Klasifikasi Identifikasi Faktor Penyebab Ketidaktepatan Masa Lulus Mahasiswa dengan Metode *Naïve Bayes Classifier*

Budanis Dwi Meilani[1], Sandra Wahyudiana[2], Anggi Yhurinda Perdana Putri[3], Adib Pakarbudi[4]

Fakultas Teknik Elektro dan Teknologi Informasi, Jurusan Sistem Informasi[1,2,3,4]
Email: budanis@itats.ac.id

ABSTRACT

Higher education is an education unit whose job is to produce quality graduates. Therefore the institution must prove its quality and ability to compete in the field of education. Related to the quality of tertiary institutions, this research focuses on the inaccuracy of the graduation period of students which still mostly occurs in the Information Systems Department of the Adhitama Institute of Technology Surabaya. The average Information Systems Department student graduates on time annually only reaches 45% while the range of graduation numbers not on time reaches 55%. So it is necessary to identify the factors causing the inaccuracy of the student's graduation period. The purpose of the study is to identify the causal factors that influence the inaccuracy of the student's graduation period. This research uses Naïve Bayes Classifier algorithm which is an algorithm of data mining techniques by applying Bayes theory for classification in data processing. Based on the tests conducted, it was found that the factors that had a big influence in determining the classification of the inaccuracy of the students' graduation period were, GPA <3.00, male gender, GPA ≥ 3.00, did not have a description of the title of thesis from the beginning, and repeated courses. These factors can be used as evaluation materials for study program managers. Testing the Naïve Bayes Classifier algorithm to determine the accuracy of the graduation period of students shows that from 79 attempts the accuracy of 91.13% was accurate with an error rate of 8.86%.

Keyword: *Data mining, Accuracy in Graduation, Naïve Bayes Classifier.*

ABSTRAK

Perguruan tinggi merupakan satuan pendidikan yang bertugas untuk mencetak lulusan- lulusan yang berkualitas. Maka dari itu institusi tersebut harus membuktikan kualitas dan kemampuannya dalam bersaing di bidang pendidikan. Berkaitan dengan kualitas perguruan tinggi, penelitian ini berfokus pada ketidaktepatan masa lulus mahasiswa yang masih sebagian besar terjadi di lingkungan Jurusan Sistem Informasi Institut Teknologi Adhitama Surabaya. Rata-rata mahasiswa Jurusan Sistem Informasi lulus tepat waktu pertahunnya hanya mencapai 45% sedangkan kisaran angka lulus tidak tepat waktu mencapai 55%. Sehingga perlu dilakukan identifikasi faktor penyebab ketidaktepatan masa lulus mahasiswa. Tujuan dari penelitian adalah untuk mengidentifikasi faktor penyebab yang berpengaruh pada ketidaktepatan masa lulus mahasiswa. Penelitian ini menggunakan algoritma *Naïve Bayes Classifier* yang merupakan algoritma teknik *data mining* dengan menerapkan teori *Bayes* untuk klasifikasi dalam mengolah data. Berdasarkan pengujian yang dilakukan diperoleh bahwa faktor yang berpengaruh besar dalam penentuan klasifikasi ketidaktepatan masa lulus mahasiswa yaitu, IPK < 3.00, Jenis kelamin laki-laki, IPK ≥ 3.00, Tidak mempunyai gambaran judul skripsi dari awal, dan Banyak mengulang mata kuliah. Faktor-faktor tersebut dapat digunakan sebagai bahan evaluasi bagi pengelola program studi. Pengujian algoritma *Naïve Bayes Classifier* untuk menentukan ketepatan masa lulus mahasiswa menunjukkan dari 79 kali percobaan mencapai hasil akurasi 91.13% akurat dengan laju *error* 8.86%.

Kata kunci: *Data mining*, Ketepatan Masa Lulus, *Naïve Bayes Classifier*.

PENDAHULUAN

Perguruan Tinggi merupakan satuan pendidikan penyelenggara pendidikan tinggi, dimana persaingan yang sangat ketat di lingkungan pendidikan membuat perguruan tinggi harus terus meningkatkan kualitasnya.[1] Banyak faktor yang dapat mempengaruhi tingkat kualitas sebuah institusi Perguruan Tinggi salah satu faktor utamanya adalah faktor sumber daya manusia

(SDM), faktor ini sangat mempengaruhi berhasil tidaknya sebuah institusi. Selain sumber daya manusia (SDM), sarana, serta prasarana yang digunakan, juga harus dimanfaatkan secara maksimal untuk menunjang kualitas Perguruan Tinggi. Sistem informasi merupakan contoh sumber daya yang dapat digunakan guna meningkatkan kemampuan dan daya saing perguruan tinggi [2]. Hal ini tidak lepas dari peranan penting mahasiswa sebagai struktur pendidikan bagi sebuah institusi [3]. Dengan mencetak lulusan-lulusan yang berkualitas dan siap bersaing di dunia kerja, maka institusi tersebut telah membuktikan kualitas dan kemampuannya dalam bersaing di bidang pendidikan. Ada berbagai jurusan keilmuan yang terdapat pada Institut Teknologi Adhi Tama Surabaya. Jurusan keilmuan tersebut memiliki standar, dengan track recordnya masing-masing. Konsentrasi yang akan diangkat sebagai topik penelitian disini adalah Jurusan Sistem Informasi. Dimana peneliti menemukan sebuah permasalahan yang umum dan rentan terjadi di kalangan mahasiswa Jurusan Sistem Informasi mengenai ketidaktepatan masa lulus mahasiswa Jurusan Sistem Informasi. Berdasarkan data yang telah diperoleh dari data independen Jurusan Sistem Informasi menemukan bahwa rata-rata kelulusan mahasiswa yang lulus tepat waktu pertahun hanya mencapai pada kisaran 45%, sedangkan rata-rata mahasiswa yang lulus tidak tepat waktu mencapai kisaran angka 55% jika dibandingkan dengan angka kelulusan mahasiswa yang tepat waktu, selisih mahasiswa yang lulus tidak tepat waktu mencapai 10%. Dari sinilah peneliti mencoba melakukan penelitian untuk mengidentifikasi faktor penyebab ketidaktepatan masa lulus mahasiswa menggunakan metode *Naive Bayes Classification*.*Naïve Bayes Classifier* merupakan salah satu algoritma dalam teknik *data mining* yang menerapkan teori Bayes dalam klasifikasi. Teorema *bayes* sendiri merupakan pendekatan statistik fundamental dalam pengenalan pola (*pattern recognition*). *Naïve Bayes* didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional, saling bebas jika diberikan nilai output [4].

TINJAUAN PUSTAKA

Pengertian *Data Mining*

Mendifinisikan *data mining* sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar[5]. *Data mining* juga dapat diartikan sebagai pengekstrakan informasi baru yang diambil dari bongkahan data besar yang membantu dalam pengambilan keputusan. Istilah *data mining* kadang disebut juga *knowledge discovery* [6]. Teknik yang diciptakan dari *data mining* adalah sebuah proses untuk menemukan hubungan, pola dan tren baru yang bermakna dengan menyaring data yang sangat besar, yang tersimpan dalam penyimpanan basis data, menggunakan teknik pengenalan pola seperti teknik Statistik dan Matematika [7]. Tujuan dari terapan *data mining* sendiri yaitu untuk dapat mengetahui pola universal dari data-data yang ada [8].Agar kemudian dapat dimanfaatkan untuk mendukung segala kegiatan operasional sebuah instansi atau perusahaan tertentu, untuk mencapai semua target yang diinginkan.

Teori Probabilitas

Dimisalkan sebuah peristiwa E dapat terjadi sebanyak n kali diantara N peristiwa yang saling eksklusif (saling asing/terjadinya peristiwa yang satu mencegah terjadinya peristiwa yang lain)dan masing-masing terjadi dengan kesempatan yang sama. Maka probabilitas terjadinya peristiwa E adalah ;

$$P(E) = \frac{n}{N} \qquad dengan\ batas - batas : 0 \leq P(E) \leq 1 \quad \ldots\ldots (1)$$

Jika P(E) = 0, maka diartikan peristiwa E pasti terjadi, sedangkan jika P(E) = 1, maka diartikan bahwa peristiwa E pasti terjadi. Apabila $\overline{E}$ menyatakan bukan peristiwa E, maka diperoleh :

$$P(\bar{E}) = 1 - P(E) \text{ atau berlaku hubungan }, P(E) + P(\bar{E}) = 1 \quad \text{..... (2)}$$

***Naïve Bayes* untuk Klasifikasi**

Hubungan antara *Naïve Bayes* dengan klasifikasi, korelasi hipotesis, dan bukti dengan klasifikasi adalah bahwa hipotesis dalam teorema *Bayes* merupakan label kelas yang menjadi target pemetaan dalam klasifikasi, sedangkan bukti merupakan atribut-atribut yang menjadi masukan dalam model klasifikasi. Jika X adalah vector masukan yang berisi atribut dan Y adalah label kelas, *Naïve Bayes* dituliskan dengan P(Y|X). Notasi tersebut mempunyai arti bahwa probabilitas label kelas Y didapatkan setelah atribut-atribut X diamati. Notasi ini disebut juga probabilitas akhir (*posterior probability*) untuk Y, sedangkan P(Y) disebut probabilitas awal (*prior probability*) Y.

Selama proses *training* harus dilakukan pembelajaran probabilitas akhir (P(Y|X) pada model untuk setiap kombinasi X dan Y berdasarkan informasi yang didapat dari data *training*. Dengan membangun model tersebut, suatu data uji X' dapat diklasifikasikan dengan mencari nilai Y' dengan memaksimalkan nilai P(Y'|X') yang diperoleh. Formula untuk *Naïve Bayes* untuk klasifikasi adalah ;

$$P(Y \mid X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i \mid Y)}{P(X)} \quad \text{.....(3-a)}$$

$$\prod_{i=1}^{q} P(X_i|Y) \quad \text{..... (3-b)}$$

P(Y|X) adalah probabilitas data dengan vektor X pada kelas Y.
P(Y) adalah probabilitas awal kelas Y.
Persamaan 3-b adalah probabilitas independen kelas Y dari semua atribut dalam vektor X.
Nilai p(X) selalu tetap, sehingga dalam perhitungan prediksi (bagian pembilang) dari persamaan 3-a merupakan hasil pemilihan dengan nilai terbesar.
Sementara probabilitas independen (persamaan 3-b) tersebut merupakan pengaruh semua atribut dari data terhadap setiap kelas Y, yang dinotasikan dengan ;

$$P(x \mid y = y) = \prod_{i=1}^{q} P(X_i \mid y = y) \quad \text{.....(4)}$$

Setiap set atribut $X = \{X_1, X_2, X_3 \ldots\ldots, X_q\}$ terdiri atas *q atribut* (*q dimensi*). Namun untuk atribut dengan tipe *numeric* (kontinu) ada perlakuan khusus sebelum dimasukkan dalam *Naïve Bayes*. Caranya adalah ;

1. Melakukan diskretisasi pada setiap atribut kontinu dan mengganti nilai atribut kontinu tersebut dengan nilai interval diskret. Pendekatan ini dilakukan dengan mentransformasi atribut kontinu ke dalam atribut ordinal.
2. Mengasumsikan bentuk tertentu dari distribusi probabilitas untuk fitur kontinu dan memperkirakan parameter distribusi dengan data *training*. Ditribusi *Gussian* biasanya dipilih untuk merepresentasikan probabilitas bersyarat dari atribut kontinu pada sebuah kelas $P(X_i|Y)$, sedangkan distribusi *Gussian* dikarakteristikkan dengan dua parameter : *mean* , μ , dan varian, σ². Untuk setiap kelas y_j , probabilitas bersyarat kelas y_j untuk fitur X_i adalah

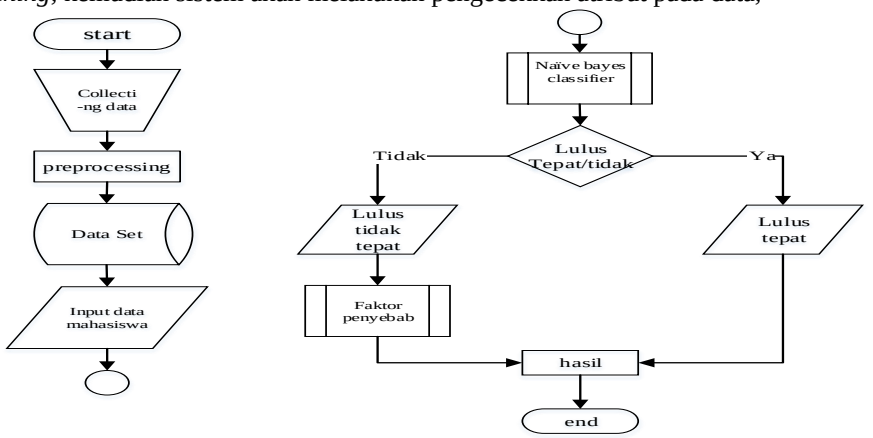
$$P(x_i = x_i \mid y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} \exp^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}} \quad \text{.....(5)}$$

Parameter μ_{ij} bisa didapat dari *mean* sampel $X_i\,(\bar{x})$ dari semua data *training* yang menjadi milik kelas y_j, sedangkan σ_{ij} dapat diperkirakan dari varian sampel (s^2) dari data *training*.

METODE

Analisa dan gambaran sistem secara umum

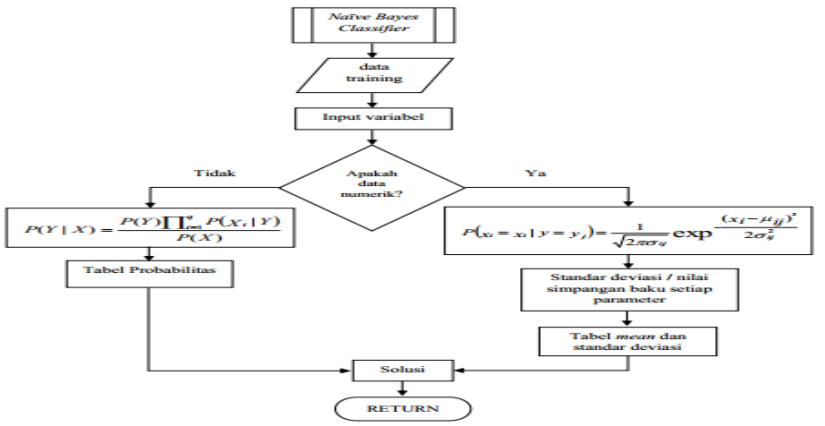
Aplikasi ini dapat membandingkan *data training* dan *data testing* untuk mengetahui hasil klasifikasi. Gambar 1 menunjukkan alur proses pada sistem yang akan dibangun. Pada tahap ini akan dijelaskan alur proses algoritma *Naïve Bayes Classifier*. Diawali dengan membaca data *training*, kemudian sistem akan melakukan pengecekkan atribut pada data,



Gambar 1. Flowchart Analisa Gambaran Sistem Secara Umum

Tahap proses Algoritma Naïve Bayes Classifier

Pada gambar 2 berikut merupakan penjelasan tahap proses alur Algoritma *Naïve Bayes Classifier*.



Gambar 2 *Flowchart* alur proses algoritma *Naive Bayes Classifier*

Data yang dipakai adalah ditunjukkan dalam tabel 1, yaitu Tabel Data Kriteria

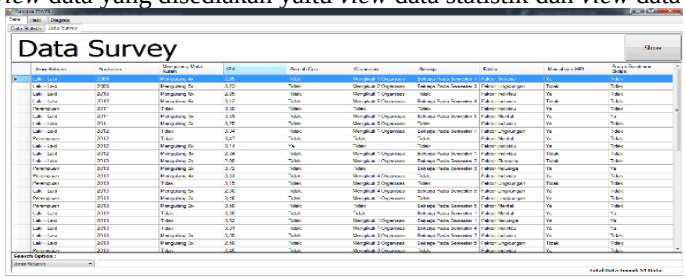
Tabel 1. Data Kriteria

NO	KRITERIA
1	Jenis Kelamin
2	Kriteria mengulang mata kuliah, apakah mahasiswa penah mengulang mata kuliah atau tidak.
3	Kriteria Indeks Prestasi Kumulatif >0 atau IPK< 4.00
4	Kriteria pernah melakukan cuti atau tidak
5	Kriteria berapa jumlah organisasi yang diikuti oleh mahasiswa
6	Kriteria mulai bekerja apakah mahasiswa tersebut kuliah sambil bekerja atau tidak dimulai semester berapa.
7	Kriteria memahami mata kuliah MPI (Metode Penelitian Ilmiah) atau tidak
8	Kriteria faktor lain yang mungkin berpengaruh, apakah mahasiswa tersebut mengalami faktor lain yang berpengaruh selain dari kriteria yang sudah di jabarkan sebelumnya, faktor lain yang berpengaruh meliputi faktor individu, ekonomi, lingkungan dll.
9	Mempunyai gambaran skripsi dari awal atau tidak, apakah mahasiswa tersebut mempunyai gambaran judul skripsi dari awal atau tidak.
10	Setelah data diuji dengan tahapan proses kriteria-kriteria yang sudah ditentukan, maka akan diperoleh hasil dan proses akan berhenti.

HASIL DAN PEMBAHASAN

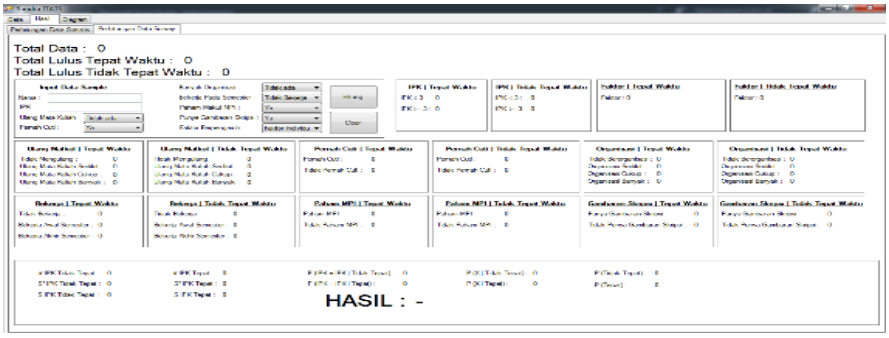
Implementasi *User Interface*

Pada halaman view data ini pengguna dapat memilih untuk melihat data yang diinginkan berikut ada dua *view* data yang disediakan yaitu *view* data statistik dan *view* data survei.



Gambar 3. *Layout view* data

Pada view hasil akan ditampilkan form isian data untuk kemudian dihitung menggunakan metode *Naïve Bayes Classifier*. Berikut *layout* form isian data *view* hasil ;



Gambar 4. *Layout view* hasil

Hasil pengujian data menunjukkan bahwa nilai kesesuaian pada variabel kategori lulus dan variabel *Bayes* mempunyai nilai *Sesuai* sejumlah 72 data dan *Tidak Sesuai* sejumlah 7 data.

Hasil yang diperoleh yaitu sebesar 91.13% nilai akurasi Metode *Naïve bayes Classifier* dinyatakan akurat, dan ketidaktepatan akurasinya sebesar 8.86% dinyatakan *error*. Jadi dari data yang telah diproses dapat dinyatakan bahwa nilai hasil akurasi yang di dapatkan 91.13% akurat atau benar. Dan untuk hasil nilai ketidaktepatan 8.86% dinyatakan *error* atau salah.

Faktor yang mempunyai pengaruh besar dalam ketidaktepatan masa lulus mahasiswa yaitu faktor Indeks Perestasi Kumulatif < 3.00 sebesar 100%, dan pengaruh terbesar kedua adalah jenis kelamin laki-lakisebesar 77%, sedangkan pengaruh ketiga terbesar yaitu Indeks Prestasi Kumulatif ≥ 3.00 dengan nilai 63%, faktor penyebab yang selanjutnya mempunyai pengaruh 49.02% adalah tidak mempunyai gambaran judul skripsi dari awal, kemudian 47.60% dipengaruhi oleh jumlah pengulangan mata kuliah, dan sebesar 45.10% dipengaruhi oleh faktor bekerja sambil kuliah.

KESIMPULAN

Hasil yang diperoleh dari pengujian yang telah dilakukan, dari 79 kali percobaan menyatakan bahwa nilai akurasi dari penggunaan Metode *Naïve Bayes Classifier* mencapai 91.13%, dengan laju error 8.86%. Identifikasi klasifikasi faktor penyebab dari keseluruhan variabel yang telah diolah menunjukkan bahwa ketidaktepatan masa lulus mahasiswa dipengaruhi oleh faktor 77% jenis kelamin laki-laki, 47.60% jumlah mengulang mata kuliah, 63%IPK ≥ 3.00, 49.02% tidakmempunyai gambaran judul skripsi dari awal, 45.10% bekerja sambil kuliah, 21.57% tidak memahami Metode Penelitian Ilmiah,7.84% jumlah organisasi yang diikuti,1.96% faktor lain yang berpengaruh.

DAFTAR PUSTAKA

[1] Kementrian Pendidikan dan Kebudayaan RI. (2013). *Standar Nasional Pendidikan Tinggi*. Direktorat Jenderal Pendidikan Tinggi dan Badan Standar Nasional Pendidikan. Jakarta.

[2] Fiyastantyo G. (2009). *Perbandingan Kinerja Metode Klasifikasi Data Mining Menggunakan Naïve Bayes dan Algoritma C4.5 Untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa.*

[3] Quardil, M.N., & Kalyankar, N.V. (2010). Drop Out Feature of Student Data for Academic Performance Using Decision Tree techniques. *Global Journal of Computer Science.*

[4] Ridwan M.,Suyono H., & Sarosa M., (2013).Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa menggunakan Algoritma Naïve Bayes Classifier.*Jurnal EECCIS vol.7 no.1 Juni 2013.*

[5] Tan, P. *et al.*(2006).*introduction to Data Mining.*Boston:Pearson Education.

[6] Prasetyo E., .(2012). *Data Mining Konsep dan Aplikasi Menggunakan MATALAB.*Edisi 1.Andi.Yogyakarta.

[7] Larose, D.T. (2005). *Discovering Knowlage in Databases*. New Jersey:Jhon Willey and Sons Inc.

[8] Saefulloh A,. Moedjiono,.(2013). Penerapan Metode Klasifikasi Data Mining Untuk Prediksi Kelulusan Tepat Waktu.*InfoSys Journal, volume 2, nomor 1, halaman : 41-54*. Jakarta.