



Implementasi Klasifikasi Naive Bayes Dalam Memprediksi Lama Studi Mahasiswa

Muhammad Fuad¹, Muhammad Arfah Wahli², Hazriani³, Yuyun⁴

^{1,2,3,4}Sistem Komputer, Program Pasca Sarjana, Universitas Handayani Makassar

⁴Badan Riset Dan Inovasi Nasional (BRIN)

Fuadmantale01@gmail.com

Abstract

Under normal conditions, undergraduate or undergraduate students from a university can complete their studies for 4 years or 8 semesters. In fact, many students complete their study period of more than 4 years. It is known that in the academic year 20XX/20XX there were 161 people who were accepted as students. Of the 161 people admitted, 100 people have completed their study period of about 4 years and the remaining 61 people have completed their studies for 5 years. Based on the problems above, this research implements a classification that can help the university predict the length of study of students who are currently studying in various study programs at the University. The method that the author presents in the classification for predicting the length of a student's study period is the Naive Bayes Algorithm. By using the Java-based Rapid Miner tool to classify graduation data. Then the implementation of data mining which is divided into 136 data training data and 25 data testing data with naive Bayes managed to obtain an accuracy rate of 82% which is also a relatively good parameter.

Keywords: data mining, Naive Bayes Rapid Miner, training data, testing data

Abstrak

Dalam kondisi normal, mahasiswa strata-1 atau S1 dari suatu Universitas, dapat menyelesaikan studi nya selama 4 tahun atau 8 semester. Pada kenyataannya, banyak mahasiswa yang menyelesaikan masa studinya lebih dari 4 tahun. Diketahui bahwa faktanya pada tahun akademik 20XX/20XX terdapat 161 orang yang diterima sebagai mahasiswa. Dari 161 orang yang diterima, 100 orang telah menyelesaikan masa studi sekitar 4 tahun dan sisanya 61 orang menyelesaikan studinya selama 5 tahun. Berdasarkan permasalahan di atas, maka penelitian ini melakukan implementasi klasifikasi yang dapat membantu pihak Universitas dalam memprediksi lama studi mahasiswa yang saat ini sedang kuliah di berbagai program studi pada Universitas. Metode yang penulis sajikan dalam klasifikasi untuk memprediksi lamanya masa studi mahasiswa adalah Algoritma naive bayes. Dengan menggunakan tool Rapid Miner yang berbasis java untuk mengklasifikasikan data kelulusan. Kemudian implementasi data mining yang terbagi atas data training sebanyak 136 data dan data testing 25 data dengan naive bayes berhasil memperoleh tingkat keakuratan sebesar 82% yang juga parameternya tergolong baik.

Kata kunci: Kata Kunci: data mining, data training, data testing, Naive Bayes, Rapid Miner

1. Pendahuluan

Dalam kondisi normal, mahasiswa strata-1 atau S1 dari suatu Universitas, dapat menyelesaikan studi nya selama 4 tahun atau 8 semester. Pada kenyataannya, dengan berbagai sebab, poly mahasiswa yang menuntaskan masa studi nya lebih 4 tahun. Lama studi mahasiswa juga menjadi salah satu kriteria dalam penilaian akreditasi baik program studi maupun Universitas negeri atau swasta menurut lembaga Badan Akreditasi Nasional Perguruan Tinggi (BAN-PT). Pengumpulan data ini bertujuan untuk mendapatkan data yang digunakan untuk penelitian. Data diperoleh melalui Url <https://www.kaggle.com/datasets>

Metode yang penulis sajikan dalam pembagian terstruktur mengenai untuk memprediksi lamanya

masa studi mahasiswa yaitu Algoritma Naive Bayes. Naive Bayes merupakan sebuah pembagian terstruktur tentang menggunakan Metode probabilitas dan statistik oleh ilmuwan Inggris Thomas Bayes. Algoritma Naive Bayes memprediksi peluang pada masa depan berdasarkan pengalaman dalam masa sebelumnya menjadi akibatnya dikenal menjadi Teorema Bayes [1] Penulis menggunakan tool Rapid Miner yang berbasis java untuk mengklasifikasikan data kelulusan. Dalam proses prediksi lamanya masa studi mahasiswa yang penulis buat, admin terlebih dahulu menyiapkan data-data yang diperlukan lalu memasukkannya ke dalam file dikomputer atau laptop. Setelah itu, maka akan dilakukan proses pengambilan data berupa microsoft excel yang ada pada file di laptop admin lalu mulai memilih label dan

tipe data, kemudian melakukan pengukuran tingkat akurasi metode Naive Bayes. Selesai melakukan pengukuran tingkat akurasi metode Naive Bayes, lalu hasilnya akan sebagai sebuah tabel. Dengan proses klasifikasi prediksi lamanya masa studi mahasiswa yang penulis buat, diharapkan dapat memudahkan mahasiswa dalam meningkatkan kualitas di bidang pendidikan dari program studi maupun universitas. Menurut latar belakang tersebut maka penulis tertarik mengajukan penelitian menggunakan judul “Implementasi Klasifikasi Naive Bayes Dalam Memprediksi Lama Studi Mahasiswa” [2]

2. Metode Penelitian

Tahapan metode yang dipakai yaitu Knowledge Discovery in Database (KDD) mempunyai arti yang berbeda, akan namun saling berkaitan satu menggunakan yang lainnya. Tahapan- tahapan KDD terlihat menjadi berikut [3]

2.1 1. Data Selection

merupakan tahapan pemilihan data, dalam suatu database data yang dipakai hanya beberapa data yang sesuai untuk dianalisa

2.2. Pre-processing/Cleaning

Pada proses ini dilakukan proses dimana data yang bermasalah dimusnahkan. Misalkan data yang diperoleh berdasarkan database eksklusif umumnya mempunyai data yang lengkap, terdapat beberapa isian yang hilang, data valid juga galat tulis. Maka data- data misalnya itu akan dibuang. Proses cleaning pula mensugesti kinerja menurut data mining karena data yang diproses akan berkurang jumlahnya dan kompleksitasnya

2.3. Data Mining

Yaitu tahapan yang dilakukan saat metode dipakai sebagai akibatnya mendapatkan keterangan baru menurut data yang ada. Pada tahapan ini juga perkumpulan data yang sudah melewati tahapan sebelumnya dianalisa menurut metode yang dipilih buat memperoleh output akhir

Naive Bayes yaitu proses pengklasifikasian probabilistik memakai cara menghitung sekumpulan probabilitas menggunakan menjumlahkan frekuensi dan kombinasi nilai data set yang diberikan. Algoritma menggunakan Teorema Bayes memakai mengasumsikan seluruh atribut independen atau nir saling ketergantungan yang diberikan berdasarkan nilai dalam variabel kelas [4]. Penjabaran pembagian terstruktur tentang Bayes adalah pembagian terstruktur mengenai statistik yang sanggup memprediksi kelas suatu anggota probabilitas [5].

Pada teorema Bayes terdapat bentuk umum, (Silvi, 2020):

$$P(C_i|X) = \frac{P(X|C_i) \times P(C_i)}{P(X)}$$

Keterangan:

X : Data menggunakan group yang belum diketahui

Ci : Hipotesis Data X adalah suatu group spesifik

P(CiX) : Probabilitas Hipotesis Ci menurut syarat X

P(Ci) : Probabilitas Hipotesis Ci

P(XCi) : Probabilitas X menurut syarat tersebut

P(X) : Probabilitas menurut X

Langkah awal, dilakukan pengelompokan data training untuk target yang dicapai yaitu lulus atau tidaknya seorang mahasiswa, lalu langkah selanjutnya yaitu menghitung nilai probabilitas menurut pengelompokan data pelatihan yang ada. Nilai probabilitas tertinggi setelah penerapan contoh sebagai penentu prediksi kelulusan mahasiswa cepat atau terlambat.

3. Hasil dan Pembahasan

Pada proses ini dilakukan dengan tahap pengambilan data yang sudah diperoleh melalui Url <https://www.kaggle.com/datasets>. sebanyak 161 data mahasiswa dan mahasiswi yang lulus tahun periode 2019-2020, kemudian pada 161 data tersebut di dilakukan sesuai dengan tahapan analisa metode Naive Bayes.

3.1. Data Selection

Proses menunjukan bahwa atribut JENIS KELAMIN diketahui jenis kelamin mahasiswa, Asal Daerah dapat diketahui mahasiswanya berasal dari lingkup universitas, lalu atribut Cuti diketahui apakah mahasiswanya mengambil cuti dalam proses studi atau tidak, lalu atribut Asal SMA mengetahui asal lulusan mahasiswa, kemudian attribute Nikah akan mengetahui dalam penyelesai studi mahasiswa sudah bersatutut apa, dan attribute Program mengetahui jalur masuknya mahasiswa.

3.2. Data Pre-procesing

Pembersihan data dilakukan menggunakan memperbaiki data kelulusan mahasiswa sebelum dilakukan proses mining, misalnya adanya missing values dalam data. Dalam hal ini, missing value pada data mahasiswa dari berdasarkan data-data yang atributnya tidak mempunyai nilai atau informasi.

3.3. Proses Pengolahan Data Mining

Pada proses ini dilakukan dengan tahap pengambilan data yang sudah dibersihkan terlebih dahulu yaitu bisa disebut dengan Data Mining sebanyak 161 data mahasiswa dan mahasiswi yang lulus tahun periode 2019-2020. Kemudian pada 161 data tersebut di Split data lagi sebagai 80 % data training dan 20% data testing pada ambil berdasarkan mahasiswa yang belum lulus. Penelitian ini juga menjadi perbandingan dari proses split data 80:20.

Tabel 3.1 Potongan Data Training

No	Jenis Kelamin	Asal Daerah	Cuti	Asal SMA	Nikah	Program	Kelas
1	Laki-Laki	Dalam Kabupaten	Tidak	SMK	Belum	Reguler	Terlambat
2	Laki-Laki	Dalam Kabupaten	Tidak	MA	Sudah	Reguler	Terlambat
3	Laki-Laki	Dalam Provinsi	Ya	MA	Belum	Reguler	Tepat
4	Perempuan	Luar Pulau	Tidak	SMA	Belum	Reguler	Terlambat
5	Laki-Laki	Dalam Kabupaten	Ya	SMA	Belum	Karyawan	Terlambat
...							
13	Laki-Laki	Dalam Provinsi	Tidak	SMA	Belum	Reguler	Tepat
13	Perempuan	Luar Pulau	Tidak	SMA	Belum	Reguler	Tepat
6							

Tabel 3.2 Potongan Data Testing

No	Jenis Kelamin	Asal Daerah	Cuti	Asal SMA	Nikah	Program	Kelas
137	Perempuan	Dalam Provinsi	Tidak	SMK	Belum	Karyawan	Tepat
138	Perempuan	Dalam Provinsi	Tidak	SMK	Belum	Reguler	Tepat
139	Laki-Laki	Dalam Kabupaten	Tidak	MA	Sudah	Reguler	Terlambat
140	Laki-Laki	Dalam Provinsi	Ya	MA	Belum	Reguler	Tepat
...							
161	Laki-Laki	Dalam Kabupaten	Tidak	SMA	Sudah	Reguler	Tepat

3.4 Proses Perhitungan Naive Bayes

Dengan menghitung probabilitas masing-masing kelas pencarian tersebut merupakan kelayakan dari metode Naive Bayes. Prediksi kelulusan ditentukan lewat dua kelas yaitu “Tepat dan “Terlambat”. Dimana ada tahapan- tahapannya sebagai berikut.

1. Proses Perhitungan Jumlah kelas

$$P(C_i) = \frac{\text{Kemunculan data masuk}}{\text{Jumlah keseluruhan data}}$$

Tabel 3.3 Probabilitas Kelas

Probabilitas Kelas	
Kelas	Nilai
Tepat	100
Terlambat	61

2. Proses Perhitungan Jumlah kelas

$$P(X|C) = \frac{\text{Jumlah Data Dengan Atribut}}{\text{Jumlah Data Masuk}}$$

Tabel 3.4 Probabilitas Atribut Jenis Kelamin

Jenis Kelamin	Tepat	Terlambat
Laki-laki	0.523809525	0.730769231
Perempuan	0.476190476	0.25

Tabel 3.5 Probabilitas Atribut Asal Daerah

Jenis Kelamin	Tepat	Terlambat
Dalam Kabupaten	0.19047619	0.730769231
Dalam Provinsi	0.642857143	0
Luar Pulau	0.166666667	0.269239769

Tabel 3.6 Probabilitas Atribut CUTI

Cuti	Tepat	Terlambat
Tidak	0.833333333	0.692307692
Ya	0.166666667	0.307692308

Tabel 3.7 Probabilitas Atribut Asal SMA

Asal SMA	Tepat	Terlambat
SMA	0.523809524	0.576923077
SMK	0.30952381	0.153846154
MA	0.166666667	0.269239769

Tabel 3.8 Probabilitas Atribut Program

Program	Tepat	Terlambat
Reguler	0.845238095	0.692307692
Karyawan	0.154761905	0.307692308

Tabel 3.10 Data Hasil Uji training dan testing

Kelas	Kelas Prediksi	Tepat	Terlambat
Tepat	Tepat	0.989240043	0
Tepat	Tepat	5.402772544	0
Terlambat	Terlambat	0.223100914	1.8552746
Tepat	Tepat	0.640020748	0
Terlambat	Tepat	2.370447213	1.3600697
Terlambat	Terlambat	0.109126423	2.1315324
Tepat	Terlambat	0.701174302	3.9755885
Tepat	Tepat	10.05746889	0
Tepat	Tepat	2.370447213	1.3600697
Tepat	Tepat	0.989240043	0
Tepat	Tepat	5.402772544	0
Terlambat	Terlambat	0.223100914	1.8552746
Tepat	Tepat	0.640020748	0
Terlambat	Tepat	2.370447213	1.3600697
Terlambat	Terlambat	0.109126423	2.1315324
Tepat	Terlambat	0.701174302	3.9755885
Tepat	Tepat	10.05746889	0
Tepat	Tepat	2.370447213	1.3600697
Tepat	Tepat	0.989240043	0
Tepat	Tepat	5.402772544	0
Terlambat	Terlambat	0.223100914	1.8552746
Tepat	Tepat	0.640020748	0
Terlambat	Tepat	2.370447213	1.3600697
Terlambat	Terlambat	0.109126423	2.1315324
Tepat	Terlambat	0.701174302	3.9755885

3.5 Interpretation/Evaluation

yaitu proses evaluasi menggunakan pengujian confusion matrix. Confusion matrix itu tabel yang memberikan output kerja berdasarkan suatu klasifikasi. Menghitung taraf akurasi adalah hal terpenting pada sebuah penelitian supaya bisa diketahui taraf keberhasilan dan kegagalan terhadap penelitian [3].

Tabel 3.11 Tabel Confusion Matrix

Aktual	Prediksi	
	Tepat	Terlambat
	TP	FP
Terlambat	FN	TN

TN = nilai true negative

TP = nilai true positive

FP = nilai false positive

FN = nilai false negative

Tabel 3.12 Tabel Hasil Confusion Matrix

Confusion Matrix		
Aktual	Prediksi	
	Tepat	Terlambat
	13	3
Terlambat	0	0

TN merupakan jumlah lulus cepat dengan berhasil diprediksi menjadi lulus cepat. TP merupakan jumlah lulus terlambat dengan berhasil diprediksi lulus terlambat. FP merupakan jumlah lulus cepat dengan diprediksi lulus terlambat. Sedangkan, FN menyatakan lulus terlambat dengan diprediksi lulus cepat [5].

3.5.1 1. Pengujian Dengan Data 80:20

Pada tahap pengujian ini memberikan hasil dari proses split data 80:20 ada pada 80% dari 100 data Tepat lulus mahasiswa serta 20% dari 61 data mahasiswa yang Terlambat lulus

Tabel 3.13 Tabel Perhitungan

AKURASI	81%	TP+TN / Jumlah Set Data
PRESISI	81%	TP / (TP + FP)
RECALL	100%	TP / (TP + FN)

4. Kesimpulan

Berdasarkan penelitian yang telah dilakukan dengan topik data mining pada proses klasifikasi dan prediksi lama studi mahasiswa di Universitas terkhususnya tahun ajaran kelulusan 2019-2020 dapat diambil kesimpulannya sebagai berikut.

1.Implementasi metode Naive Bayes pada prediksi using studinya mahasiswa berhasil dengan data training sebesar 136 data dan data testing 25 data bahwa mahasiswa dan mahasiswi studinya Cepat tidak Terlambat

2.Hasil implementasi metode Naive Bayes dalam prediksi lama studinya mahasiswa berhasil dengan memperoleh ke akuratan/accuracy sebesar 81% , Presisi 81%, dan recall sebesar 100%

3.Dari hasil perhitungan confusion matrix terdapat juga analisa error atau kesalahan dalam dataset sehingga mendapatkan kesalahan error accuracy 19%, Presisi 19% yang dimana seharusnya mendapat data perhitungan confusion matrix accuracy 100%, Presisi 100%

Ucapan Terimakasih

Terima kasih diucapkan kepada; Ibu Hazriani Zainudin dan Bapak Yuyun Wabula selaku Author yang telah membimbing dalam penyelesaian peper ini.

Daftar Rujukan

- [1] Alim Murtopo, A., Kelulusan Tepat Waktu, P., Alim Murtopo STMIK YMI TEGAL, A., & Pendidikan No, J. (n.d.). Prediksi Kelulusan Tepat Waktu Mahasiswa STMIK YMI Tegal Menggunakan Algoritma Naïve Bayes Time Graduation Prediction by Using Naïve Bayes Algorithm at STMIK YMI Tegal.
- [2] Armansyah, A., & Ramli, R. K. (2022). Model Prediksi Kelulusan Mahasiswa Tepat Waktu dengan Metode Naïve Bayes. *Edumatic: Jurnal Pendidikan Informatika*, 6(1), 1–10. <https://doi.org/10.29408/edumatic.v6i1.4789>
- [3] Hennry, K., Malelak, L., Made, I., Ardiada, D., Feoh, G., Informatika, T., Sains, K., Teknologi, D., Dhyana, U., Jalan, P., Padang, R., Tegaljaya, L., Utara, K., & Badung, K. (n.d.). Implementasi Klasifikasi Naive Bayes Dalam Memprediksi Lama Studi Mahasiswa (Studi Kasus: Universitas Dhyana Pura). <https://doi.org/10.31598>
- [4] Qisthiano, M. R., Kurniawan, T. B., Negara, E. S., & Akbar, M. (2021). Pengembangan Model Untuk Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu dengan Metode Naïve Bayes. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 5(3), 987. <https://doi.org/10.30865/mib.v5i3.3030>
- [5] Setiyani, L., Wahidin, M., Awaludin, D., & Purwani, S. (2020). Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Data Mining Naïve Bayes : Systematic Review. *Faktor Exacta*, 13(1), 35. <https://doi.org/10.30998/faktorexacta.v13i1.5548>