Perancangan Model Prediksi Kelulusan Mahasiswa Tepat Waktu pada UIN Raden Fatah

Gusmelia Testiana, M. Kom
gusmeliatestiana_uin@radenfatah.ac.id

Sistem Informasi, Fakultas Sains dan Teknologi
Universitas Islam Negeri Raden Fatah Palembang

Abstract : Associated with one of the functions of Higher Education in education, teaching and subject this becomes one of the items of accreditation that is timely graduation for students. So far, UIN Raden Fatah Palembang does not yet have timely graduation prediction patterns as a reference to predict the number of graduated on time. Something very unfortunate if the data so large is not used to explore what information is contained therein. In addition, so far there is the assumption that to predict the graduation rate on time simply by looking at the IPK data only. Departing from the above problems then conducted this research is to perform data mining on student data UIN Raden Fatah Palembang to get information about the timely graduation of students UIN Raden Fatah Palembang.

Keywords: graduation, data mining

Abstrak : Terkait dengan salah satu fungsi dari Perguruan Tinggi dalam pendidikan, pengajaran dan perihal ini menjadi salah satu butir akreditasi yaitu kelulusan tepat waktu bagi mahasiswa. Selama ini UIN Raden Fatah Palembang belum memiliki pola–pola prediksi kelulusan tepat waktu sebagai acuan untuk memprediksi jumlah lulus tepat waktu. Sesuatu yang sangat disayangkan jika data yang begitu besar tidak dimanfaatkan untuk digali informasi apa yang terdapat didalamnya. Selain itu, selama ini ada anggapan bahwa untuk memprediksi tingkat kelulusan tepat waktu cukup dengan melihat data IPK saja. Berangkat dari permasalahan di atas maka dilakukanlah penelitian ini yaitu untuk melakukan *data mining* terhadap data mahasiswa UIN Raden Fatah Palembang sehingga didapatkan informasi mengenai kelulusan tepat waktu dari mahasiswa UIN Raden Fatah Palembang.

Kata Kunci: kelulusan, data mining

1. PENDAHULUAN

UIN Raden Fatah Palembang memiliki data yang besar, yaitu terdapat database untuk Sistem Informasi Akademik (SIMAK Online) yang dijadikan pasokan data untuk mengelola Indeks Prestasi Semester (IPS) dan Indeks Prestasi Kumulatif Mahasiswa (IPK), IPS dan IPK hanya berupa data belum memberikan informasi yang sangat bermanfaat selain hanya untuk mengetahui nilai mahasiswa per semester, selama ini untuk memperkirakan kelulusan tepat waktu mahasiswa dengan melihat pengaruh dari IPS dan IPK hanya berupa perkiraan.

Dewasa ini *data mining* berkembang digunakan untuk menyelesaikan masalah menyangkut pendidikan. Beberapa penelitian

terkait *data mining* digunakan sebagai upaya meningkatkan akreditasi perguruan tinggi, oleh karena itu untuk mengatasi masalah tersebut diterapkan *data mining* dengan algoritma *Naïve Bayes* untuk mencari karakteristik mahasiswa yang berkesempatan untuk lulus tepat waktu.

alam melakukan prediksi kelulusan tepat waktu mahasiswa UIN Raden Fatah Palembang terdapat berbagai macam masalah yaitu :

ก. Belum adanya metode yang digunakan di UIN Raden Fatah dalam memprediksi kelulusan mahasiswa.

ข. Dosen Pembimbing Akademik (PA) melakukan prediksi hanya berdasarkan IPK mahasiswa tersebut sehingga tidak bisa mengetahui apakah mahasiswanya akan lulus tepat waktu atau terlambat.

Penelitian ini bertujuan untuk :

ก. Menentukan algoritma yang dapat memprediksi kelulusan setiap mahasiswa UIN Raden Fatah Palembang.

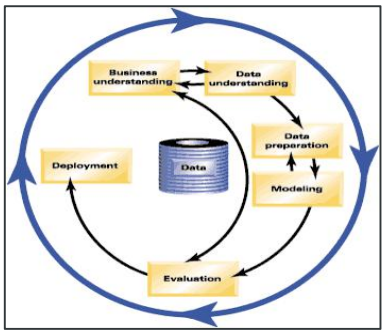
ข. Mengetahui atribut-atribut apa saja yang dibutuhkan untuk menentukan kelulusan mahasiswa tepat waktu pada UIN Raden Fatah Palembang.

ค. Menampilkan data prediksi hasil data mining untuk kelulusan tepat waktu.

2. METODOLOGI PENELITIAN

Metodologi atau tahapan penelitian diperlukan sebagai kerangka dan panduan proses penelitian, sehingga rangkaian proses penelitian dapat dilakukan secara terarah, teratur dan sistematis.

Penelitian ini didesain dengan menggunakan model CRISP-DM (*Cross Industy Standard Process for Data Mining*), dalam metode ini terdapat 6 tahapan [Larose, 2005].



Gambar 1. Tahap CRISP-DM (*Cross Industy Standard Process for Data Mining*)

Proses data mining berdasarkan CRISP-DM terdiri dari 6 fase. Yaitu:

1. *Business/Research Understanding Phase*
2. *Data Understanding Phase* (Fase Pemahaman Data)
3. *Data Preparation Phase* (Fase Pengolahan Data)
4. *Modeling Phase* (Fase Pemodelan)
5. *Evaluation Phase* (Fase Evaluasi)
6. *Deployment Phase* (Fase Penyebaran).

2.1 Business/Research Understanding Phase

Data yang diperoleh dari database SIMAK pada UIN Raden Fatah Palembang,

ternyata selama ini belum pernah dilakukan penggalian kekayaan terhadap data tersebut. Data tersebut sudah cukup banyak dan dilakukan *data mining* untuk menghasilkan kelulusan tepat waktu tentunya akan sangat bermanfaat.

Diketahui bahwa UIN Raden Fatah Palembang belum memanfaatkan database tersebut, dan dalam menentukan prediksi kelulusan masih menggunakan metode asumsi-asumsi dengan tingkat subyektifitas yang tinggi. Sampai saat ini, masih belum ditemukan algoritma yang paling akurat dalam prediksi kelulusan tepat waktu. Untuk itu, dalam penelitian ini akan mengkaji dan membuat model dengan algoritma *Naïve Bayes* dan *Neural Network* dalam menghasilkan rule prediksi kelulusan tepat waktu.

2.2 *Data Understanding Phase* (Fase Pemahaman Data)

Data yang diperoleh dari database SIMAK UIN Raden Fatah Palembang Fakultas Sains dan Teknologi pada tahun 2010 sampai dengan 2015 sebanyak 1425 mahasiswa. Attribute atau variable yang digunakan sebanyak 7 atribut: NPM, Nama Mahasiswa, Jenjang Pendidikan, Status, Jenis Kelamin, IPK (Indek Prestasi Kumulatif) dan IP semester. Dilakukan pemrosesan terhadap data tersebut sehingga digunakan sebanyak 7 atribut atau variable yang digunakan dalam prediksi kelulusan tepat waktu adalah: Nim, Nama Mahasiswa, Jenjang Pendidikan, IPK dan IP semester. Dari 7 atribut 2 adalah *predictor* yaitu IPK dan IP semester dan 1 attribut tujuan yaitu kelulusan tepat waktu.

Table 1. Attribut dan Nilai Kategori IPSemster, IPK dan Prediksi

No.	Atribut	Nilai
1.	IP Semester	Sangat Baik Baik Cukup Kurang
2.	IPK	Sangat Baik Baik Cukup Kurang
3.	Prediksi	Lulus tepat waktu
		Lulus terlambat

2.3 *Data Preparation Phase* (Fase Pengolahan Data)

Dari 1425 data mahasiswa diambil data dari angkatan 2010 sampai 2013 dengan pertimbangan sebagai data fakta kelulusan mahasiswa. Mahasiswa angkatan 2014 dan 2015 merupakan mahasiswa yang akan diprediksi kelulusannya.

Untuk selanjutnya dilakukan teknik *preprocessing* agar kualitas data yang diperoleh lebih baik dengan cara, (Vecellis, 2009):

1. *Data validation*, untuk mengidentifikasi dan menghapus data yang ganjil (*outlier/noise*), data yang tidak konsisten, dan data yang tidak lengkap (*missing value*).

2. *Data integration and transformation*, untuk meningkatkan akurasi dan efisiensi algoritma. Data yang digunakan dalam penulisan ini bernilai kategorikal. Untuk model *neural network*, data ditransformasi ke dalam angka menggunakan *software Rapidminer*.

Data size reduction and dicretization, untuk memperole data set dengan jumlah atribut dan *record* yang lebih sedikit tetapi bersifat informatif. Di dalam data *training* yang digunakan dalam penelitian ini, dilakukan seleksi atribut dan penghapusan data duplikasi menggunakan *software* Rapidminer.

2.4 *Modeling Phase* (Fase Pemodelan)

Pada tahapan ini merupakan tahapan pemrosesan data training yang diklasifikasikan oleh model dan kemudian menghasilkan sejumlah aturan. Pada penelitian ini menggunakan *algoritma Naïve Bayes* dan *Neural Network*.

2.5 *Evaluation Phase* (Fase Evaluasi)

Pada fase ini dilakukan pengujian terhadap model-model yang bertujuan untuk mendapatkan model yang paling akurat. Evaluasi dan validasi dilakukan dengan menggunakan metode *Confusion Matrix* untuk *algoritma Naïve Bayes* dan *Neural Network*.

2.6 *Deployment Phase* (Fase Penyebaran)

Setelah pembentukan model selanjutnya dilakukan analisa dan pengukuran pada tahap sebelumnya, pada tahap ini diterapkan model atau rule yang paling akurat dalam prediksi kelulusan tepat waktu dan selanjutnya dapat digunakan untuk mengevaluasi data baru.

2.6 Teorema Bayes

Bayes merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema Bayes (atau aturan Bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat (naïf). Dengan kata lain, Naïve Bayes, model yang digunakan adalah "model fitur independen".

Dalam Bayes (terutama Naïve Bayes), maksud independensi yang kuat pada fitur adalah bahwa sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama.

Prediksi Bayes didasarkan pada teorema Bayes dengan formula umum sebagai berikut :Penjelasan dari formula tersebut adalah sebagai berikut :

2.7 Parameter Keterangan

P(H|E) Probabilitas akhir bersyarat *(conditional probability)* suatu hipotesis H

terjadi jika diberikan bukti (evidence) E terjadi.

P(E|H) Probabilitas sebuah bukti E terjadi akan memengaruhi hipotesis H.

P(H) Probabilitas awal (priori) hipotesis H terjadi tanpa memandang bukti apapun.

P(E) Probabilitas awal (priori) bukti E terjadi tanpa memandang hipotesis/bukti yang lain.

Ide dasar dari aturan Bayes adalah bahwa hasil dari hipotesis atau peristiwa (H) dapat diperkirakan berdasarkan pada beberapa bukti (E) yang diamati. Ada beberapa hal penting dari aturan Bayes tersebut, yaitu:

1. Sebuah probabilitas awal/prior H atau P(H) adalah probabilitas dari suatu hipotesis sebelum bukti diamati.
2. Sebuah probabilitas akhir H atau P(H|E) adalah probabilitas dari suatu hipotesis setelah bukti diamati.

2.8 Naïve Bayes Untuk Klasifikasi

Kaitan antara *Naïve Bayes* dengan klasifikasi, korelasi hipotesis dan bukti klasifikasi adalah bahwa hipotesis dalam teorema *Bayes* merupakan label kelas yang menjadi target pemetaan dalam klasifikasi, sedangkan bukti merupakan fitur-fitur yang menjadikan masukkan dalam model klasifikasi.

Jika X adalah vektor masukkan yang berisi fitur dan Y adalah label kelas, *Naïve Bayes* dituliskan dengan P(X|Y). Notasi tersebut berarti probabilitas label kelas Y didapatkan setelah fitur-fitur X diamati. Notasi ini disebut juga probabilitas akhir (*posterior probability*) untuk Y, sedangkan P(Y) disebut probabilitas awal (*prior probability*) Y.

Selama proses pelatihan harus dilakukan pembelajaran probabilitas akhir P(Y|X) pada model untuk setiap kombinasi X dan Y bedasarkan informasi yang didapat dari data latih. Dengan membangun model tersebut, suatu data uji X' dapat diklasifikasikan dengan mencari nilai Y' dengan memaksimalkan nilai P(X'|Y') yang didapat.

Formulasi *Naïve Bayes* untuk klasifikasi adalah:

Setiap set fitur X = {X1, X2, X3,…, Xq} terdiri atas *q* atribut (*q* dimensi).

Umumnya, *Bayes* mudah dihitung untuk fitur bertipe kategoris seperti pada kasus klasifikasi hewan dengan fitur "penutup kulit dengan nilai {bulu, rambut, cangkang} atau kasus fitur "jenis kelamin"

dengan nilai {pria, wanita}. Namun untuk fitur dengan tipe numerik (kontinu) ada perlakuan khusus sebelum dimasukkan dalam *Naïve Bayes*. Caranya adalah:

1. Melakukan diskretisasi pada setiap fitur kontinu dan mengganti nilai fitur kontinu tersebut dengan nilai interval diskret. Pendekatan ini dilakukan dengan mentransformasikan 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(Xi|Y), sedangkan distribusi Gaussian dikarakteristikkan dengan dua parameter : *mean*, μ dan *varian*, . Untuk setiap kelas *yj*, probabilitas bersyarat kelas *yj* untuk fitur Xi adalah :

Parameter bisa didapat dari mean sampel Xi dari semua data latih yang menjadi milik kelas yj, sedangkan dapat diperkirakan dari varian sampel (s2) dari data latih.

2.9 Karakteristik *Naïve Bayes*

Klasifikasi dengan *Naïve Bayes* bekerja berdasarkan teori probabilitas yang memandang semua fitur dari data sebagai bukti dalam probabilitas. Hal ini memberikan karakteristik *Naïve Bayes* sebagai berikut:

1. Metode *Naïve Bayes* bekerja teguh (*robust*) terhadap data-data yang terisolasi yang biasanya merupakan data dengan karakteristik berbeda (*outliner*). *Naïve Bayes* juga bisa menangani nilai atribut yang salah dengan mengabaikan data latih selama proses pembangunan model dan prediksi.
2. Tangguh menghadapi atribut yang tidak relevan.
3. Atribut yang mempunyai korelasi bisa mendegradasi kinerja klasifikasi *Naïve Bayes* karena asumsi independensi atribut tersebut sudah tidak ada.

Naive Bayes merupakan algoritma klasifikasi yang sederhana dimana setiap atribut bersifat *independent* dan memungkinkan berkontribusi terhadap keputusan akhir.

Dasar dari teorema Naïve Bayes yang dipakai dalam pemrograman adalah rumus bayes yaitu sebagai berikut:

(|)=(|)()()

dimana *P(H|X)* merupakan probabilitas H di dalam X atau dengan bahasa lain *P(H|X)* adalah persentase banyaknya *H* di dalam *X*, *P(X|H)* merupakan probabilitas *X* di dalam

H, P(H) merupakan probabilitas prior dari *H* dan *P(X)* merupakan probabilitas prior dari *X*.

2.10 Clustering

Berbeda dengan *association rule mining* dan *classification* dimana kelas data telah ditentukan sebelumnya, *clustering* melakukan pengelompokan data tanpa berdasarkan kelas data tertentu. Bahkan *clustering* dapat dipakai untuk memberikan label pada kelas data yang belum diketahui itu. Karena itu *clustering* sering digolongkan sebagai metode *unsupervised learning*. Prinsip dari *clustering* adalah memaksimalkan kesamaan antar anggota satu kelas dan meminimumkan kesamaan antar kelas/*cluster*. *Clustering* dapat dilakukan pada data yang memiliki beberapa atribut yang dipetakan sebagai ruang multidimensi. Banyak algoritma *clustering* memerlukan fungsi jarak untuk mengukur kemiripan antar data, diperlukan juga metode untuk normalisasi bermacam atribut yang dimiliki data. Beberapa kategori algoritma *clustering* yang banyak dikenal adalah metode partisi dimana pemakai harus menentukan jumlah *k* partisi yang diinginkan lalu setiap data diuji untuk dimasukkan pada salah satu partisi, metode lain yang telah lama dikenal adalah metode hierarki yang terbagi dua lagi: *bottom-up* yang menggabungkan *cluster* kecil menjadi *cluster* lebih besar *dan top-down* yang memecah *cluster* besar menjadi *cluster* yang lebih kecil. Kelemahan metode ini adalah bila salah satu penggabungan/pemecahan dilakukan pada tempat yang salah, tidak dapat didapatkan *cluster* yang optimal. Pendekatan yang banyak diambil adalah hierarki dengan metode *clustering* lainnya seperti yang dilakukan oleh Chameleon[1].

Clustering dengan pendekatan hirarki mengelompokkan data yang mirip dalam hirarki yang sama dan yang tidak mirip di hirarki yang agak jauh. Ada dua metode yang sering diterapkan yaitu *agglomerative hiearchical clustering* dan *divisive hierarchical clustering*. Agglomerative melakukan proses clustering dari *N* cluster menjadi satu kesatuan cluster, dimana *N* adalah jumlah data, sedangkan divisive melakukan proses clustering yang sebaliknya yaitu dari satu cluster menjadi *N* cluster.

Beberapa metode *hierarchical clustering* yang sering digunakan dibedakan menurut cara mereka untuk menghitung tingkat kemiripan. Ada yang menggunakan *Single Linkage*, *Complete Linkage*, *Average Linkage*, *Average Group Linkage* dan lain-lainnya. Seperti juga halnya dengan *partition-based clustering*, kita juga bisa memilih jenis jarak yang

digunakan untuk menghitung tingkat kemiripan antar data.

Salah satu cara untuk mempermudah pengembangan dendogram untuk hierarchical clustering ini adalah dengan membuat similarity matrix yang memuat tingkat kemiripan antar data yang dikelompokkan. Tingkat kemiripan bisa dihitung dengan berbagai macam cara seperti dengan Euclidean Distance Space. Berangkat dari similarity matrix ini, kita bisa memilih lingkage jenis mana yang akan digunakan untuk mengelompokkan data yang dianalisa.

Menurut Jiawei Han (2006: 401-430) secara umum metode pada clustering dapat digolongkan ke dalam beberapa metode berikut:

1. Metode partisi (Partitioning Method)

Langkah kerja metode partisi, yaitu apabila terdapat basis data sejumlah n objek atau data tupelo, selanjutnya data di partisi menjadi k partisi dari data, dimana setiap partisi mewakili sebuah cluster dan $k \leq n$. Adapun syarat yang harus terpenuhi sebagai berikut: (1) setiap kelompok harus berisi setidaknya satu objek, dan (2) setiap objek harus memiliki tepat satu kelompok. Awalnya basis data dipartisi menjadi k partisi. Kemudian menggunakan teknik relokasi berulang, mencoba untuk memperbaiki partisi dengan memindahkan dari satu kelompok ke kelompok lain. Kriteria umum dari partisi yang baik adalah bahwa objek dalam satu cluster memiliki kemiripan yang sangat dekat, sedangkan objek dalam cluster yang berbeda memiliki kemiripan yang jauh berbeda.

Pencapaian optimalitas global dalam pengelompokan berbasis partisi akan memerlukan penghitungan lengkap dari semua partisi yang memungkinkan. Sebaliknya, sebagian besar aplikasi mengadopsi salah satu dari beberapa metode heuristik yang popular, seperti (1) algoritma k-means, dimana setiap segmen diwakili oleh nilai rata-rata dari objek dalam cluster, dan (2) algoritma k-medoids, dimana setiap segmen diwakili oleh salah satu objek yang terletak didekat centroid. Metode pengelompokan heuristik ini bekerja dengan baik untuk menemukan cluster berbentuk bola kecil untuk basis data yang berukuran sedang.

2. Metode Hirarki (Hierarchi Method)

Metode hirarki menciptakan dekomposisi hirarki dari himpunan objek data yang diberikan. Sebuah metode hirarki dapat diklasifikasikan sebagai salah satu agglomerative atau memecah belah, berdasarkan cara dekomposisi hirarki terbentuk. Pendekatan agglomerative memiliki dua cara pendekatan, yaitu bottom–up dan top-down. Pendekatan bottom-up berlangsung seperti berikut, awalnya setiap

objek membentuk kelompok tersendiri. Berturut-turut menggabungkan objek atau kelompok yang dekat satu sama lain, sampai semua kelompok digabung menjadi satu (tingkat teratas dari hirarki), atau sampai terjadinya kondisi pemutusan hubungan. Sedangkan pendekatan topdown, dimulai dengan semua objek dalam cluster yang sama dibagi menjadi kelompok yang lebih kecil, sampai akhirnya setiap objek dalam satu cluster atau sampai terjadi kondisi pemutusan hubungan.

Metode hirarki memuat fakta bahwa setelah langkah penggabungan atau split dilakukan, proses memecah belah tidak dapat dibatalkan. Ada dua pendekatan untuk meningkatkan kualitas pengelompokan hirarki: (1) melakukan analisis yang cermat terhadap objek "linkage" pada setiap partisi hirarki, seperti di Chameleon, atau (2) mengintegrasikan aglomerasi hirarki dan pendekatan-pendekatan lain dengan terlebih dahulu menggunakan algoritma agglomerative hirarki objek kelompok ke dalam microclusters, dan kemudian melakukan macroclustering pada microclusters menggunakan metode pengelompokan lain seperti relokasi berulang. Salah satu algoritma yang tergolong kedalam metode hirarki yaitu BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies). BIRCH merupakan salah satu algoritma pengelompokan hirarki yang terintegrasi. BIRCH memperkenalkan dua konsep, clustering feature dan clustering feature tree (CF tree), yang mana digunakan untuk menggambarkan ringkasan cluster.

3. Metode Berbasis Kerapatan (Density Based Method)

Sebagian besar metode cluster mempartisi objek berdasarkan jarak antara objek. Metode semacam itu hanya dapat menemukan cluster berbentuk bola dan mengalami kesulitan dalam menemukan cluster berbentuk sembarang. Metode pengelompokan lain telah dikembangkan berdasarkan gagasan kerapatan. Ide secara umumnya adalah terus tumbuhnya cluster yang diberikan selama densitas (jumlah objek atau pusat massa) di "neighborhood (lingkungan)" melebihi ambang batas tertentu, yaitu untuk setiap titik data dalam cluster tertentu, lingkungan radius tertentu setidaknya harus memuat minimal jumlah titik. Metode tersebut dapat digunakan untuk menyaring outlier dan menemukan bentuk kelompok sembarang.

Beberapa algoritma yang termasuk kedalam metode berbasis kerapatan, yaitu: DBSCAN, OPTICS, DENCLUE. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) merupakan algoritma yang memperluas wilayah dengan kepadatan yang tinggi ke dalam cluster dan menempatkan cluster irregular pada basis

data spasial dengan noise. Algoritma ini mendefiniskan cluster sebagai kumpulan maksimal dari titik-titik kepadatan yang terkoneksi. OPTICS (Ordering Points To Identify the Clustering Structure) merupakan algoritma pada metode hirarki yang diusulkan untuk mengatasi kesulitan user dalam menentukan parameter yang digunakan untuk menemukan cluster yang bisa diterima. DENCLUE (Density Based Clustering) merupakan algoritma clustering yang berdasarkan suatu set fungsi distribusi kerapatan.

4. Metode berbasis Grid

Metode berbasis grid mengkuantisasi ruang objek kedalam jumlah sel terbatas yang membentuk struktur jaringan. Semua operasi pengelompokan dilakukan pada struktur jaringan (yaitu, pada ruang terkuantisasi). Keuntungan utama dari pendekatan ini adalah waktu proses yang cepat, yang biasanya tergantung pada jumlah data objek dan bergantung hanya pada jumlah sel dalam setiap dimensi dalam ruang terkuantisasi. Algoritma yang termasuk kedalam metode berbasis grid diantaranya: STING, Wave Cluster, dan lainnya. STING (Statistical Information Grid) merupakan algoritma clustering yang bekerja dengan membagi daerah spatial menjadi sel-sel rectanguleri. Wave Cluster merupakan algoritma clustering yang melakukan sumarisasi data yang dilakukan dengan menentukan sruktur grid multidimensional terhadap space data.

5. Metode Berbasis Model

Metode berbasis model membuat hipotesis sebuah model untuk masingmasing kelompok dan menemukan yang terbaik sesuai data yang diberikan model. Algoritma berbasis model dapat menemukan cluster dengan membangun fungsi kerapatan yang menggambarkan distribusi spasial titik data. Hal ini menyebabkan cara otomatis untuk menentukan jumlah cluster berdasarkan standar statistik, membawa "noise" atau outlier kedalam perhitungan dan dengan demikian menghasilkan metode pengelompokan yang kuat. Algoritma yang termasuk ke dalam metode berbasis model yaitu COBWEB dan SOM. COBWEB adalah algoritma pembelajaran konseptual yang melakukan analisis probabilitas dan mengambil konsep sebagai model untuk cluster. SOM (mengorganisir diri berfitur peta) adalah algoritma berbasis jaringan saraf yang clusternya dengan memetakan data dimensi tinggi ke peta fitur 2D atau 3D, yang juga berguna untuk visualisasi data.

Secara sederhana, clustering dapat dikonsentrasikan pada jarak Euclidean antar record :

dimana x = $x_1,x_2,\ldots,x_m$ dan y = $y_1,y_2,\ldots,y_m$ yang melambangkan nilai atribut m dari dua catatan. Fungsi perhitungan matrik lainnya juga ada, seperti jarak cityblock :

atau jarak Minkowski, yang merupakan kasus umum dari dua metrik sebelumnya untuk eksponen q secara umumnya:

untuk kategori atribut, dapat didefinisikan "berbeda dari" fungsi untuk membandingkan nilai atribut ke i dari sepasang catatan :

dimana x_i dan y_i adalah nilai kategorik. Kemudian dapat mengganti (x_i,y_i) untuk *i*, istilah ini dalam metrik jarak Euclidean diatas.

Performa yang optimal dari algoritma clustering sama seperti algoritma klasifikasi. Algoritma ini membutuhkan data yang akan dinormalisasi sehingga tidak ada variabel tertentu atau bagian dari variabel yang mendominasi analisis. Analisis dapat menggunakan salah satu dari min-max normalisasi atau standar ZScore.

Secara umum teknik clustering memiliki tujuan pada identifikasi kelompok data dimana kesamaan dalam suatu kelompok data sangat tinggi sedangkan kesamaan dengan kelompok data lain sangat rendah (Larose, 2005:148-149). Algoritma k-means adalah algoritma klasik untuk menyelesaikan masalah clustering, yang relatif sederhana dan cepat (Zhang & Fang, 2013: 193). Algoritma k-means clustering lebih sering dikenal karena kemampuannya dalam mengelompokkan data dalam jumlah besar dengan cepat dan efisien. Algoritma kmean sangat rawan dipusat-pusat cluster awal, karena pusat cluster awal diproduksi secara acak. Algoritma k-means tidak menjanjikan hasil pengelompokan yang khas. Efisiensi keaslian algoritma k-mean sangat bergantung pada titik pusat cluster (centroid) awal (Yedla, et al, 2010: 121-122). Langkah kerja dari algoritma k-means adalah sebagai berikut :

1. Menanyakan kepada pengguna berapa banyak k cluster dataset yang akan dipartisi.
2. Menetapkan secara acak k record yang menjadi lokasi pusat cluster awal.
3. Setiap record dicari centroid cluster terdekatnya. Artinya setiap centroid cluster "memiliki" subset dari record, sehingga merepresentasikan sebuah partisi dari dataset. Didapatkan k cluster, C1, C2, …, Ck.
4. Setiap k cluster dicari centroidnya dan memperbarui lokasi setiap pusat cluster untuk nilai centroid baru.

5. Ulangi langkah 3 sampai 5, sampai terjadi konvergensi atau terjadi penghentian.

Algoritma berakhir ketika titik pusat cluster tidak lagi berubah. Dengan kata lain, algoritma berakhir ketika dari seluruh cluster C1, C2, …, Ck, semua record yang dimiliki oleh masing-masing pusat cluster tetap dalam cluster itu. Atau, algoritma dapat berhenti ketika beberapa kriteria konvergensi terpenuhi, seperti ada penyusutan yang tidak signifikan dalam jumlah kuadrat error (sum of squared errors):

dimana $p \tilde{o} \; Ci$ melambangkan setiap titik dalam cluster i dan m_i merupakan pusat cluster i (Larose, 2005:153).

3. HASIL

Dalam pengumpulan data terdapat sumber data, sumber data yang dihimpun langsung oleh peneliti disebut dengan sumber primer, sedangkan apabila melalui tangan kedua disebut sumber sekunder (Riduwan, 2008). Data yang diperolah adalah data sekunder karena diperoleh dari database mahasiswa yang dimiliki oleh Universitas Islam Negeri Raden Fatah Palembang. Data yang diperoleh dalam penelitian ini adalah data kualitatif dan kuantitatif. Data yang dikumpulkan adalah data mahasiswa UIN Raden Fataj Palembang pada Fakultas Sains dan Teknologi (FST) untuk tahun angkatan 2010 sampai dengan 2015. Data terkumpul sebanyak 1.425 data, dengan atribut nim, nama, IP semester 1, IP semester 2, IP semester 3 sampai dengan IP Semester 8 dan IPK (Indek Prestasi Kumulatif). Mahasiswa kelulusan tahun 2014 sampai 2016 dijadikan sebagai data fakta kelulusan dan angkatan 2014 dan 2015 adalah mahasiswa yang akan diprediksi kelulusannya.

Tabel. 2 Data mahasiswa (10 dari 1425 mahasiswa)

NIM	Nama	Kelamin	Tanggal Lahir
1522810002	Aldini Caesar Seftiani	W	16/09/1997
1522810003	Ali Ma`Ruf Saputra	P	16/02/1996
1522810004	Indah Farodilah	W	17/12/1997
1522810005	Kurnia Fajria Oksyarina	W	21/10/1996
1522810006	Kusumawardhani Fildzah Hani	W	24/03/1996
1522810007	Nurdiah Hasana	W	13/08/1997
1522810008	Nurhalimah	W	29/09/1997
1522810009	Reni Afriani	W	22/04/1997
1522810010	Rima Vivian Sari	W	30/05/1998
1522810011	Wiza Shabrina	W	25/06/1997

Dilihat dari data yang diperoleh dari database Sistem Informasi Akademik terlihat bahwa data tersebut masih ada yang tidak valid seperti tanggal lahir yang salah. Tanggal lahir tidak diperhitungkan dalam kelulusan maka data tersebut tidak perlu

dilakukan *cleansing* terlebih dahulu.

Tabel 3 Data Nilai Mahasiswa (10 nilai dari 66278 nilai KRS)

MhswID	Nama	SKS	NilaiAkhir
1522810001	FILSAFAT UMUM	2	75.00
1522810001	BAHASA INGGRIS	2	64.10
1522810001	KIMIA DASAR	2	80.00
1522810001	BAHASA INDONESIA	2	48.7
1522810001	BAHASA ARAB	2	80.00
1522810001	PANCASILA	2	81.00
1522810001	STUDI KEISLAMAN	4	10
1522810001	IAD/ISD/IBD	2	83.00
1522810001	BIOLOGI UMUM	2	65.10
1522810001	PRAKTIKUM KIMIA DASAR	1	81.00

Total keseluruhan data nilai Fakultas Saintek UIN Raden Fatah untuk angkatan 2010 sampai dengan 2015 sebanyak 66278 untuk 1425 mahasiswa. Data ini masih mentah dan belum dijadikan IP semester masing-masing mahasiswa. Perlu dilakukan *preprocessing* data untuk mengubah bentuk KRS tersebut ke bentuk IP semester masing-masing mahasiswa dan IPKnya.

Proses dilakukan dengan bantuan aplikasi Microsoft Excel mengingat kemampuan Microsoft Excel untuk mengolah data berupa tabelyang sangat besar. Hasil dari *preprocessing* data tersebut adalah nilai mahasiswa sudah dalam bentuk IP semester masing masing mahasiswa.

Metode Naïve Bayes menggunakan data training sejumlah 876 record. Perhitungan pemilihan prediksi kelulusan tepat waktu dengan nilai prediksi lulus tepat waktu dan lulus tidak tepat waktu terlihat pada Tabel 4. Baris-baris berikutnya adalah hasil perhitungan nilai probabilitas prior, yaitu probabilitas nilai lulus tepat waktu dan lulus tidak tepat waktu masing-masing atribut terhadap total lulus tepat waktu dan lulus tidak tepat waktu dari seluruh data. Dalam data training terdapat 876 record dengan 795 kasus lulus tepat waktu dan 81 kasus lulus terlambat, untuk menentukan prior probability dengan menggunakan rumus [1]:

Bayes :

Naïve Bayes :

P(Tepat Waktu) = 560/876 = 0,639269

P(Tidak Tepat Waktu) = 316/876 = 0,36073

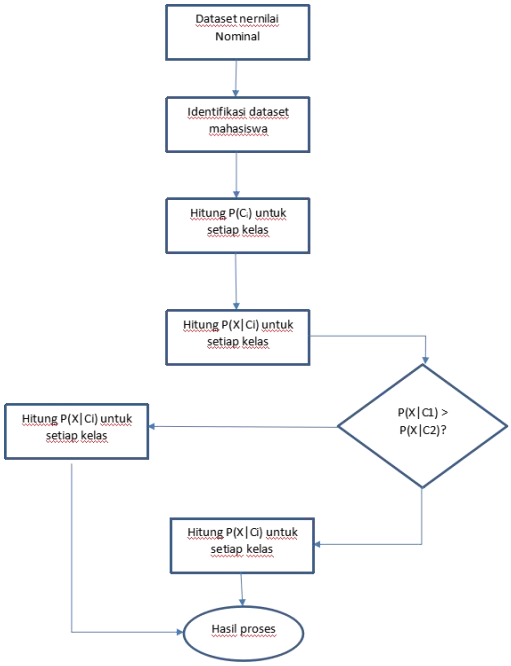
Tabel 4 Perhitungan *prior probability.*

P	Jumlah	TW	TTW	P(X|Ci)	
				TW	TTW

	mahasiswa				
IP Semester 1	876	560	316		
>=3.75	41	22	19	0.073214	0.129747
3.74 – 2.75	527	357	170	0.941071	1.66
2.74 – 2.00	152	96	56	0.271429	0.481
<2.00	156	85	71	0.278571	0.494
IP Semester 2	872	560	312		
>=3.75	33	30	3	0.053571	0.094937
3.74 – 2.75	634	468	166	0.835714	1.481013
2.74 – 2.00	123	50	73	0.089286	0.158228
<2.00	82	12	70	0.021429	0.037975
IP Semester 3	852	550	302		
>=3.75	33	30	3	0.053571	0.094937
3.74 – 2.75	614	458	116	0.835714	1.481013
2.74 – 2.00	123	50	73	0.089286	0.158228
<2.00	82	12	70	0.021429	0.037975
IP Semester 4	802	530	282		
>=3.75	33	30	3	0.053571	0.094937
3.74 – 2.75	624	468	156	0.835714	1.481013
2.74 – 2.00	113	50	63	0.089286	0.158228
<2.00	72	12	60	0.021429	0.037975
IP Semester 5	672	460	212		
>=3.75	33	30	3	0.053571	0.094937
3.74 – 2.75	534	468	66	0.835714	1.481013
2.74 – 2.00	23	50	73	0.089286	0.158228
<2.00	82	12	70	0.021429	0.037975
IP Semester 6	872	560	312		
>=3.75	33	30	3	0.053571	0.094937
3.74 – 2.75	634	468	166	0.835714	1.481013
2.74 – 2.00	123	50	73	0.089286	0.158228
<2.00	82	12	70	0.021429	0.037975
IP Semester 7	772	500	302		
>=3.75	33	30	3	0.053571	0.094937
3.74 – 2.75	634	468	166	0.835714	1.481013
2.74 – 2.00	123	50	73	0.089286	0.158228
<2.00	82	12	70	0.021429	0.037975
IP Semester 8	872	560	312		
>=3.75	33	30	3	0.053571	0.094937
3.74 – 2.75	634	468	166	0.835714	1.481013
2.74 – 2.00	123	50	73	0.089286	0.158228
<2.00	82	12	70	0.021429	0.037975

Pemilihan teknik pemodelan

Tool yang digunakan adalah *RapidMiner* versi 5.3:



Gambar 1. Proses *data mining*

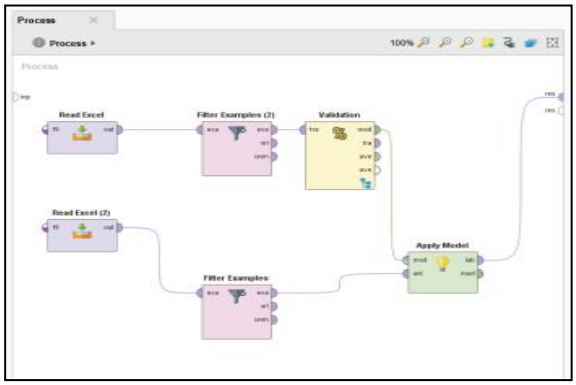
Dari data atribut nominal, kemudian identifikasi dataset mahasiswa. Hitung (Ci) untuk setiap atribut, dalam kasus dataset pada penelitian ini yaitu atribut tahun kelulusan yang terdiri dari 2 kelas yaitu kelas lulus tepat waktu dinyatakan "YES" dan tidak lulus tepat waktu dinyatakan "NO".

Kemudian hitung P(X|Ci), i=1,2 untuk setiap kelas atau atribut. Setelah itu bandingkan, jika P(X|C1) > (P{X|C2) maka kesimpulannya C1 adalah kelas lulus tepat waktu = "YES". Jika P(X|C1) < (P{X|C2) maka kesimpulannya C2 tidak lulus tepat waktu = "NO".

3.1 Implementasi dengan *RapidMiner*

Berikut adalah pengolahan data dengan menggunakan *naïve bayes* pada *RapidMiner :*

1. Pemodelan adalah tahapan (langkah) dalam membuat model dari suatu sistem nyata (realitas).
2. Rapidminer adalah sebuah lingkungan machine learning data mining, text mining dan predictive analytics.
3. Jumlah dataset mencapai 928 record maka penulis menggunakan tools Rapidminer untuk membantu proses perhitungan.
4. X Validation untuk membantu mengahasilkan tingkat keakurasian berdasarkan dataset yang telah di lakukan proses klasifikasi.



Gambar 2. Proses pada Rapidminer

Salah satu tujuan penelitian ini adalah untuk mengetahui nilai akurasi dari algoritma naïve bayes yang digunakan untuk mengklasifikasi kelulusan.

Di dalam kolom training terdapat algoritma klasifikasi yang diterapkan yaitu Naïve bayes. Sedangkan di dalam kolom testing terdapat Apply Model untuk menjalankan model naïve bayes dan Performance untuk

mengukur performa dari model Naïve bayes tersebut.

Pada percobaan dengan algoritma naïve bayes dengan menggunakan tools Rapidminer diperoleh waktu komputasi adalah 0 second. 0 second disini artinya komputasi menggunakan naïve bayes berjalan cukup cepat. Hal ini sesuai dengan kelebihan naïve bayes dibandingkan beberapa algoritma lain seperti neural network yang membutuhkan waktu cukup lama untuk melakukan komputasi data.

Gambar 3. Simpel distribution.

Hasil akurasi model naïve bayes menunjukkan tingkat akurasinya 82.08% artinya model klasifikasi kelulusan menggunakan naïve bayes terbukti baik hal ini dilihat dari tingkat akurasinya yang mencapai 82.08% akan tetapi hal ini perlu di tinjau ulang dari sudut pandang kompleksitas dan jumlah datasetnya.

Percobaan pada penelitian ini menggunakan Rapidminer 5.3.008. Algoritma yang digunakan adalah *naive bayes. Validasinya* menggunakan *x-validation* dan untuk testing menggunakan *Apply Model* untuk menjalankan algoritma atau model *naïve bayes* serta *Performance* untuk mengukur performa dari model *naïve bayes* tersebut.

Evaluasi *(Evaluation)*

Evaluasi adalah fase lanjutan terhadap tujuan *data mining*. Evaluasi dilakukan secara mendalam dengan tujuan agar hasil pada tahap pemodelan sesuai dengan sasaran yang ingin dicapai dalam tahap *business understanding*.

3.2 Evaluasi Hasil *(Evaluation Results)*

Tahap ini menilai sejauh mana hasil pemodelan *data mining* memenuhi tujuan *data mining* yang telah ditentukan pada tahap *business understanding*.

3.3 Pengecekan Ulang Proses *(Review Process)*

Pada tahapan ini penulis memastikan bahwa semua tahapan / faktor penting yang telah dilakukan dalam pengolahan data tidak ada yang terlewatkan.

3.4 Menentukan Langkah Selanjutnya *(Determine Next Steps)*

Pada tahap ini adalah tahapan dalam menentukan langkah selanjutnya yang dilakukan. Terdapat 2 pilihan yaitu kembali pada tahap awal (*business understanding*) atau melanjutkan ke tahap akhir (*deployment*).

4. Hasil Analisis

Deployment merupakan tahapan akhir dalam pembuatan laporan hasil kegiatan data mining. Laporan akhir yang berisi mengenai pengatahuan yang diperoleh atau pengenalan pola pada data dalam proses *data mining*.

Berdasarkan penelitian yang dilakukan, telah dihasilkan suatu pola, informasi, dan pengetahuan baru dalam proses data mining untuk klasifikasi kelulusan mahasiswa berdasarkan data mahasiswa Fakultas Sains dan Teknologi UIN Raden Fatah. Dari penelitian tersebut dihasilkan suatu pola, informasi, dan pengetahuan baru sesuai dengan tujuan *data mining* yaitu pola perhitungan *data mining* yang berisi data *training* dan data *testing* serta mencari probabilitas dari setiap atribut berdasarkan data *training* dan data *testing* untuk menghasilkan suatu informasi baru, apakah pada data mahasiswa Fakultas Sains dan Teknologi UIN Raden Fatah lebih banyak kelas tahun lulus yang tepat waktu atau kelas tahun lulus tidak tepat waktu. Kemudian untuk menguji tingkat keakurasiannya maka digunakan Rapidminer sebagai alat bantu dalam proses pengujian tingkat akurasi dari klasifikasi tersebut.

Dari proses perhitungan *data mining* menggunakan algoritma *naïve bayes* dan tingkat keakurasian, dihasilkan suatu informasi baru yaitu perhitungan *data mining* berdasarkan mahasiswa Fakultas Sains dan Teknologi UIN Raden Fatah, menunjukkan kelas tahun lulus “yes” / tepat waktu dengan total perkalian *prior probability* senilai 0, sedangkan kelas tahun lulus “no” / tidak tepat waktu dengan total perkalian *prior probability* senilai 0.00055. Untuk tingkat akurasi berdasarkan proses klasifikasi menggunakan algoritma *naïve bayes*, dengan melalui semua tahapan dipastikan tidak ada bagian – bagian penting yang terlewatkan, dihasilkan tingkat akurasi sebesar 82.08 %.

Berdasarkan hasil perhitungan data mining dan proses pengujian tingkat akurasi dengan menggunakan Rapidminer, dapat ditarik kesimpulan bahwa angkatan 2010 kelas tahun lulus “no” / tidak tepat waktu lebih besar dari kelas tahun lulus “yes” / tepat waktu. Sedangkan analisa yang dilakukan terhadap tingkat akurasi menggunakan algoritma naïve bayes menunjukkan bahwa nilai yang dihasilkan oleh algoritma *naïve bayes* memiliki tingkat kekuatan yang cukup tinggi. Hal ini di buktikan dengan hasil perhitungan yang mencapai nilai 82.08 %, Nilai 82. 08 % membuktikan bahwa model yang dibangun dapat digunakan untuk melakukan klasifikasi kelulusan mahsiswa. Nilai 82.08 % bisa juga di sebabkan oleh kekurang kompleksan data yang mengakibatkan model dapat memprediksi

dengan akurat.

5. SIMPULAN

A. Berdasarkan perhitungan data mining menggunakan algoritma naïve bayes, dapat ditarik kesimpulan bahwa kelas tahun lulus "no" / tidak lulus tepat waktu lebih kecil daripada kelas tahun lulus "yes" / lulus tepat waktu.

B. Dari hasil observasi terhadap dataset mahasiswa Fakultas Sians dan Teknlogi UIN Raden Fatah Palembang dan melalui proses perhitungan menggunakan metode klasifikasi naïve Bayes dengan atribut yang telah dijelaskan di pembahasan sebelumnya, didapatkan sebuah hasil bahwa nilai akurasi terhadap klasifikasi kelulusan sebesar 82.08 %. Dimana 82.08 % bisa juga disebabkan oleh kurang kompleksitas data yang mengakibatkan model dapat memprediksi cukup akurat.

6. REFERENSI

Adeyemo B. A dan Kuye G, 2006, Mining Students' Acedemic Performance Using Decision tree Algorithms, *Journal of Information Technology Impact*, Vol. 6 No. 3 pp. 161-170

Al-Radaideh Q.A Al-Shawakfa E.M. dan Al-Najjar I.M, 2006, Mining Stuents Data using Decision Trees, *International Arab Conference on Information Technology (ACIT'2006),* pp. 1-5.

Dubes R.C. and Jain, A.K., 1998, *Algorithms for Clustering Data*, Prentice-Hall.

Karypis, George. Han, Eui-Hong(Sam). Kumar, Vipin. *CHAMELEON:A hierarchical clustering algorithm using dynamic modeling*. 2007.

Santosa Budi. 2007, *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*, Graha Ilmu, Yogyakarta

Sajadin S, Embong. A, Mohammad, M. A, Furqan.M, 2009 *" Improving Student Academik Performance Using Data Mining Techniques"*. Proceeding The 5th IMTGT International Conference on Mathematic, Statistics and Their Application (ICMSA 2009), ISBN 978-602-95343-0-6, page 390-394.

Turban, E., Aronson, J. E. dan Liang, T., 2005, *Decision Support Sistems and Intelliget Sistems* (Sistem Pendukung Keputusan

dan Sistem Cerdas), Edisi Ketujuh, Andi, Yogyakarta.

Vapnik V.N., 2007. *The Nature of Statistical Learning Theory*, 2nd edition, Springer-Verlag, New York Berlin Heilderberg.

Waiyamai,K, 2003. "*Improving Quality Graduate Student by Data Mining*". Departement of Computer engineering. Faculty of Engineering. Kasetsart University, Bangkok Thailand.