Penerapan Teknik *Data Mining* Untuk Klasifikasi Ketepatan Waktu Lulus Mahasiswa Teknik Informatika Universitas Telkom Menggunakan Algoritma *Naive Bayes Classifier*

Application of Data Mining Techniques for Classification the Graduation on Time of Informatic Engineering Telkom University Students using Naive Bayes Classifier Algorithm

Naziah Amalia[1], Shaufiah, ST., MT.[2], Siti Sa'adah, ST., MT.[3]
Program Studi Teknik Informatika, School of Computing, Telkom University, Bandung
[1]naziahamalia@gmail.com, [2]shaufiah@gmail.com, [3]tisataz@gmail.com

Abstrak

Ketepatan waktu lulus mahasiswa merupakan hal yang penting bagi sebuah universitas karena merupakan salah satu syarat akreditasi, tak terkecuali Universitas Telkom. Universitas Telkom merupakan salah satu universitas swasta di Indonesia yang menawarkan berbagai program studi, salah satunya Teknik Informatika. Program studi tersebut menerima kurang lebih 640 mahasiswa setiap tahunnya. Namun, jumlah mahasiswa tersebut tidak diimbangi dengan mahasiswa yang lulus tepat waktu. Mahasiswa yang lulus tepat waktu hanya mencapai kisaran 15% setiap tahunnya. Banyak aspek yang menyebabkan mahasiswa lulus tidak tepat waktu. Beberapa diantaranya adalah kota asal, jumlah SKS, IPK, berapa kali mengambil mata kuliah TA, dan masih banyak lagi aspek yang bisa mempengaruhi ketepatan waktu lulus. Ketepatan waktu lulus tersebut dapat diklasifikasikan berdasarkan aspek yang ada.

Pada Tugas Akhir ini, dilakukan klasifikasi ketepatan waktu lulus dengan menggunakan algoritma *Naive Bayes Classifier*. Algoritma tersebut dibangun dengan menggunakan Java Netbeans. Pengujian dilakukan untuk mengetahui titik optimum dari algoritma *Naive Bayes Classifier*. Hasil pengujian dengan perbandingan data tersebut didapatkan akurasi sebesar 91%. Titik optimum algoritma tersebut berada pada pembagian data dengan perbandingan 90% data training dan 10% data testing. Berdasarkan pengujian yang dilakukan per atribut, faktor yang mempengaruhi ketepatan waktu lulus merupakan atribut TA1, karena memiliki prosentase yang paling tinggi dibandingkan atribut yang lain.

Kata kunci: ketepatan waktu lulus, klasifikasi, Data Mining, *Naive Bayes Classifier*

1. Pendahuluan

Universitas Telkom atau disingkat Tel-U adalah sebuah perguruan tinggi swasta di Indonesia. Universitas Telkom merupakan penggabungan dari empat institusi yang berada di bawah badan penyelenggara Yayasan Pendidikan Telkom (YPT), yaitu Institut Teknologi Telkom (IT Telkom), Institut Manajemen Telkom (IM Telkom), Politeknik Telkom, dan Sekolah Tinggi Seni Rupa dan Desain Indonesia Telkom (STISI Telkom). Dalam proses penggabungan menjadi Universitas Telkom pada tahun 2013, IT Telkom ditransformasikan menjadi Fakultas Teknik (FT) atau Telkom Engineering School (TES). Selanjutnya pada tahun 2014 Fakultas Teknik dikembangkan menjadi tiga fakultas, yaitu: 1). Fakultas Teknik Elektro (FTE) atau *School of Electrical Engineering* (*SEE*), 2). Fakultas Rekayasa Industri (FRI) atau *School of Industrial Engineering* (*SIE*), dan 3). Fakultas Informatika (FIF) atau *School of Computing* (*SC*).

Pada Tugas Akhir ini penulis melakukan penelitian pada salah satu Program Studi yang ada di Fakultas Informatika (FIF) atau *School of Computing* (*SC*). Fakultas Informatika (FIF) atau *School of Computing* (*SC*) terdiri dari 3 Program Studi yaitu Program Studi S1 Teknik Informatika, S1 Ilmu Komputasi, dan S2 Teknik Informatika, dan Program Studi yang penulis ambil sebagai acuan penelitian Tugas Akhir ini adalah Teknik Informatika.

Teknik Informatika merupakan salah satu program studi yang ada di Universitas Telkom yang telah meluluskan lebih dari 3500 mahasiswa dan hanya 48,12% diantaranya yang lulus tepat waktu [Laporan Manajemen Universitas Telkom Triwulan I 2014]. Data tersebut menunjukkan bahwa masih banyak mahasiswa yang lulus tidak tepat waktu. Hal ini bertentangan dengan tujuan prodi yang ingin meningkatkan akreditasi, yang mana ketepatan waktu lulus merupakan salah satu syarat akreditasi. Keterlambatan waktu lulus mahasiswa tersebut dapat disebabkan oleh beberapa hal, salah satunya adalah terganggunya proses perkuliahan. Terganggunya proses perkuliahan mahasiswa dapat dipengaruhi oleh berbagai aspek, seperti misalnya kurangya kemampuan mahasiswa terhadap mata kuliah yang diambil, kurangnya kehadiran mahasiswa dalam perkuliahan, dan berbagai aspek lainnya yang tidak dapat diukur.

Ketepatan waktu lulus mahasiswa dapat diklasifikasikan dengan menggunakan teknik data mining. Teknik

data mining yang digunakan dalam penelitian ini adalah algoritma *Naive Bayes Classifier*. *Naive bayes classifier* digunakan untuk menentukan probabilitas masing-masing parameter, sehingga dapat diketahui parameter yang menjadi penyebab paling berpengaruh terhadap ketepatan waktu lulus mahasiswa. *Naive Bayes Classifier* merupakan sebuah pengklasifikasi probabilitas sederhana yang mengaplikasikan Teorema Bayes. Ide dasar dari Teorema Bayes adalah menangani masalah yang bersifat hipotesis yakni mendesain suatu fungsi klasifikasi untuk memisahkan objek. Dari hasil studi perbandingan algoritma klasifikasi, didapatkan bahwa hasil klasifikasi bayesian atau lebih dikenal dengan *Naive Bayes Classification* dari segi performa lebih baik dari algoritma *Decision Tree* dan algoritma *Selected Neural Networks Classifiers*. *Naive Bayes Classifiers* juga memilki kecepatan dan keakuratan yang tinggi bila di implementasikan ke dalam *database* yang ukurannya besar.

2. Landasan Teori

2.1. Klasifikasi pada Data Mining

Data mining, sering juga disebut *knowledge discovery in database* (KDD), adalah kegiatan meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar [2].

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi, terdapat dua pekerjaan utama yang dilakukan, yaitu, pembangunan model sebagai prototype untuk disimpan sebagai memori dan penggunaan model tersebut untuk melakukan pengenalan/klasifikasi/prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut dalam model yang sudah disimpannya [4].

2.2. Metode-Metode Pilihan pada Klasifikasi

Berikut merupakan beberapa metode yang digunakan pada klasifikasi secara umum, diantaranya adalah [8]:

1. Klasifikasi berdasarkan Pohon Keputusan (*Decission Tree*)
   Pohon keputusan atau Decission Tree merupakan proses pelatihan data set yang memiliki atribut dengan besaran nominal, yaitu bersifat kategoris dan setiap nilai tidak bisa dijumlahkan atau dikurangkan. Pada umumnya, ciri kasus berikut cocok untuk diterapkan pada decision tree [2]:
   a. Data/example dinyatakan dengan pasangan atribut dan nilainya.
   b. Label/output data biasanya berniali diskrit.
   c. Data mempunyai missing value.
2. Klasifikasi *Bayesian*
   Klasifikasi Bayesian merupakan klasifikasi berdasarkan statistic classifiers. Metode ini dapat mngklasifikasikan sebuah kelas dengan probabilitas dari setiap atribut. Klasifikasi Bayesian didasarkan pada Bayes Theorem. Beberapa penelitian yang membandingkan algoritma klasifikasi telah menemukan sebuah klasifikasi Bayesian sederhana yang dikenal dengan nama Naïve Bayes Classifier. Algoritma ini telah dibandingkan dengan decision tree dan selected neural network secara performansi. Klasifikasi Bayesian juga memiliki tingkat akurasi yang tinggi dan cepat jika diterapkan pada database yang besar. Naïve Bayes Classifier mengenali setiap atribut pada data set sebagai atribut yang independent, sehingga disebut algoritma yang naïve.
3. Klasifikasi berdasarkan Propagasi Balik (*Backpropagation*)
   Propagasi balik atau Backpropagation merupakan sebuah algoritma pembelajaran dari neural network. Secara umum, neural network merupakan satu set input/output yang terhubung dan pada setiap koneksi memiliki *weight.* Input/output yang terhubung tersebut mengadopsi sistem saraf manusia, yang pemrosesan utamanya ada di otak. Bagian terkecil dari otak manusia adalah sel saraf yang disebut unit dasar pemroses informasi atau neuron. Ada sekitar 10 miliar neuron dalam otak manusia dan sekitar 60 triliun koneksi. Dengan menggunakan neuron-neuron tersebut secara simultan, otak manusia dapat memproses informasi secara parallel dan cepat, bahkan lebih cepat dari computer tercept saat ini. Dengan analogi system kinerja otak tersebut, neural network terdiri dari unit pemroses yang disebut neuron yang berisi penambah dan fungsi aktivasi, sejumlah bobot, sejumlah vector masukan. Fungsi aktivasi berguna untuk mengatur keluaran yang diberikan oleh neuron.
   Propagasi balik (*Backpropagation*) mempelajari data dengan memprediksi setiap jaringan pada setiap atribut dan kemudian mengklasifikasikannya kedalam kelas target. Kelas target dapat diketahui melalui training pada data set. Besar ‘weight’ akan dimodifikasi untuk mendapatkan mean square error yang

paling kecil antara kelas prediksi dan kelas target. Modifikasi ini dilakukan dengan langkah mundur dari output layer, setiap hidden layer hingga ke hidden layer pertama. Itulah asal penamaan metode Propagasi Balik (*Backpropagation*).

4. *Support Vector Machine*

Support Vector Machine (SVM) merupakan metode klasifikasi yang berakar dari teori pembelajaran statistic yang hasilnya sangat menjanjikan untuk memberikan hasil yang lebih baik daripada metode yang lain. SVM juga bekerja dengan baik pada set data berdimensi tinggi, bahkan SVM yang menggunakan teknik kernel harus memetakan data asli dari dimensi asalnya menjadi dimensi lain yang relative lebih tinggi. Pada SVM, data latih yang akan dipelajari hanya data terpilih saja yang berkontribusi untuk membentuk model yang digunakan dalam klasifikasi yang akan dipelajari. Hal ini menjadi kelebihan SVM karena tidak semua data latih akan dipandang untuk dilibatkan dalam setiap iterasi pelatihannya. Data yang berkontribusi tersebut disebut *support vector* sehingga metodenya disebut *Support Vector Machine*.

2.3. Algoritma Naive Bayes Classifier

Klasifikasi Bayesian adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class. Klasifkasi bayesian didasarkan pada teorema bayes. Teorema keputusan bayes adalah pendekatan statistik yang fundamental dalam pengenalan pola (*pattern recoginition*). Pendekatan ini didasarkan pada kuantifikasi *trade-off* antara berbagai keputusan klasifikasi dengan menggunakan probabilitas dan ongkos yang ditimbulkan dalam keputusan-keputusan tersebut. Ide dasar dari bayes adalah menangani masalah yang bersifat hipotesis yakni mendesain suatu klasifikasi untuk memisahkan objek [2].

Misalkan terdapat beberapa alternatif hipotesis h € H. Dalam *bayes learning*, dimaksimalkan hipotesis yang paling mungkin, h, atau *maximum a priori* (MAP), jika diberi data, x. Secara matematis ini bisa dirumuskan [2]

$$\begin{aligned} h_{MAP} &= arg\ max\ P(h|x) \\ &= \arg\max \frac{(P(x|h)P(h))}{P(x)} \\ &= \arg\max P(x|h)P(h) \end{aligned} \quad \text{.....(2.3)}$$

Dalam banyak kasus diasumsikan bahwa setiap hipotesis h dalam H mempunyai peluang prior yang sama (P(hj) = P(hi) untuk semua hi dan hj dalam H). P(x|h) sering disebut *likelihood* dari data x diberikan h dan sembarang hipotesis yang memaksimalkan P(x|h) dinamakan hipotesis maximum *likelihood*, yang dinotasikan

$$h_{ML} = \arg max_{h\epsilon H} P(h|x) \quad \text{.....(2.4)}$$

Dari hasil studi perbandingan algoritma klasifikasi, didapatkan bahwa hasil klasifikasi bayesian atau lebih dikenal dengan *Naive Bayes Classification* dari segi performa lebih baik dari algoritma *Decision Tree* dan algoritma *Selected Neural Networks Classifiers*. *Naive Bayesian Classifiers* juga memilki kecepatan dan keakuratan yang tinggi bila di implementasikan ke dalam *database* yang ukurannya besar.

Naive Bayesian Classifiers berasumsi bahwa efek dari status pada kelas yang diberikan adalah *independent* terhadap nilai atribut yang lainnya. Asumsi ini biasa disebut dengan *class conditional independence*. Itu dibuat untuk menyederhanakan komputasi yang terkait dan dalam hal ini disebut sebagai '*naive*'.

Umumnya, Bayes mudah dihitung untuk fitur bertipe kategoris. Namun untuk fitur dengan tipe numerik (kontinu) ada perlakuan khusus sebelum dimasukkan dalam *Naive Bayes*. Caranya adalah [4]

1. melakukan diskretisasi pada setiap fitur kontinu dan mengganti nilai fitur kontinu tersebut dengan nilai interval diskret. Pendekatan ini dilakukan dengan mentransformasi fitur kontinu ke dalam fitur ordinal.
2. Mengasumsikan bentuk tertentu dari distribusi probabilitas untuk fitur kontinu dan memperkirakan parameter distribusi dengan data pelatihan. Distribusi Gaussian biasanya dipilih untuk merepresentasikan probabilitas bersyarat dari fitur kontinu pada sebuah kelas $P(X_i|Y)$, sedangkan distribusi Gaussian dikarakteristikkan dengan dua parameter: mean, μ, dan varian, σ^2. Untuk setiap kelas y_j, probabilitas bersyarat kelas y_j untuk fitur Xi adalah

$$P(X_i = x_i|Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} exp^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}} \quad \text{.....(2.5)}$$

2.4. Pengukuran Akurasi

Sebuah sistem yang melakukan klasifikasi diharapkan dapat melakukan klasifikasi semua set data dengan benar, tetapi tidak dapat dipungkiri bahwa kinerja suatu sistem tidak bisa 100% benar sehingga sebuah sistem klasifikasi juga harus diukur kinerjanya. Umumnya, pengukuran kinerja klasifikasi dilakukan dengan matriks konfusi (*confusion matrix*). Matrix konfusi merupakan tabel pencatat hasil kerja klasifikasi.

Table 2.1 Matriks Konfusi untuk Klasifikasi dua kelas

f_{ij}		Kelas hasil prediksi (*j*)	
		Kelas = 1	Kelas = 0
Kelas asli (*i*)	Kelas = 1	f_{11}	f_{10}
	Kelas = 0	f_{01}	f_{00}

Setiap sel f_{ij} dalam matriks menyatakan jumlah rekord/data dari kelas *i* yang hasil prediksinya masuk ke kelas *j*.

Kuantitas matriks konfusi dapat diringkas menjadi dua nilai, yaitu akurasi dan laju eror. Dengan mengetahui jumlah data yang diklasifikasikan secara benar, kita dapat mengetahui akurasi hasil prediksi, dan dengan mengetahui jumlah data yang diklasifikasikan secara salah, kita dapat mengetahui laju eror dari prediksi yang dilakukan. Dua kuantitas ini digunakan sebagai metrik kinerja klasifikasi.
Untuk menghitung akurasi digunakan formula

$$Akurasi = \frac{jumlah\ data\ yang\ diprediksi\ secara\ benar}{jumlah\ prediksi\ yang\ dilakukan} = \frac{f_{11}+f_{00}}{f_{11}+f_{10}+f_{01}+f_{00}} \quad \text{.....(2.1)}$$

Untuk menghitung laju eror (kesalahan prediksi) digunakan formula

$$Laju\ eror = \frac{jumlah\ data\ yang\ diprediksi\ secara\ salah}{jumlah\ prediksi\ yang\ dilakukan} = \frac{f_{10}+f_{01}}{f_{11}+f_{10}+f_{01}+f_{00}} \quad \text{.....(2.2)}$$

Semua algoritma klasifikasi berusaha membentuk model yang mempunyai akurasi tinggi (laju eror yang rendah). Umumnya, model yang dibangun dapat memprediksi dengan benar pada semua data yang menjadi data latihnya, tetapi ketika model berhadapan dengan data uji, barulah kinerja model dari sebuah algoritma klasifikasi ditentukan [3].

2.5. Data Preprocessing

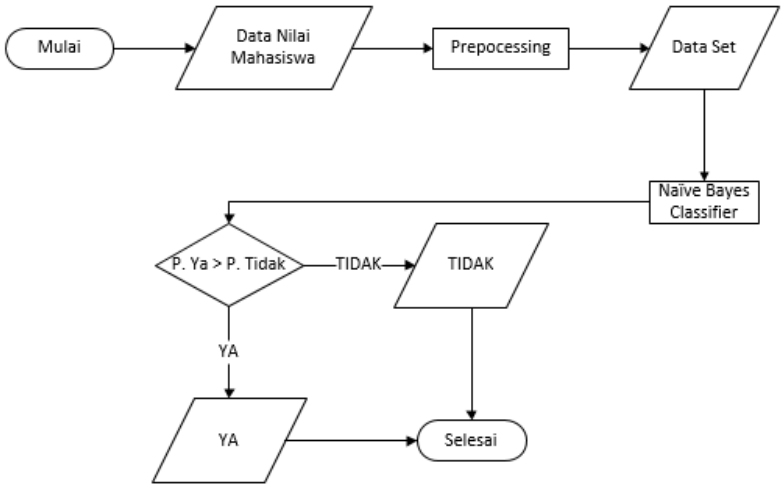
Data preprocessing adalah serangkaian proses yang dilakukan terhadap data mentah sehingga data tersebut bisa digunakan dalam proses data mining secara efisien.

1. *Data Cleaning*
   Data cleaning adalah proses pembersihan data dari atribut-atribut yang tidak akan digunakan dalam proses data mining.
   a. Dalam penelitian kali ini atribut yang akan digunakan dalam proses klasifikasi adalah kota asal, SKS, IPK, KK(Kelompok Keahlian), berapa kali ambil MK TA1 Dan TA2 dari tiap mahasiswa, sehingga atribut-atribut yang tidak berhubungan dengan keenam hal tersebut akan dibersihkan.
   b. Beberapa data yang telah tersedia masih terdapat missing value, maka missing value tersebut diisi dengan nilai-nilai yang dapat membantu proses mining.
2. *Data Summarization*
   Data summarization yaitu proses penggabungan data mahasiswa, sehingga didapatkan jumlah SKS, IPK, berapa kali mengambil mata kuliah TA1 Dan TA2 dari masing-masing mahasiswa.
3. *Data Transformation*
   Data transformaion adalah proses merubah data ke dalam bentuk yang dapat di olah dalam proses data mining. beberapa contohnya diantaranya adalah proses generalisasi dan pembuatan atribut. Atribut untuk label kelas lulus tepat waktu dan lulus tidak tepat waktu belum ada, untuk itulah atribut tersebut dibuat dari atribut tgl_keluar dari data set yang tersedia.

3. Perancangan Sistem

3.1 Gambaran Sistem Secara Umum

Secara umum sistem yang akan dibangun dapat digunakan untuk mengklasifikasikan ketepatan waktu lulus mahasiswa S1 Teknik Informatika Universitas Telkom dengan menggunakan algoritma *Naive Bayes Classifier*. Aplikasi ini dapat membandingkan data *training* dan data *testing* untuk mengetahui hasil klasifikasi.

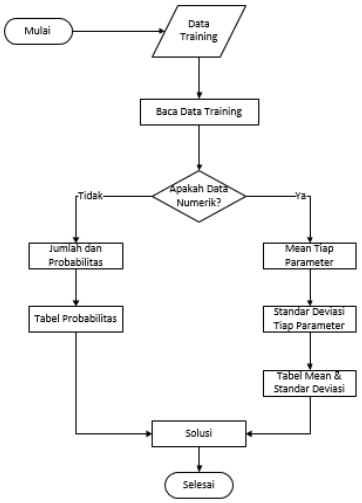


Gambar 3.1 Flowchart Gambaran Umum Sistem

Gambar 3.1 menunjukkan alur proses pada sistem yang dibangun. Diawali dengan data nilai mahasiswa yang kemudian dilakukan preprocessing terhadap data tersebut, sehingga menghasilkan data yang siap diolah. Kemudian data set tersebut, diolah untuk mengklasifikasikan ketepatan waktu lulus mahasiswa dengan menggunakan algoritma *Naive Bayes Classifier*. Pada proses *Naive Bayes Classifier*, masing-masing atribut dihitung nilai probabilitasnya, kemudian hasil probabilitas tersebut dibandingkan untuk mengklasifikasikan data.

3.2 Tahap proses Algoritma *Naive Bayes Classifier*

Untuk tahapan proses ini akan dijelaskan pada gambar 3.2:



Gambar 3.2 Flowchart Proses Algoritma *Naïve Bayes Classifier*

Gambar 3.2 menunjukkan alur pada proses algoritma *Naive Bayes Classifier*. Dimulai dengan membaca data training, kemudian system melakukan pengecekan atribut pada data, atribut bertipe numeric atau bukan. Jika atribut bukan bertipe numerik, maka atribut tersebut dihitung jumlah dan probabilitasnya. Jika atribut bertipe numerik, maka atribut tersebut dihitung mean dan standar deviasinya, kemudian dihitung probabilitasnya.

4. Pengujian Sistem dan Analisis

4.1. Pengujian Sistem

Setelah melakukan tahap implementasi, tahap selanjutnya adalah melakukan pengujian terhadap performansi sistem yang telah dibuat. Performansi sistem yang dimaksud adalah dengan mengukur nilai akurasi dari setiap skenario pengujian.

4.1.1 Tujuan dan Strategi Pengujian

Pengujian yang dilakukan bertujuan untuk mengetahui performansi algoritma *Naïve Bayes Classifier* pada titik optimalnya. Hal ini penting, sebagai acuan penelitian selanjutnya. Untuk mengetahui titik optimal algoritma, jumlah data dibagi menjadi 2 bagian, yaitu data training dan data testing.

4.1.2 Skenario Pengujian

Berdasarkan strategi pengujian diatas, jumlah perbandingan data training dan data testing dibagi menjadi beberapa bagian, dimulai dari 90% data training dan 10% data testing, hingga 10% data training dan 90% data testing. Setelah didapatkan hasil akurasi dari perbandingan tiap bagian data, didapatkan titik optimal dari algoritma *Naïve Bayes Classifier*. Pengukuran akurasi tersebut menggunakan matriks konfusi. Selanjutnya, hasil tersebut dianalisis untuk mengetahui atribut yang berpengaruh terhadap ketepatan waktu lulus.

4.1.3 Hasil Pengujian

Setelah pengujian dilakukan terhadap skenario yang telah dirancang, didapatkan hasil pengujian sebagai berikut:

Table 4.1 Detail Akurasi perbandingan data

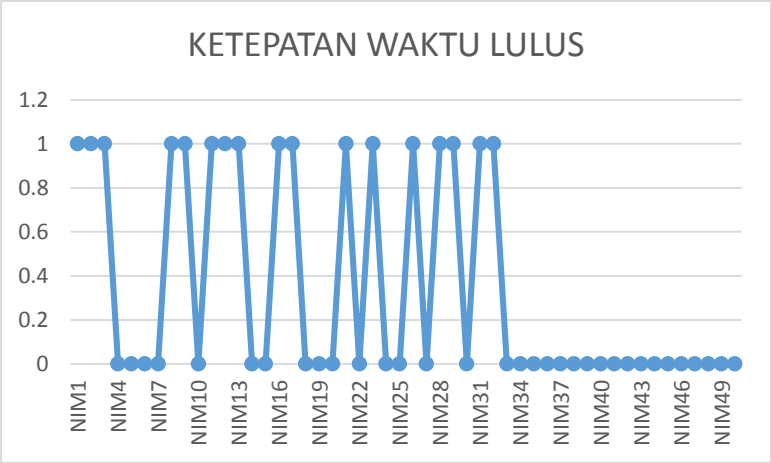
Detail Akurasi	90:10	80:20	70:30	60:40	50:50	40:60	30:70	20:80	10:90
Correctly Classified	91%	88%	84%	85%	84%	84%	82%	81%	81%
Incorrectly Classified	9%	12%	16%	15%	16%	16%	18%	19%	19%
True Positive Rate	0.75	0.710	0.706	0.702	0.671	0.682	0.651	0.656	0.657
True Negative Rate	0.975	0.951	0.898	0.895	0.904	0.900	0.892	0.876	0.868
False Positive Rate	0.025	0.049	0.102	0.105	0.096	0.100	0.108	0.124	0.132
False Negative Rate	0.25	0.290	0.294	0.298	0.329	0.318	0.349	0.344	0.343
Positive Predictive Value	0.923	0.846	0.75	0.690	0.710	0.709	0.704	0.677	0.647
Negative Predictive Value	0.279	0.253	0.298	0.234	0.231	0.236	0.25	0.258	0.245
False Discovery Rate	0.023	0.046	0.099	0.105	0.094	0.098	0.105	0.123	0.133
False Omission Rate	0.308	0.346	0.312	0.293	0.348	0.329	0.378	0.355	0.338
Accuracy	0.911	0.885	0.840	0.847	0.843	0.844	0.824	0.813	0.811

4.2. Analisis Hasil Pengujian

Berdasarkan pengujian yang telah dilakukan di dapatkan hasil sebagai berikut:

1. Analisis klasifikasi dari kecenderungan data

   Berdasarkan kecenderungan data aktual, penjelasan data ditampilkan pada gambar 4.1. Angka 1 menunjukkan bahwa mahasiswa tersebut lulus tepat waktu, sedangkan angka 0 menunjukkan bahwa mahasiswa tersebut lulus tidak tepat waktu.



Gambar 4.1 Grafik data ketepatan waktu lulus mahasiswa

Pada gambar 4.1, data menunjukkan bahwa mahasiswa yang lulus tidak tepat waktu lebih banyak daripada mahasiswa yang lulus tepat waktu. Sehingga perlu diketahui penyebab keterlambatan waktu lulus mahasiswa. Keterlambatan waktu lulus mahasiswa tersebut, dapat diklasifikasika dari atribut yang ada, yaitu SKS, IPK, TA1 dan TA2. Keterkaitan antara SKS dan ketepatan waktu lulus adalah dengan melihat dari berapa banyak SKS yang diambil. Setelah dilakukan pengujian, diketahui bahwa probabilitas ketepatan waktu lulus yang dihasilkan untuk atribut SKS tidak terlalu besar, hanya berkisar 40%. Hasil ini dapat menunjukkan bahwa jumlah SKS yang diambil tidak memiliki pengaruh yang besar terhadap ketepatan waktu lulus mahasiswa. Atribut SKS berpengaruh terhadap ketepatan waktu lulus, jika pada tiap semester, diketahui berapa jumlah SKS yang diambil dan sudah lulus. Pada atribut IPK, probabilitas ketepatan waktu lulus yang dihasilkan cukup besar, yaitu berkisar 85%. Hasil ini menunjukkan bahwa atribut IPK memiliki pengaruh yang cukup besar terhadap ketepatan waktu lulus. Namun, mahasiswa yang lulus tidak tepat waktu, tidak selalu mendapatkan IPK kurang dari 3.30, bahkan pada data, terdapat mahasiswa yang memiliki IPK sebesar 3.70 tetapi lulus tidak tepat waktu. Namun, data memang menunjukkan bahwa 85% mahasiswa yang lulus tidak tepat waktu memiliki IPK kurang dari 3.30. IPK yang relatif kecil tersebut dapat disebabkan oleh beberapa hal, diantaranya adalah nilai di setiap mata kuliah yang diambil. Sebagai contoh, pada tabel 4.2 dijabarkan mata kuliah yang diulang dari salah satu mahasiswa.

Table 4.2 Detail Mata Kuliah yang diulang

Mata Kuliah yang diulang	**Jumlah Berapa Kali mengulang**
DESAIN DAN ANALISIS ALGORITMA	3
LOGIKA MATEMATIKA	2
JARINGAN KOMPUTER	3
TEORI KOMPUTASI	2
PROBABILITAS DAN STATISTIKA	3
MANAJEMEN PROYEK TEKNOLOGI INFORMASI	3
KALKULUS II	2
KERJA PRAKTEK	2
ORGANISASI DAN ARSITEKTUR KOMPUTER	4
PRAKTIKUM SISTEM OPERASI	2
PEMROGRAMAN KOMPUTER	2
TUGAS AKHIR II	2
KECERDASAN MESIN DAN ARTIFISIAL	2
SISTEM BASISDATA	2
MATEMATIKA DISKRET	4
GRAFIKA DAN PENGOLAHAN CITRA	3

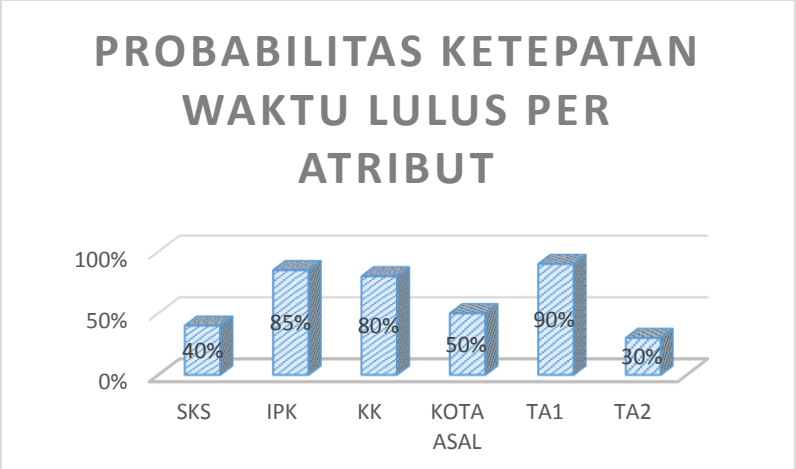
Pada tabel 4.2, terlihat bahwa mata kuliah yang diulang oleh mahasiswa yang bersangkutan relatif

banyak dan menyebabkan mahasiswa tersebut harus menambah semester untuk mengulang mata kuliah yang belum lulus tersebut. Hal ini dapat menjadi salah satu alasan ketidak tepatan waktu lulus mahasiswa.

Pada atribut KK atau Kelompok Keahlian, data menunjukkan bahwa sebanyak 384 mahasiswa memilih SIDE, 109 mahasiswa memilih TELE, dan 70 mahasiswa memilih ICM. Jumlah tersebut menunjukkan ketidakseimbangan data antara mahasiswa yang memilih KK SIDE, TELE dan ICM. Data juga menunjukkan bahwa mahasiswa yang paling banyak lulus tidak tepat waktu adalah dari KK TELE, hanya berkisar 17% mahasiswanya yang lulus tepat waktu. Sedangkan, probabilitas ketepatan waktu lulus yang dihasilkan untuk atribut KK berkisar 80%. Hasil tersebut menunjukkan bahwa atribut KK memiliki pengaruh yang cukup besar terhadap ketepatan waktu lulus.

Pada atribut Kota Asal, data menunjukkan sebanyak 226 mahasiswa berasal dari Luar Jawa, 141 mahasiswa berasal dari Jawa Barat, 95 berasal dari Jawa Tengah, 36 mahasiswa berasal dari Jawa Timur, 36 mahasiswa berasal dari DKI Jakarta, 17 mahasiswa berasal dari Banten, dan 11 mahasiswa berasal dari DI Yogyakarta. Data juga menunjukkan bahwa mahasiswa yang lulus tidak tepat waktu adalah mahasiswa yang berasal dari daerah Jawa Barat. Hanya sekitar 17% mahasiswa yang berasal dari Jawa Barat yang lulus tepat waktu. Atribut kota asal ini memiliki probabilitas ketepatan waktu lulus yang berkisar 50%. Hal tersebut menunjukkan bahwa atribuat kota asal tidak memiliki pengaruh yang besar terhadap ketepatan waktu lulus.

Atribut selanjutnya adalah atribut TA1 dan TA2. Atribut ini menunjukkan berapa kali mahasiswa mengambil mata kuliah TA1 dan TA2. Data menunjukkan mahasiswa yang mengambil mata kuliah TA1 lebih dari 2 kali, belum tentu lulus tidak tepat waktu. Sama halnya dengan atribut TA2, mahasiswa yang mengambil mata kuliah TA2 lebih dari 1 kali belum tentu juga lulus tidak tepat waktu. Sedangkan untuk probabilitas ketepatan waktu lulus, atribut TA1 memiliki probabilitas berkisar 90% dan untuk atribut TA2 memiliki probabilitas berkisar 30%. Hasil tersebut menunjukkan bahwa antara atribut TA1 dan TA2, yang memiliki pengaruh besar terhadap ketepatan waktu lulus mahasiswa adalah atribut TA1. Gambar 8.1 menunjukkan grafik probabilitas ketepatan waktu lulus per atribut.



Gambar 4.2 Grafik Probabilitas Ketepatan Waktu Lulus per Atribut

Berdasarkan analisis tiap atribut diatas, sampel hasil klasifikasi tiap atribut dapat dilihat pada tabel 4.4.

Table 4.4 Sampel Hasil Klasifikasi

NIM	KOTA ASAL	SKS	IPK	ANGKATAN	KK	TA1	TA2	TEPAT WAKTU	HASIL	CATEGORY
1C0BD5C27A358B935ED80A9B19D3DB0A	LUAR JAWA	157	3.17	2007	SIDE	5	2	TIDAK	TIDAK	RED
1C0BD5C27A358B935D23E5CA4346DCC8	LUAR JAWA	146	2.94	2007	SIDE	4	6	TIDAK	TIDAK	RED
1C0BD5C27A358B935CD94BDF16AABA06	LUAR JAWA	146	2.92	2007	SIDE	8	2	TIDAK	TIDAK	RED

NIM	KOTA ASAL	SKS	IPK	ANGKATAN	KK	TA1	TA2	TEPAT WAKTU	HASIL	CATEGORY
1C0BD5C27A358B93529BEB413B722266	LUAR JAWA	148	2.84	2007	SIDE	7	4	TIDAK	TIDAK	RED
1C0BD5C27A358B934BB0CDAE4C70E4E2	JAWA BARAT	149	3.01	2007	TELE	3	4	TIDAK	TIDAK	RED
1C0BD5C27A358B9346AFDA3E9FCCAD11	BANTEN	149	2.72	2007	SIDE	3	5	TIDAK	TIDAK	RED
1C0BD5C27A358B93456134DD273CDD65	JAWA TENGAH	145	2.99	2007	SIDE	5	4	TIDAK	TIDAK	RED
1C0BD5C27A358B93431AF3676935B43C	JAWA TENGAH	146	2.78	2007	SIDE	4	5	TIDAK	TIDAK	RED
1C0BD5C27A358B9340B2E4FBB5E32017	LUAR JAWA	146	2.73	2007	TELE	3	3	TIDAK	TIDAK	RED
1C0BD5C27A358B9337621BFDD8CE9726	JAWA BARAT	155	2.6	2007	SIDE	3	5	TIDAK	TIDAK	RED
1C0BD5C27A358B9336DB6BCC4D60ED53	JAWA BARAT	146	2.55	2007	ICM	5	4	TIDAK	TIDAK	RED
1C0BD5C27A358B9336CB094DA6B5DAD8	LUAR JAWA	148	2.55	2007	SIDE	2	2	TIDAK	TIDAK	RED
1C0BD5C27A358B9334C9CC4C3273CC53	JAWA BARAT	152	3.01	2007	SIDE	4	3	TIDAK	TIDAK	RED
1C0BD5C27A358B93345617A71D9AD1D3	JAWA BARAT	149	2.67	2007	ICM	4	3	TIDAK	TIDAK	RED
1C0BD5C27A358B93336FA7432DF66FFB	JAWA BARAT	146	3.7	2007	SIDE	3	7	TIDAK	TIDAK	RED
1C0BD5C27A358B93303741B207F5CD7B	DKI JAKARTA	147	2.66	2007	ICM	6	3	TIDAK	TIDAK	RED
1C0BD5C27A358B932E26D4428DE8A99D	JAWA BARAT	147	2.81	2007	ICM	6	3	TIDAK	TIDAK	RED
1C0BD5C27A358B93147FF15F859E2BFF	LUAR JAWA	154	2.6	2007	TELE	2	2	TIDAK	TIDAK	RED
099298D2DAE2C261F1EC88F170F35CEF	JAWA BARAT	145	3.26	2008	TELE	2	3	TIDAK	TIDAK	RED
099298D2DAE2C261B347E3FE19747F22	LUAR JAWA	145	3.13	2008	SIDE	2	3	TIDAK	TIDAK	RED
099298D2DAE2C261AB50268F445879F9	BANTEN	145	3.34	2008	SIDE	1	1	YA	YA	YELLOW
0987F678C1322D37FE7A4F7497A0338F	JAWA BARAT	145	3.19	2008	ICM	1	1	YA	TIDAK	RED

Pada tabel 4.4, atribut Hasil merupakan hasil dari klasifikasi semua atribut. Sedangkan atribut Category merupakan hasil probabilitas kelas Ya. Category GREEN memiliki P(Ya) > 70%, Category YELLOW memiliki P(Ya) dengan kisaran antara 40% - 70%, dan untuk Category RED memiliki P(Ya) < 40%.

2. Analisis terhadap pembagian data

   Berdasarkan skenario pengujian, data yang telah dibagi menjadi 2 bagian, yaitu data training dan data testing, dimulai dari 90% data training dan 10% data testing hingga 10% data training dan 90% data testing menghasilkan akurasi yang cukup bagus, yaitu diatas 80% dengan penjabaran 91% untuk data

training:data testing sebesar 90:10, 88% untuk 80:10, 84% untuk 70:30, 85% untuk 60:40, 84% untuk 50:50, 84% untuk 40:60, 82% untuk 30:70, 81% untuk 20:80, 81% untuk 10:90. Hasil tersebut menunjukkan semakin sedikit data training, akurasi yang dihasilkan juga semakin menurun, sedangkan data testing semakin banyak jumlahnya.

Pembagian data menjadi 2 bagian dilakukan sesuai dengan tujuan dan skenario pengujian. Hasil yang didapatkan tersebut menunjukkan bahwa titik optimum dari algoritma *Naive Bayes Classifier* berada pada pembagian 90% data training dan 10% data testing.

3. Analisis terhadap matriks konfusi

   Berdasarkan hasil pengujian yang telah dilakukan, pembagian data yang memiliki akurasi yang paling tinggi adalah perbandingan 90% data training dan 10% data testing dengan *correctly classified* sebesar 86%. *Correctly classified* adalah prosentasi jumlah kelas yang diprediksi sesuai dengan kelas aktual. Dengan detail akurasi *true positive rate* (*sensitivity*) sebesar 0.895, *true negative rate* (*specificity*) sebesar 0.838, *positive predictive value* (*precision*) sebesar 0.739, *accuracy* sebesar 0.857. *Sensitivity* digunakan untuk membandingkan jumlah *true positive* terhadap jumlah *record* yang positif sedangkan *Specifity*, *precision* adalah perbandingan jumlah *true negative* terhadap jumlah *record* yang negatif. *Accuracy* yang menghasilkan nilai dengan kisaran 0.80 – 0.90 menunjukkan bahwa algoritma *Naïve Bayes Classifier* termasuk ke dalam *good classification.*

5. Kesimpulan dan Saran

5.1. Kesimpulan

Berdasarkan hasil pengujian, maka kesimpulan yang didapatkan adalah sebagai berikut:

1. Algoritma *Naive Bayes Classifier* dapat diimplementasikan untuk klasifikasi ketepatan waktu lulus mahasiswa dan menghasilkan prosentase klasifikasi yang cukup bagus, lebih dari 80% dari skenario pengujian yang dilakukan.
2. Performansi algoritma *Naïve Bayes Classifier* menunjukkan hasil yang cukup bagus dari pengujian yang telah dilakukan. Dari pengujian tersebut, pengujian yang menghasilkan *correctly classified* yang yang paling bagus adalah pengujian dengan pembagian data dengan perbandingan 90% data training dan 10% data testing, yaitu mencapai 86%. Angka tersebut menunjukkan bahwa algoritma *Naïve Bayes Classifier* termasuk ke dalam *good classification*.
3. Faktor yang mempengaruhi ketepatan waktu lulus berdasarkan atribut yang digunakan adalah atribut TA1 karena memiliki probabilitas yang paling besar dibandingkan atribut yang lain, yaitu sebesar 90%.

5.2. Saran

Berikut ini saran penulis tentang pengembangan yang dapat dilakukan pada penelitian klasifikasi ketepatan waktu lulus mahasiswa:

1. Selain atribut yang digunakan oleh penulis dalam melakukan penelitian, akan lebih baik jika menambahkan atribut lain berupa atribut non-akademik, seperti hobi, jumlah organisasi yang diikuti, mengikuti kegiatan lab atau tidak, pernah cuti atau tidak, dan atribut lain yang dapat mempengaruhi ketepatan waktu lulus.
2. Terdapat beberapa algoritma yang dapat digunakan dalam klasifikasi, sehingga penulis menyarankan untuk menggunakan algoritma lain atau menggabungkan dua atau lebih algoritma dalam klasifikasi agar mendapatkan hasil yang lebih baik.

6. Daftar Pustaka

[1] A. G. Mabrur and R. Lubis, "Penerapan Data Mining untuk Memprediksi Kriteria Nasabah Kredit," *Jurnal Komputer dan Informatika (KOMPUTA)*, vol. 1, pp. 53-57, 2012.

[2] B. Santosa, "Data Mining Teknik Pemanfaata Data untuk Keperluan Bisnis," Yogyakarta: Graha Ilmu, 2007.

[3] D.T. Larose, "Discovering Knowledge in Data," New Jersey: John Willey & Sons, 2005.

[4] E. Prasetyo, "Data Mining: Konsep dan Aplikasi menggunakan Matlab," Yogyakarta: Andi, 2012.

[5] E.T. Luthfi dan Kusrini. "Algoritma Data Mining," Yogyakarta: Andi, 2009.

[6] G.W. Dekker, "Predicting Students Dropout: A Case Study," In International Conference on Data Mining, Cordoba, Spain, 41-50, 2009.

[7] I. Tahyudin, E. Utami, dan A. Amborowati. "Comparing Clasification Algorithm Of Data Mining to Predict thr Graduation Students on Time," Purwokerto: Jurnal Sistem Informasi. STIMIK AMIKOM, 2013.

[8] J. Han dan M. Kamber, "Data Mining: Concepts and Technique," San Fransisco : Morgan Kaufmann

Publishers, 2006.
[9] J. Lin and J. Yu, "Weighted Naive Bayes Classification Algorithm Based on Particle Swarm Optimization," *IEEE,* pp. 444-447, 2011.
[10] K. Hastuti, "Comparative Analysisi of Classification Algorithm for Data Mining Prediction of Non-Active Students," Semarang: Seminar Nasional Applied Information and Communication 2012 (SEMANTIK 2012).
[11] M. J. Islam, Q. M. J. Wu, M. Ahmadi and M. A. Sid-Ahmed, "Investigating the Performance of Naive-Bayes Classifiers and K- Nearest Neighbor Classifiers," *International Conference on Convergence Information Technology*, pp. 1541-1546, 2007.
[12] M. Ridwan, H. Suyono, dan M. Sarosa, "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma *Naïve Bayes Classifier*," Malang: Jurnal EECCIS Universitas Brawijaya Vol.7, No.1, Juni 2013.
[13] P. Nancy, Ramani dan R. Geetha. 2011. "A Comparation on Performance of Data Mining Algorithms in Classification of Social Network Data," International Journal of Computer Aplications (0975 – 8887) Vol. 32 No. 8 October 2011.
[14] S. Kotsiantris, "Educational Data Mining: A Case Study for Predicting Dropout-Prone Students," Int. J of Knowledge Engineering and Soft Data Paradigms, Vol. x, 2010.
[15] S.M. Suhartinah dan Ernastuti, "Graduation Prediction of Guna Darma University Students Using Algorithm Naïve Bayes and C4.5 Algorithm," Jakarta: Jurnal Magister Sistem Informasi. Universitas Guna Darma, 2010
[16] Zlatco J. Kovacic, "Early Prediction of Student Success: Mining Students Enrolment Data," In Proceedings of Informing Science & IT Education Conference (InSITE), 2010.