Jurnal Mobile and Forensics (MF)
Vol. 2, No. 1, Maret 2020, pp. 41-50
P-ISSN: 2656-6257, E-ISSN: 2714-6685, DOI: https://doi.org/10.12928/mf.v2i1.2137

PREDIKSI LAMA STUDI MENGGUNAKAN NAÏVE BAYES BERDASARKAN ASPEK SOSIAL EKONOMI MAHASISWA

[1,*]Desy Pratiwi Ika Putri, [2]Desi Anggreani, [3]Aji Prasetya Wibawa
[123]Universitas Negeri Malang, Jl. Semarang No. 5, (0341) 551312
e-mail: [1]desypratiwi407@gmail.com, [2]desi.anggreanis12@gmail.com, [1]aji.prasetya.ft@um.ac.id
*corresponding email

Abstrak

Perguruan tinggi merupakan satuan penyelenggara pendidikan tinggi sebagai tingkat lanjut jenjang pendidikan menengah di jalur pendidikan formal. Kualitas perguruan tinggi, khususnya perguruan tinggi di Indonesia diukur berdasarkan 9 standar utama. Salah satu aspek yang berpengaruh ialah mahasiswa dan lulusan. Ketepatan waktu studi mahasiswa adalah hal yang penting dalam perguruan tinggi. Ketepatan waktu mahasiswa dalam menyelesaikan studi menjadi salah satu penunjang penilaian kualitas perguruan tinggi. Metode Naïve Bayes dapat digunakan untuk memprediksi ketepatan lama studi. Klasifikasi Naïve Bayes dalam penelitian ini menggunakan beberapa variabel yang sangat erat kaitannya dalam menyelesaikan studi khususnya pada aspek sosial ekonomi mahasiswa. Adapun variable dari sisi sosial dan ekonomi tersebut diantaranya jenis kelamin, nilai IPK, tempat lahir, tipe sekolah, jumlah keikutsertaan organisasi, tingkat ekonomi, dan dukungan orang tua. Pada penelitian ini, metode Naïve Bayes diimplementasikan pada kasus prediksi lama studi mahasiswa menggunakan 200 data set. Hasil penelitian menunjukkan tingkat rata-rata akurasi sebesar 80,5% dengan menggunakan *K-Fold Cross Validation* diperoleh standar deviasi 3,02%.

Higher education is a higher education provider unit as an advanced level of secondary education in the formal education pathway. The quality of tertiary institutions, especially tertiary institutions in Indonesia, is measured according to 9 main standards. One influential aspect is students and graduates. Timeliness of student studies is important in higher education. Timeliness of students in completing their studies is one of the supports for assessing the quality of higher education. The Naïve Bayes method can be used to predict the accuracy of the study duration. Naïve Bayes classification in this study uses several variables that are very closely related in completing studies, especially on the social economic aspects of students. The social and economic variables include gender, GPA, birthplace, type of school, number of organizational participations, economic level, and parent support. In this study, the Naïve Bayes method is implemented in the case of prediction of student study duration using 200 data sets. The results showed an average level of accuracy of 80.5% using K-Fold Cross Validation obtained a standard deviation of 3.02%.

Kata Kunci: Prediksi Lama Studi, Data Mining,K-fold Cross Validation Naïve Bayes, Mahasiswa

PENDAHULUAN

Perguruan tinggi merupakan satuan penyelenggara pendidikan tinggi sebagai tingkat lanjut jenjang pendidikan menengah di jalur pendidikan formal. Perguruan tinggi sebagai lembaga yang melaksanakan fungsi Tridharma Perguruan Tinggi, yaitu pendidikan, penelitian dan pengabdian kepada masyarakat, serta mengelola ipteks harus mampu mengatur diri sendiri dalam upaya meningkatkan dan menjamin mutu secara terus menerus[1].

Kualitas perguruan tinggi, khususnya perguruan tinggi di Indonesia diukur berdasarkan akreditasi yang dilakukan oleh Badan Akreditasi Nasional Perguruan Tinggi atau BAN-PT. Kualitas diukur berdasarkan 9 standar utama, salah satunya ialah mahasiswa dan lulusan[2]. Mahasiswa adalah salah satu aspek penting dalam

Received Januari 19, 2020; Revised Februari 3, 2020; Accepted Maret 27, 2020

menunjang kesuksesan program studi. Ketepatan waktu mahasiswa dalam menyelesaikan studi menjadi salah satu penunjang penilaian kualitas perguruan tinggi. Banyak faktor yang mempengaruhi ketepatan waktu studi mahasiswa di antaranya dari sisi sosial, ekonomi, akademik, dan lainnya. Adapun faktor tersebut dijabarkan berupa dukungan orang tua, jumlah kegiatan non akademik yang diikuti, tingkat ekonomi mahasiswa, Indeks Prestasi Kumulatif (IPK), jumlah mata kuliah yang diambil[3]. Namun, beberapa faktor tersebut sangat kurang disadari sehingga mempengaruhi lama studi mahasiswa.

Data mining digunakan sebagai bahan untuk menganalisis dan pengambilan keputusan atas faktor faktor yang mempengaruhi ketepatan waktu studi mahasiswa. Data mining adalah proses ekstraksi informasi untuk memperoleh pengetahuan dan menemukan pola pada tumpukan data dalam database berskala besar[4]. Klasifikasi Naïve Bayes telah banyak digunakan sebagasi salah satu metode untuk memprediksi ketepatan waktu studi mahasiswa. Jurnal "Predicting Patterns of Student Graduation Rates Using Naïve Bayes Classifier and Support Vector Machine" menggunakan 7 variabel dan meghasilkan tingkat akurasi klasifikasi tertinggi 69,51%[5]. Pada penelitian "Intelegent Computing System to Predict Vocational High School Student Learning Achievement Using Naïve Bayes Algorithm" menggunakan 5 kategori menghasilkan nilai akurasi tertinggi 83% dan terendah 68%[6]. [7]; [8]; dan [9] menggunakan algoritma Naïve Bayes dengan jumlah data, jumlah dan jenis kategori yang berbeda-beda menunjukkan bahwa algoritma ini dapat digunakan dalam memprediksi kelulusan siswa dengan tingkat akurasi diatas 70%.

Lolo dan Ricambi, menggunakan algoritma Naïve Bayes untuk memprediksi minat lulusan SMA yang ingin melanjutkan studi ke Universitas menggunakan sosial media Twitter[10]. Berbeda dengan Wang yang menggabungkan RFM dan Naïve Bayes dalam klasifikasi loyalitas pelanggan untuk pengambilan keputusan[11]. Klasifikasi Naïve Bayes membuktikan dampak variabel-variabel yang digunakan dalam prediksi lama studi mahasiswa[12]. [13] melakukan penelitian prediksi lama studi mahasiswa menggunakan Naïve Bayes dengan 4 variabel dan 50 data. Penelitian tersebut menghasilkan tingkat akurasi 60%.

Berdasarkan permasalahan yang telah dikemukakan, peneliti tertarik mengangkat judul penelitian yang berkaitan dengan klasifikasi Naïve Bayes dalam memprediksi lama studi mahasiswa dengan menggunakan variable yang diambil dari sisi faktor sosial dan ekonomi mahasiswa dengan data berskala besar.

METODE PENELITIAN

Metode Naïve Bayes

Naïve Bayes menghitung serangkaian probabilitas dengan menjumlahkan kombinasi frekuensi dan nilai *dataset* yang diberikan[14]. Algoritma ini mengasumsikan semua atribut independen atau tidak adanya hubungan karakteristik tertentu dari suatu kelas dengan karakterisktik kelas lainnya[15]. Penggunaan Naïve Bayes membutuhkan jumlah data pelatihan (*Training* Data) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian dan dapat bekerja jauh lebih baik dalam situasi dunia nyata yang kompleks[16]. Pada proses klasifikasi, selain *training data* terdapat satu jenis data yang digunakan yaitu *testing data*. *Training data* dan *testing data* saling mempengaruhi satu sama lain dimana *training data* digunakan sebagai acuan dalam menghitung nilai probabilitas *testing data* [13]. *Training data* atau data set digunakan untuk melatih algoritma klasifikasi [17].

Bentuk umum teorema Bayes dapat dilihat pada persamaan (1) berikut [18]:

$$P(H \mid X) = \frac{P(X \mid H)\, P(H)}{P(X)} \quad (1)$$

Dimana X adalah data dengan kelas yang tidak diketahui, H adalah data hipokripsi X yang merupakan kelas spesifik, P(H|X) adalah probabilitas dari hipokripsi H berdasarkan kondisi X (probabilitas posteriori), P(H) adalah probabilitas hipokripsi H (probabilitas sebelumnya), P(X|H) adalah probabilitas X berdasarkan kondisi H, dan P(X) adalah probabilitas dari X. Perhitungan peluang dengan menggunakan persamaan umum berikut,

$$P(A) = \frac{n(A)}{n(S)} = \frac{\sum Kejadian\ A}{\sum Ruang\ Sampel\ S} \quad (2)$$

Berikut ini *pseudecode* dari metode Naïve Bayes:

Input: *Dataset* Latih P(H), X = Variabel yang digunakan dalam prediksi (IPK, Jenis Kelamin, Tempat lahir, Tipe Sekolah, Jumlah Organisasi, Tingkat Ekonomi, Dukungan Orang tua)

Output : Hasil klasifikasi atau prediksi (Lama studi = Tepat Waktu atau Lama Studi = Tidak Tepat Waktu)

Proses :

1. Membaca data *training* yaitu P(H)
2. Menjumlahkan variabel X pada data *training*
3. Melakukan perulangan probabilitas P (x_1 | H), P (x_2 | H), …, P (x_n | H) hingga mencapai pada Xn
4. Melakukan perhitungan probabilitas akhir dengan P(x | H) = P(x | H) • P(H) di setiap kelas
5. Mendapatkan variabel H selaku kemungkinan terbaik yang menjadi hasil presiksi

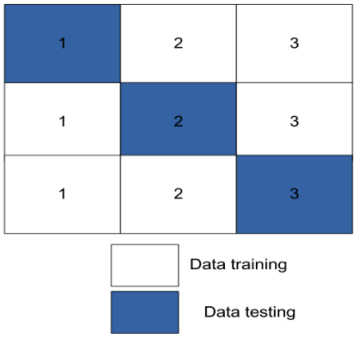
Pengujian Akurasi

Pengukuran performa klasifikasi dilakukan dengan membandingkan seluruh data uji yang diklasifikasikan benar dengan banyaknya data uji. Persamaan (3) dapat digunakan sebagai model yang digunakan untuk mengukur kinerja.

$$Akurasi = \frac{\sum Klasifikasi\ Benar}{\sum Data\ Uji} \times 100\% \quad (3)$$

K-FOLD CROSS VALIDATION

Cross-validation atau estimasi rotasi merupakan pendekatan umum untuk pemilihan model dalam pengembangan data. Teknik ini digunakan untuk membandingkan jumlah besar model yang sesuai dan untuk mencegah *overfitting*, situasi di mana model terlalu disesuaikan dengan *dataset* empiris sehingga tidak mungkin untuk mereplikasi dalam *dataset* baru [19]. Teknik ini membagi data menjadi *k* bagian (*folds*), menggunakan satu bagian untuk pengujian dan sisanya (*k-1 folds*) untuk pemasangan model. Proses iterasi estimasi rotasi *k-folds* dilakukan sampai semua *folds* digunakan untuk pengujian [20].



Gambar 1. Mode 3-*fold cross validation*

Pada gambar 1 merupakan penggunaan 3-*fold cross validation*. Dimana setiap data akan dieksekusi sebanyak 3 kali dan setiap subset data akan mempunyai kesempatansebagai data *testing* atau data *training*. Model pengujian seperti berikut dengan diasumsikan nama setiap pembagian data yaitu A1, A2, dan A3:

1. Percobaan pertama data A1 sebagai data *testing* sedangkan data A2 dan A3 sebagai data *training*.
2. Percobaan kedua data A2 sebagai data *testing* sedangkan data A1 dan A3 sebagai data *training*.
3. Percobaan ketiga atau percobaan terakhir data A3 sebagai data *testing* sedangkan data A1 dan A2 sebagai data *training*.

Simpangan baku adalah ukuran penyebaran data yang menunjukkan jarak rata-rata dari nilai tengah ke suatu titik nilai. Jika simpangan baku yang dihasilkan semakin besar, maka penyebaran dari nilai tengah juga besar, begitu pula sebaliknya. Tujuan menghitung simpangan baku dalam penelitian ini untuk melihat jarak antara rata-rata akurasi dengan akurasi setiap percobaan[21]. Simpangan baku dapat dihitung menggunakan persamaan (4).

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}} \quad (4)$$

Dimana N mewakili banyaknya percobaan. μ adalah rata-rata (*mean*); X_i mewakili percobaan ke-*i*.

HASIL DAN PEMBAHASAN

Dataset

Dalam penelitian ini digunakan 200 jumlah data mahasiswa. Dalam proses mining, untuk memudahkan pengujian maka dilakukan konversi data kedalam bentuk yang sederhana sehingga dapat diolah oleh alat bantu data *mining*. Pada penelitian ini menggunakan 7 variabel. Variabel-variabel tersebut diambil dari sisi aspek sosial dan ekonomi yang berkaitan dengan mahasiswa itu sendiri. Berikut hasil konversi data.

1. Jenis Kelamin
   Cara berfikir antara laki-laki dan perempuan memiliki perbedaan[22], sehingga jenis kelamin memiliki dampak pada pendidikan. Pada Variabel jenis kelamin tidak dilakukan konversi karena jenis kelamin terbagi dua yaitu perempuan dan laki-laki.
2. Indeks Prestasi Komulatif (IPK)

Indeks Prestasi Komulatif (IPK) memiliki rentang nilai antara 0.00 sampai 4.00 sehingga dilakukan konversi yang ditunjukkan pada table 1 berikut.

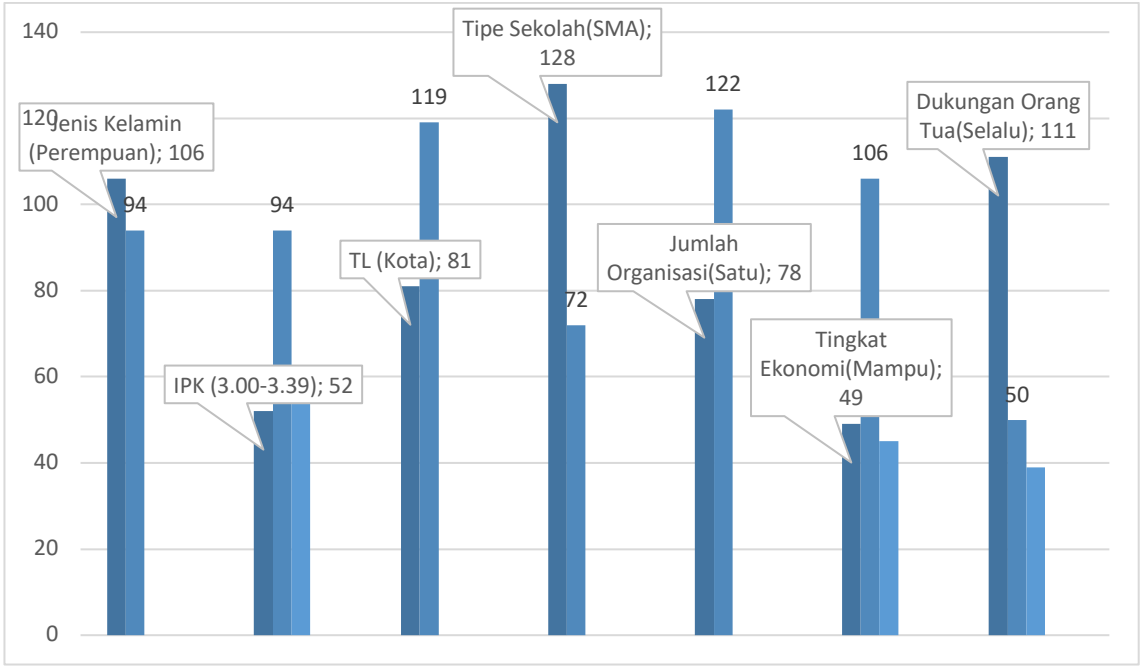
Tabel 1. Konversi Nilai IPK

IPK	IPK Konversi
$3.00 \leq IPK \leq 3.39$	1
$3.40 \leq IPK \leq 3.69$	2
$3.70 \leq IPK \leq 4.00$	3

3. Tempat Lahir
   Tempat lahir setiap mahasiswa berbeda-beda dan sangat berpengaruh pada pendidikan. Ras dan budaya yang membentuk kebiasaan mahasiswa berdasarkan tempat lahir[23]. Tempat lahir dikonversi menjadi dua kategori yaitu Kota dan Desa.
4. Tipe Sekolah
   Variabel tipe sekolah merupakan tipe sekolah sebelum menginjak bangku perkuliahan yang dibagai menjadi dua kategori, yaitu Sekolah Menengah Atas (SMA) dan Sekolah Menengah Kejuruan (SMK).
5. Jumlah Organisasi
   Dalam dunia perkuliahan, organisasi menjadi point penting bagi mahasiswa sebagai salah satu tempat dalam membangun karakter. Keikutsertaan dalam berorganisasi memiliki kekurangan dan kelebihan. Kelebihan mengikuti organisasi dapat meningkatkan percaya diri, sedangkan kekurangannya sulit membagi waktu antara kuliah dan organisasi[24]. Variabel jumlah organisasi dikonversi menjadi dua kategori yaitu Satu dan Lebih dari satu.
6. Tingkat Ekonomi
   Tingkat ekonomi adalah masalah dalam Pendidikan[25]. Tingkat ekonomi dalam hal ini adalah pendapatan orang tua yang dikonversi menjadi tiga kategori yaitu Kurang Mampu, Menengah, dan Mampu.
7. Dukungan Orangtua
   Dukungan orang tua yang dimaksudkan dalam penelitian ini adalah seberapa sering orang tua memantau aktivitas atau perkembangan pendidikan mahasiswa. Dukungan orang tua dikategorikan menjadi tiga, yaitu Tidak Pernah, Jarang dan Selalu.

Dataset yang digunakan untuk analisis diambil dari lulusan Universitas di Kota Makassar berdasarkan pengisian form. Form tersebut berisi variabel Jenis Kelamin (JK), IPK, Tempat Lahir (TL), Tipe Sekolah (TS), Jumlah Organisasi (JO), Tingkat Ekonomi (TE) dan Dukungan Orang Tua (DOT). Pada Gambar 2 menunjukkan data mahasiswa yang digunakan dalam penelitian ini.

Pada Gambar 2 terdapat sebuah data yang menjadi data testing. Data tersebut mengacu pada data *training* yang telah dijelaskan diatas, yaitu dengan inisial D jenis kelamin perempuan, ipk 3.91, tempat lahir kota, tipe sekolah sma, jumlah organisasi lebih dari satu, tingkat ekonomi menengah, dukungan orang tua selalu.



Gambar 2. Data *Training*

Dari hasil *dataset* diketahui bahwa jumlah data latih dengan lama studi "Tepat Waktu" sebanyak 104 dan lama studi "Tidak Tepat Waktu" sebanyak 96.

$$P(\text{Lama Studi} = \text{"Tepat Waktu"}) = 104$$

$$P(\text{Lama Studi} = \text{"Tidak Tepat Waktu"}) = 96$$

maka nilai probabilitas setiap *variable* X terhadap *variable* H P(H|X) sebagai berikut:

$$P(JK = \text{P}|Lama\ Studi = \text{"}Tepat\ Waktu\text{"}) = \frac{\sum Jenis\ Kelamin\ P(Tepat\ Waktu)}{\sum Lama\ Studi\ Tepat\ Waktu} = \frac{35}{104}$$

$$P(JK = \text{P}|Lama\ Studi = \text{"}Tidak\ Tepat\ Waktu\text{"}) = \frac{\sum Jenis\ Kelamin\ P(Tidak\ Tepat\ Waktu)}{\sum Lama\ Studi\ Tidak\ Tepat\ Waktu} = \frac{19}{96}$$

$$P(IPK = 3|Lama\ Studi = \text{"}Tepat\ Waktu\text{"}) = \frac{\sum IPK\ 3(Tepat\ Waktu)}{\sum Lama\ Studi\ Tepat\ Waktu} = \frac{59}{104}$$

$$P(IPK = 3|Lama\ Studi = \text{"}Tidak\ Tepat\ Waktu\text{"}) = \frac{\sum IPK\ 3(Tidak\ Tepat\ Waktu)}{\sum Lama\ Studi\ Tidak\ Tepat\ Waktu} = \frac{35}{96}$$

$$P(TL = \text{Kota}|Lama\ Studi = \text{"}Tepat\ Waktu\text{"}) = \frac{\sum TL\ Kota(Tepat\ Waktu)}{\sum Lama\ Studi\ Tepat\ Waktu} = \frac{37}{104}$$

$$P(TL = Kota|Lama\ Studi = \text{"}Tidak\ Tepat\ Waktu\text{"}) = \frac{\sum TL\ Kota(Tidak\ Tepat\ Waktu)}{\sum Lama\ Studi\ Tidak\ Tepat\ Waktu} = \frac{44}{96}$$

$$P(TS = \text{SMA}|Lama\ Studi = \text{"}Tepat\ Waktu\text{"}) = \frac{\sum TS\ SMA(Tepat\ Waktu)}{\sum Lama\ Studi\ Tepat\ Waktu} = \frac{77}{104}$$

$$P(TS = SMA|Lama\ Studi = \text{"}Tidak\ Tepat\ Waktu\text{"}) = \frac{\sum TS\ SMA(Tidak\ Tepat\ Waktu)}{\sum Lama\ Studi\ Tidak\ Tepat\ Waktu} = \frac{51}{96}$$

$$P(JO = \text{Lebih Satu}|Lama\ Studi = \text{"}Tepat\ Waktu\text{"}) = \frac{\sum JO\ Lebih\ Satu(Tepat\ Waktu)}{\sum Lama\ Studi\ Tepat\ Waktu} = \frac{55}{104}$$

$$P(JO = Lebih\ Satu|Lama\ Studi = \text{"}Tidak\ Tepat\ Waktu\text{"}) = \frac{\sum JO\ Lebih\ Satu(Tidak\ Tepat\ Waktu)}{\sum Lama\ Studi\ Tidak\ Tepat\ Waktu} = \frac{66}{96}$$

$$P(TE = \text{Menengah}|Lama\ Studi = "Tepat\ Waktu") = \frac{\sum TE\ Menengah(Tepat\ Waktu)}{\sum Lama\ Studi\ Tepat\ Waktu} = \frac{56}{104}$$

$$P(TE = Menengah|Lama\ Studi = "Tidak\ Tepat\ Waktu") = \frac{\sum TE\ Menengah(Tidak\ Tepat\ Waktu)}{\sum Lama\ Studi\ Tidak\ Tepat\ Waktu} = \frac{50}{96}$$

$$P(DOT = \text{Selalu}|Lama\ Studi = "Tepat\ Waktu") = \frac{\sum DOT\ Selalu(Tepat\ Waktu)}{\sum Lama\ Studi\ Tepat\ Waktu} = \frac{63}{104}$$

$$P(DOT = Selalu|Lama\ Studi = "Tidak\ Tepat\ Waktu") = \frac{\sum DOT\ Selalu(Tidak\ Tepat\ Waktu)}{\sum Lama\ Studi\ Tidak\ Tepat\ Waktu} = \frac{48}{96}$$

Setelah nilai probabilitas dari setiap variabel X didapatkan, kemudian nilai probabilitas dari setiap variabel X dengan kondisi lama studi "Tepat Waktu" digabungkan untuk mengetahui probabilitas X secara keseluruhan. Dan didapatkan nilai dari probabilitas X dengan kondisi lama studi "Tepat Waktu" adalah sebesar 0.0087, dengan perhitungan sebagai berikut:

$$\begin{aligned}
&P(X|\text{ Lama Studi = "Tepat Waktu"}) = P(\text{Jenis Kelamin = "P", IPK = "3",}\\
&\text{Tempat Lahir = "Kota", Tipe Sekolah = "SMA", Jumlah Organisasi ="Lebih dari satu",}\\
&\text{Tingkat Ekonomi = "Menengah", Dukungan Orang Tua = "Selalu")}\\
&= \frac{35}{104}\times\frac{59}{104}\times\frac{37}{104}\times\frac{77}{104}\times\frac{55}{104}\times\frac{56}{104}\times\frac{63}{104}\\
&= 0.3365\times 0.5673\times 0.3558\times 0.7404\times 0.5288\times 0.5385\times 0.6058\\
&= 0.0087
\end{aligned}$$

Proses dilakukan kembali dengan probabilitas dari setiap variabel X dengan kondisi lama studi "Tidak Tepat Waktu" digabungkan untuk mengetahui nilai probabilitas X secara keseluruhan. Sehingga di dapatkan nilai dari probabilitas X dengan kondisi lama studi "Tidak Tepat Waktu "adalah sebesar 0.0031, sebagai berikut:

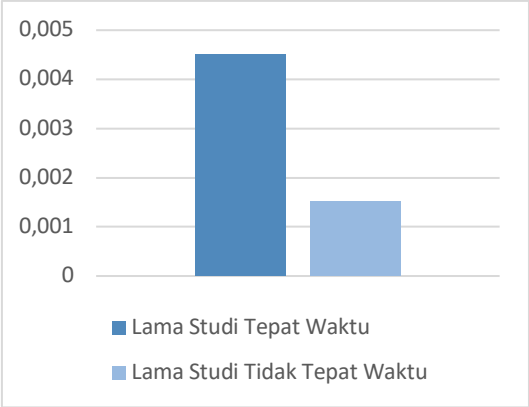
$$\begin{aligned}
&P(X|\text{ Lama Studi = "Tidak Tepat Waktu"}) = P(\text{Jenis Kelamin = "P", IPK = "3",}\\
&\text{Tempat Lahir = "Kota", Tipe Sekolah = "SMA", Jumlah Organisasi = "Lebih dari satu",}\\
&\text{Tingkat Ekonomi = "Menengah", Dukungan Orang Tua = "Selalu")}\\
&= \frac{19}{96}\times\frac{35}{96}\times\frac{44}{96}\times\frac{51}{96}\times\frac{66}{96}\times\frac{50}{96}\times\frac{48}{96}\\
&= 0.1979\times 0.3646\times 0.4583\times 0.5313\times 0.6875\times 0.5208\times 0.5000\\
&= 0.0031
\end{aligned}$$

Dengan demikian nilai probabilitas hipokripsi H berdasarkan kondisi lama studi "Tepat Waktu" adalah 0.004511

$$\begin{aligned}
&P(\text{H}|\text{ Lama Studi = "Tepat Waktu"}) = \frac{P(X|H)\times P(H)}{P(X)}\\
&= \frac{0.0087\times 104}{200}\\
&= 0.004511
\end{aligned}$$

Sedangkan nilai probabilitas hipokripsi H berdasarkan kondisi lama studi "Tidak Tepat Waktu" adalah 0.001510

$$\begin{aligned}
&P(\text{H}|\text{ Lama Studi = "Tidak Tepat Waktu"}) = \frac{P(X|H)\times P(H)}{P(X)}\\
&= \frac{0.0031\times 96}{200}\\
&= 0.001510
\end{aligned}$$



Gambar 3. Hasil Prediksi Lama Studi dari Data *Testing*

Dilakukan pengujian akurasi dan validasi data dengan *K-Fold Cros Validadation*, jumlah keseluruhan data dibagi menjadi beberapa bagaian yang sama banyaknya yaitu dari 200 data di bagi menjadi 4 bagian sehingga masing-masing 50 data dengan diberi label LS1,LS2, LS3, dan LS4. Pada percobaan pertama LS1 selaku data uji dan LS2, LS3, dan LS4 selaku data latih. Pergujian akan dilakukan dengan mengukur kinerja klasifikasi pada sistem dengan menggunakan persamaan (3). Maka didapatkan tingkat akurasi LS1 sebagai berikut:

$$\text{Akurasi} = \frac{\sum Klasifikasi\ Benar}{\sum Data\ Uji} \times 100\%$$

$$Akurasi = \frac{43}{50} \times 100\% = 86\%$$

Pada percobaan perhitungan akurasi kedua, LS2 digunakan sebagai data uji dengan LS1, LS3, LS4 sebagai data latih. Percobaan kedua menghasilkan akurasi sistem sebesar 78%. Pada percobaan ketiga LS3 Sebagai data uji dan LS1, LS2, dan LS4 sebagai data latih diperoleh akurasi sistem sebesar 78%. Sedangkan percobaan terakhir dengan LS4 selaku data uji dan LS1, LS2, LS3 sebagai data latih diperoleh akurasi sistem sebesar 80%.

Fold	Data	Bagian			
Fold-1	LS1 (*Testing*)	LS1(86%)	LS2	LS3	LS4
Fold-2	LS2 (*Testing*)	LS1	LS2(78%)	LS3	LS4
Fold-3	LS3 (*Testing*)	LS1	LS2	LS3(78%)	LS4
Fold-4	LS4 (*Testing*)	LS1	LS2	LS3	LS4(80%)

Gambar 4. Proses *Fold Cross Validation*

Dari keseluruhan percobaan yang dilakukan diperoleh rata-rata akurasi sistem dengan menjumlahkan keseluruhan akurasi data kemudian dibagi dengan banyaknya pembagian data maka dihasilkan 80,5%. Selanjunya menghitung standar deviasi dengan tujuan mengetahui nilai sebaran data dalam sebuah sample data dan melihat jarak akurasi setiap percobaan. Semakin nilai dari standar deviasi mengarah pada nilai nol maka akan menunjukan bahwa nilai dalam data bernilai sama. Dengan menggunakan persamaan (4), didapatkan nilai standar deviasi sebagai berikut:

$$\sigma = \sqrt{\frac{(86-80{,}5)^2 + (78-80{,}5)^2 + (78-80{,}5)^2 + (80-80{,}5)^2}{4}}$$

$$\sigma = 3{,}02$$

KESIMPULAN

Dari penelitian yang telah dilakukan mengenai prediksi lama studi mahasiswa dengan berdasarkan tujuh variabel. Dengan metode Naïve Bayes dalam implementasi, semua variabel memiliki pengaruh satu sama lain terhadap hasil prediksi. Variabel yang mendominasi dalam memprediksi lama studi mahasiswa adalah IPK, jumlah keikutsertaan dalam organisasi, dan dukungan orang tua. Tingkat ekonomi hanya memiliki pengaruh yang kecil dalam memprediksi lama studi mahasiswa baik tepat waktu maupun tidak tepat waktu. Dari hasil pengujian yang dilakukan diperoleh tingkat rata-rata akurasi sebesar 80,5% dan standar deviasi 3,02%. Dengan melihat standar deviasi yang tergolong rendah maka jarak akurasi setiap percobaan terlihat dekat. Untuk peningkatan akurasi perlu adanya pegembangan baik dari sisi penambahan jumlah data set maupun dari sisi metode Naïve Bayes yang kombinasikan dengan metode-metode lainnya.

DAFTAR PUSTAKA

[1] BAN-PT, "Instrumen Akreditasi Perguruan Tinggi," no. April, pp. 7–9, 2019.
[2] I. B. A. Peling, I. N. Arnawan, I. P. A. Arthawan, and I. G. N. Janardana, "Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm," *Int. J. Eng. Emerg. Technol.*, vol. 2, no. 1, p. 53, 2017.
[3] B. D. F. Kurniatullah and Y. T. C. Pramudi, "Estimation of Students' Graduation Using Multiple Linear Regression Method," *J. Appl. Intell. Syst.*, vol. 2, no. 1, pp. 29–36, 2017.
[4] M. I. Zulfa, A. Fadli, and Y. Ramadhani, "Classification model for graduation on time study using data mining techniques with SVM algorithm," *AIP Conf. Proc.*, vol. 2094, no. April, 2019.
[5] A. Kesumawati and D. T. Utari, "Predicting patterns of student graduation rates using Naïve bayes classifier and support vector machine Predicting Patterns of Student Graduation Rates Using Naïve Bayes Classifier and Support Vector Machine," vol. 060005, no. October 2018, 2021.
[6] A. D. Herlambang, S. H. Wijoyo, and A. Rachmadi, "Intelligent Computing System to Predict Vocational High School Student Learning Achievement Using Naive Bayes Algorithm," vol. 4, no. 1, pp. 15–25, 2019.
[7] A. Kasem, "Learning Analytics in Universiti Teknologi Brunei : Predicting Graduates Performance," *2018 Fourth Int. Conf. Adv. Comput. Commun. Autom.*, vol. 2017, no. 4, pp. 1–5, 2018.
[8] F. Razaque *et al.*, "Using naïve bayes algorithm to students' bachelor academic performances analysis," *4th IEEE Int. Conf. Eng. Technol. Appl. Sci. ICETAS 2017*, vol. 2018-Janua, pp. 1–5, 2018.
[9] O. Moscoso-zea, P. Saa, and S. Luján-mora, "Evaluation of algorithms to predict graduation rate in higher education institutions by applying educational data mining," *Australas. J. Eng. Educ.*, vol. 00, no. 00, pp. 1–10, 2019.
[10] S. Lolo and S. Ricambi, "AN ANALYSIS OF BAYESIAN ALGORITHM IN GRADUATES TO STUDY FURTHER INTO UNIVERSITY," pp. 871–879, 2019.

[11] G. Wang, "Applying Customer Loyalty Classification with RFM and Naïve Bayes for Better Decision Making," *2019 Int. Semin. Appl. Technol. Inf. Commun.*, pp. 564–568, 2019.
[12] Y. Ko, "How to use negative class information for Naive Bayes classification," *Inf. Process. Manag.*, vol. 53, no. 6, pp. 1255–1268, 2017.
[13] W. Astuti, "Kinerja Metode Naïve Bayes dalam Prediksi Lama Studi Mahasiswa Fakultas Ilmu Komputer," vol. 3, no. 2, pp. 107–111, 2018.
[14] M. Sivasakthi, "Classification and prediction based data mining algorithms to predict students' introductory programming performance," *Proc. Int. Conf. Inven. Comput. Informatics, ICICI 2017*, no. Icici, pp. 346–350, 2018.
[15] J. Y. Hutapea, Y. T. Samuel, and H. Sitorus, "a Comparison Of Accuracy Between Two Methods Naive Bayes Algorithm And Decision Tree-J48 To Predict The Stock Price Of Pt Astra Internasional Tbk Using Data From Indonesia Stock Exchange," *Abstr. Proc. Int. Sch. Conf.*, vol. 7, no. 1, pp. 1244–1258, 2019.
[16] T. Imandasari, E. Irawan, A. P. Windarto, and A. Wanto, "Algoritma Naive Bayes Dalam Klasifikasi Lokasi Pembangunan Sumber Air," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, no. November, p. 750, 2019.
[17] A. P. Wibawa *et al.*, "Naïve Bayes Classifier for Journal Quartile Classification," *Int. J. Recent Contrib. from Eng. Sci. IT*, vol. 7, no. 2, p. 91, 2019.
[18] Y. A. Gerhana, I. Fallah, W. B. Zulfikar, D. S. Maylawati, and M. A. Ramdhani, "Comparison of naive Bayes classifier and C4.5 algorithms in predicting student study period," *J. Phys. Conf. Ser.*, vol. 1280, no. 2, 2019.
[19] K. J. Grimm, G. L. Mazza, and P. Davoudzadeh, "Model Selection in Finite Mixture Models: A k-Fold Cross-Validation Approach," *Struct. Equ. Model.*, vol. 24, no. 2, pp. 246–256, 2017.
[20] R. Valavi, J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Arroita, "blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models," *Methods Ecol. Evol.*, vol. 10, no. 2, pp. 225–232, 2019.
[21] F. Tempola, M. Muhammad, and A. Khairan, "Perbandingan Klasifikasi Antara Knn Dan Naive Bayes Pada Penentuan Status Gunung Berapi Dengan K-Fold Cross Validation Comparison of Classification Between Knn and Naive Bayes At the Determination of the Volcanic Status With K-Fold Cross," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, pp. 577–584, 2018.
[22] B. Cahyono, "ANALISIS KETRAMPILAN BERFIKIR KRITIS DALAM," vol. 8, no. 1, pp. 50–64, 2017.
[23] A. A. Rondiyah, N. E. Wardani, and K. Saddhono, "UNTUK MENINGKATKAN PENDIDIKAN KARAKTER KEBANGSAAN DI ERA MEA (MASAYARAKAT EKONOMI ASEAN)," pp. 141–147, 2015.
[24] T. Carmelia, S. Tiatri, E. Wijaya, and L. Belakang, "Akademik Dengan JOB Performance Pada Mahasiswa Aktif Organisasi Kemahasiswaan," pp. 184–197, 2017.
[25] S. Haq, "Economic Education Analysis Journal," vol. 5, no. 3, pp. 1034–1045, 2016.