh algoritma. Terdapat 3 langkah yang akan dilakukan pada proses data transformasi untuk melakukan klasifikasi menggunakan algoritma *Naïve Bayes* yaitu:
      a) Grouping
         Grouping merupakan tahap untuk menggolongkan variabel asal sekolah, dan kota sekolah. Variabel asal sekolah akan digolongkan menjadi 3 yaitu SMA, SMK, dan Lain. Untuk kota sekolah akan digolongkan menjadi 2 jenis, yaitu Dalam Kota untuk kota Yogyakarta, dan Luar Kota untuk selain kota Yogyakarta.
      b) Menghitung IPK
         Untuk mengetahui nilai IPK adalah dengan menghitung rata-rata dari data nilai IP Semester 1, IP Semester 2, IP Semester 3, dan IP Semester 4.
      c) Discretization
         Discretization merupakan tahap mengubah data berbentuk kontinu ke dalam bentuk diskret. Seluruh data nilai yaitu IP Semester 1, IP Semester 2, IP Semester 3, IP Semester 4, dan IPK akan diubah ke dalam bentuk diskret (A, B, C, D, dan E). Untuk aturan konversinya, diambil berdasarkan Buku Panduan Akademik Tahun 2016-2017 yang dapat diunduh pada laman website AMIKOM. Aturan untuk mengkonversi nilai angka ke dalam nilai huruf digambarkan pada Tabel 4.

Tabel 4. Aturan Konversi Nilai

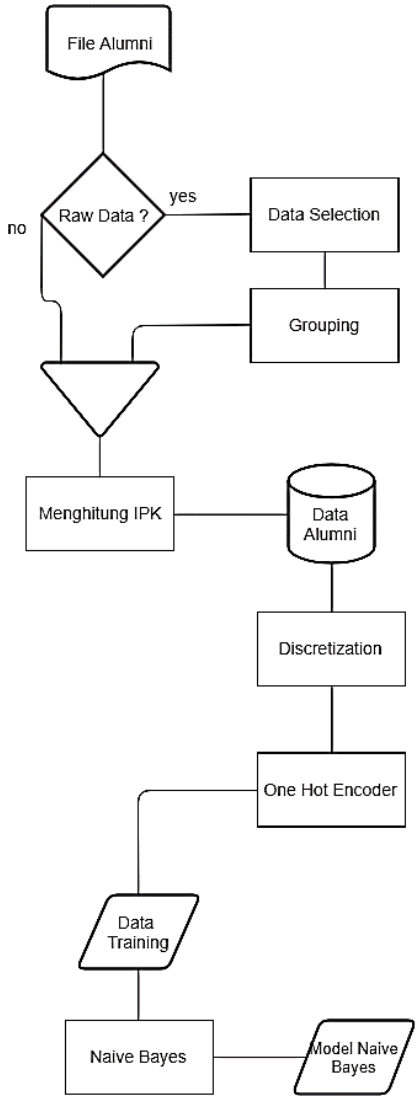
Nilai Huruf	Nilai Angka
A	≥ 3.5
B	≥ 3
C	≥ 2.5
D	≥ 2
E	< 2

d. One Hot Encoding
Tahap ini merupakan proses mengubah data alumni menjadi *binary* data. Masing-masing nilai unik pada suatu variabel akan digunakan sebagai variabel baru yang disebut *dummy variable*. Bilangan binary menunjukkan kemunculan nilai dari suatu data. Sebagai contoh, apabila suatu data memiliki nilai SMA pada variabel tipe sekolah, maka pada *dummy variable* tipe_sekolah_SMA akan memiliki nilai 1 dan memiliki nilai 0 pada *dummy variable* tipe_sekolah_SMK, dan seterusnya seperti yang terlihat pada Gambar 2.

ket_lulus	school_type_Lain	school_type_SMA	school_type_SMK	gender_L	gender_P	school_city_Dalam Kota	school_city_Luar Kota	...
0	0	0	1	1	0	0	1	...
0	0	1	0	1	0	0	1	...
0	1	0	0	1	0	0	1	...
0	0	0	1	0	1	0	1	...
0	0	1	0	1	0	0	1	...
0	0	0	1	1	0	0	1	...
0	0	0	1	1	0	0	1	...
0	0	0	1	1	0	0	1	...
0	0	0	1	1	0	0	1	...
0	0	0	1	1	0	0	1	...
0	0	1	0	1	0	0	1	...
0	0	1	0	1	0	0	1	...
0	1	0	0	1	0	1	0	...
0	0	1	0	1	0	1	0	...
0	0	0	1	0	1	1	0	...
0	0	1	0	1	0	0	1	...
0	0	1	0	1	0	0	1	...
0	0	1	0	1	0	0	1	...
0	0	1	0	1	0	0	1	...
0	0	1	0	1	0	0	1	...
0	0	0	1	1	0	0	1	...
0	0	1	0	1	0	0	1	...

Gambar 2. Data Alumni Hasil One Hot Encoding

e. Implementasi Naïve Bayes
Tahap ini merupakan tahap implementasi dari metode yang telah diuraikan sebelumnya. Pertama akan dilakukan proses pembersihan data secara manual, kemudian memilih data, mengubah data ke dalam bentuk diskret, dan mengubah data ke dalam bentuk *binary* menggunakan metode *One Hot Encoding* untuk dapat diimplementasikan ke dalam proses komputasi. Proses ini dapat disebut juga proses *training*, dimana model *Naïve Bayes* akan dibuat menggunakan 300 dataset yang kemudian akan langsung diuji menggunakan metode *K-Fold Cross Validation*, dan *Confusion Matrix* untuk diperoleh performanya. Hasil dari tahap ini berupa sebuah model *Naïve Bayes* yang sudah dapat digunakan untuk melakukan klasifikasi. Alur dari proses *training* model *Naïve Bayes* dirangkum pada Gambar 3.



Gambar 3. Proses Training Data

3. HASIL DAN PEMBAHASAN

Berdasarkan hasil pengujian dari model yang telah dibuat menggunakan metode *10-Fold Cross Validation*, dan *Confusion Matrix* diperoleh nilai akurasi, *precision, recall*, dan *f1-score* pada masing-masing fold yang dirangkum paada Tabel 5.

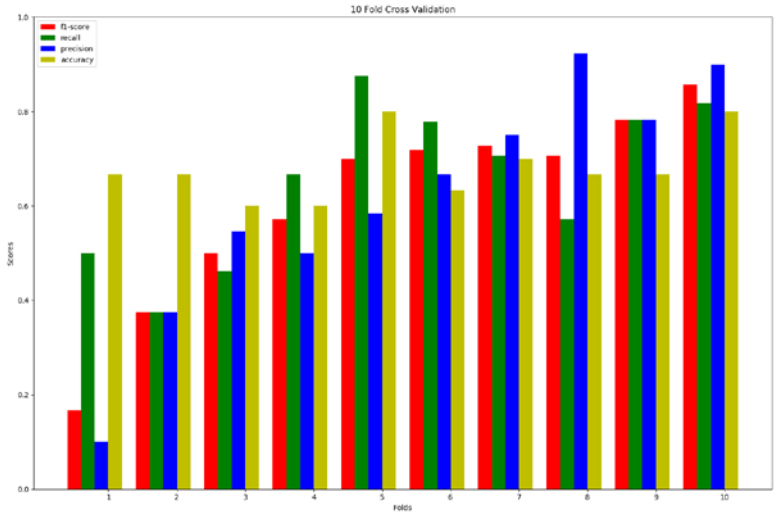
Tabel 5. Hasil 10 Fold Cross Validation

Fold	Parameter	Hasil (%)
Fold 1	Accuracy	66
	Precision	10
	Recall	50
	F1-Score	16
	Jumlah Data	30
Fold 2	Accuracy	66
	Precision	37.5
	Recall	37.5

	F1-Score	37.5
	Jumlah Data	30
Fold 3	Accuracy	60
	Precision	54.5
	Recall	46.2
	F1-Score	50
	Jumlah Data	30
Fold 4	Accuracy	60
	Precision	50
	Recall	66.7
	F1-Score	57.1
	Jumlah Data	30
Fold 5	Accuracy	80
	Precision	58.3
	Recall	87.5
	F1-Score	70
	Jumlah Data	30
Fold 6	Accuracy	63.3
	Precision	66.7
	Recall	77.8
	F1-Score	71.8
	Jumlah Data	30
Fold 7	Accuracy	70
	Precision	75
	Recall	70.6
	F1-Score	72.7
	Jumlah Data	30
Fold 8	Accuracy	66.7
	Precision	92.3
	Recall	57.1
	F1-Score	70.6
	Jumlah Data	30
Fold 9	Accuracy	66.7
	Precision	78.3
	Recall	78.3
	F1-Score	78.3
	Jumlah Data	30
Fold 10	Accuracy	80
	Precision	90
	Recall	81.8
	F1-Score	85.7
	Jumlah Data	30
Rata-Rata	Accuracy	68
	Precision	61.3
	Recall	65.3
	F1-Score	61
	Jumlah Data	300

Dari Tabel 5, akurasi tertinggi didapatkan pada fold ke-5 dan fold ke-10 dengan nilai 80%. Pada fold ke-5, nilai *precision, recall,* dan *f1-score* nya adalah 58.3%, 87.5%, dan 70%, sedangkan pada fold ke-10 memiliki nilai 90%, 81.8%, dan 85.7%. Adapun hasil terendah yang

dihasilkan oleh model ini adalah 60% yang diperoleh pada fold ke-3, dan fold ke-4. Pada fold ke-3 memiliki nilai *precision*, *recall*, dan *f1-score* sebesar 54.5%, 46.2%, 50%, sedangkan pada fold ke-4 memiliki nilai sebesar 50%, 66.7%, dan 57.1%. Rata-rata performa yang diperoleh dari hasil *confusion matrix* pada 10-*Fold Cross Validation* memiliki nilai *accuracy*, *precision*, *recall*, dan *f1-score* 68%, 61.3%, 65.3%, dan 61%. Untuk lebih mempermudah dalam pembacaan data pada Tabel 3 data dapat dibuat ke dalam bentuk grafik bar chart yang digambarkan pada Gambar 4.



Gambar 4. Grafik Performa Model Hasil 10 Fold Cross Validation

Dari total keseluruhan fold, diperoleh pula nilai hasil *confusion matrix* yang mencatat hasil dari klasifikasi model. *Confusion matrix* dapat digunakan untuk mengetahui letak kesalahan klasifikasi dari model Naïve Bayes. Tabel 6 menjelaskan *confusion matrix* yang diperoleh dari hasil evaluasi model Naïve Bayes.

Tabel 6. Confusion Matrix Naïve Bayes

	Predicted Tidak Tepat Waktu	**Predicted Tepat Waktu**	**Jumlah**
Actual Tidak Tepat Waktu	105	51	156
Actual Tepat Waktu	45	99	144
Jumlah	150	150	300

Pada Tabel 6 terdapat 150 data yang diprediksi tidak tepat waktu, dan 150 data yang diprediksi tepat waktu. Namun dari 150 data yang diprediksi tidak tepat waktu, terdapat kesalahan prediksi sebanyak 45 data. Data ini seharusnya mempunyai kelas tepat waktu, namun model memprediksi tidak tepat waktu. Untuk 150 data yang diprediksi tepat waktu, terdapat 51 data yang salah diprediksi. Data ini seharusnya mempunyai kelas tidak tepat waktu, namun diprediksi tepat waktu oleh model. Tabel 6 merupakan total penjumlahan dari seluruh fold yang ada. Nilai *precision*, *recall*, dan *f1-score* diperoleh dari tabel *confusion matrix* yang terbentuk pada masing-masing fold.

4. KESIMPULAN

Berdasarkan hasil percobaan yang telah dilakukan maka dapat diambil kesimpulan sebagai berikut:

1. Penelitian terhadap 300 data alumni menggunakan Algoritma Naïve Bayes yang digunakan untuk klasifikasi waktu kelulusan mahasiswa menghasilkan model klasifikasi dengan rata-rata nilai akurasi, *precision, recall,* dan *f1-score* sebesar 68%, 61.3%, 65.3%, dan 61% yang dihitung menggunakan metode 10-*Fold Cross Validaiton*, dan *Confusion Matrix*.
2. Penentuan data *training* yang digunakan dapat mempengaruhi hasil pengujian, karena probabilitas yang dimiliki oleh model akan digunakan untuk menentukan kelas pada data *testing*, sehingga besar kecilnya nilai akurasi, *precision, recall,* dan *f1-score* juga dipengaruhi oleh penentuan data *training*.

5. SARAN

Berdasarkan kesimpulan dari penelitian, maka penulis ingin memberikan saran sebagai berikut:

1. Menggunakan dataset dengan jumlah yang lebih banyak agar pola yang didapat oleh model lebih bervariasi.
2. Menggunakan algoritma atau metode pengujian yang lain, seperti algoritma c4.5, algoritma forward chaining, algoritma backward chaining, sehingga penelitian ini dapat digunakan sebagai pembanding.
3. Menggunakan gabungan beberapa algoritma atau metode pengujian yang berbeda seperti ROC Curve sehingga dapat diperoleh hasil performa yang lebih baik.

DAFTAR PUSTAKA

[1] Jananto, A., 2013, Algoritma Naïve Bayes untuk Mencari Perkiraan Waktu Studi Mahasiswa, *Jurnal Teknologi Informasi DINAMIK*, No. 1, Vol. 18, Hal. 9 – 16.

[2] Salmu, S., Solichin, A., 2017, Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naïve Bayes: Studi Kasus UIN Syarif Hidayatullah Jakarta, *Seminar Nasional Multidisiplin Ilmu 2017*, Jakarta, 27 April.

[3] Pratiwi, R. W., Nugroho, Y. S., 2016, Prediksi Rating Film Menggunakan Metode Naïve Bayes, *Jurnal Teknik Elektro*, No. 2, Vol. 8, Hal. 60 – 63.

[4] Han, J., Kamber, M., Pei, J., 2012, *Data Mining Concepts and Techniques 3rd Edition*, Morgan Kauffman, San Fransisco.

[5] Santosa, B., 2007, *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*, Graha Ilmu, Yogyakarta.

[6] Hastie, T., Tibshirani, R., Friedman, J., 2009, *The Elements of Statistical Learning Data Mining, Inference, Prediction 2nd Edition*, Springer-Verlag, New York.