Prediksi Kelulusan Tepat Waktu Mahasiswa STMIK YMI Tegal Menggunakan Algoritma Naïve Bayes

Time Graduation Prediction by Using Naïve Bayes Algorithm at STMIK YMI Tegal

Aang Alim Murtopo
STMIK YMI TEGAL, Jl. Pendidikan No. 1 Tegal
Jurusan Teknik Informatika
Email: aang.alim@gmail.com

Abstrak

Kualitas perguruan tinggi, khususnya program studi di Indonesia diukur berdasarkan akreditasi yang dilaksanakan oleh Badan Akreditasi Nasional Perguruan Tinggi atau BAN PT. Kualitas tersebut diukur berdasarkan 7 standar utama, salah satu nya adalah Mahasiswa dan Lulusan. Perguruan tinggi memiliki data akademik dan biodata mahasiswa sejak mereka mendaftar hingga lulus kuliah. Algoritma klasifikasi data mining Naïve Bayes dapat digunakan untuk prediksi kelulusan mahasiswa yang nantinya bisa di kategorikan tepat waktu atau tidak tepat waktu, dari hasil prediksi bisa di manfaatkan untuk dasar pengambilan keputusan sehingga dapat meningkatkan kualitas dari keputusan manajerial institusi. Banyak variabel yang mempengaruhi mahasiswa bisa lulus secara tepat waktu, sehingga dalam penelitian ini menggunakan faktor internal (faktor dari dalam diri sendiri) dan faktor eksternal (faktor dari luar diri sendiri). Faktor eksternal yang digunakan untuk menjadi penentu dalam model ini antara lain status kerja dan status perkawinan.Berdasar faktor tersebut apakah faktor eksternal berpengaruh pada kelulusan mahasiswa secara tepat waktu. Hasil dari penelitian ini adalah pengukuran akurasi, dimana sebelum didapatkan nilai akurasi dilakukan pengujian dengan memanfaatkan ROC Curva dan k-fold cross validation,pengujian dilakukan sebanyak 10 fold. Dari hasil pengujian didapat nilai akurasi rata-rata sebesar 91,29%, sedangkan nilai akurasi tertinggi dari hasil pengujian 10-fold cross validation sebesar 94,34%.

Kata kunci *—Naïve Bayes, Lulus tepat waktu, Faktor internal, Faktor eksternal, ROC Curve, K-fold cross validation*

Abstract

The quality of higher education, particularly the study program in Indonesia is measured based accreditation conducted by the Badan Akreditasi Nasional Perguruan Tinggi or BANPT. Quality is measured by the 7 main standards,one of them is the Students and Graduates. Colleges have student academic data and personal data since they signed up to college. Algorithms Naïve Bayes classification of data mining can be used for prediction of graduation students will be categorized on time or not timely, the results of prediction can be utilized for the basic decision-making so as to improve the quality of managerial decisions institution. Many variables affect the student can graduate in a timely manner, so that in this study using the internal factors (factors of the self) and external factors (factors outside your self). External factors which used to be decisive in this model others work status and marital status. Based on these factors whether external factors affect student graduation after on time.The results of this study is the measurement accuracy, which prior to the testing accuracy values obtained by using ROC Curves and k-fold cross validation where testing is done as much as 10-fold. From the test results it can be an average accuracy rate of 91.29%, while the value of the highest accuracy of the results of testing10-fold cross validation of 94.34%.

Keywords—*Naïve Bayes, Graduate on time,Internal factors, External factors, ROC Curves, K-fold cross validation*

1. PENDAHULUAN

1.1 Latar Belakang

Perguruan tinggi adalah satuan pendidikan penyelenggara pendidikan tinggi[1]. Kualitas perguruan tinggi, khususnya program studi di Indonesia diukur berdasarkan akreditasi yang dilaksanakan oleh Badan Akreditasi Nasional Perguruan Tinggi atau BAN PT. Kualitas tersebut diukur berdasarkan 7 standar utama, salah satu nya adalah Mahasiswa dan Lulusan. Khusus mengenai evaluasi standar mahasiswa dan lulusan, komponen yang dinilai adalah: sistem rekrutmen mahasiswa baru, dan lulusan (rata- rata masa studi dan IPK)[2]. Salah satu permasalahan yang banyak di hadapi oleh perguruan tinggi untuk meningkatkan kualitas pendidikan mahasiswa dan untuk meningkatkan kualitas dari keputusan manajerial institusi. Pencapaian kualitas lavel mutu tertinggi dari system perguruan tinggi adalah dengan menggali pengetahuan dari data bidang pendidikan sebagai atribut pembelajaran utama yang mempengaruhi pencapaian mahasiswa[3].

Banyak penelitian membahas mengenai prediksi kelulusan atau prestasi dengan berbagai model algoritma data mining. Pada penelitian terdahulu telah dilakukan teknik prediksi kelulusan, seperti:

Dalam penelitian[5] menyebutkan bahwa data proses masuk, asal sekolah, kota asal dan program studi menjadi pertimbangan dalam menghitung tingkat kelulusan. Dalam penelitiannya menggunakan metoda *associasion rule* serta *algoritma apriori*. Dimana dalam metoda dan algoritma ini masing-masing faktor tersebut dicari nilai *support* dan *confidence* nya sehingga akan diperoleh faktor mana yang paling berperan atau paling mempunyai pengaruh yang cukup besar berkaitan dengan tingkat kelulusan mahasiswa.

Dalam Penelitian [6] tentang prediksi kinerja akademik mahasiswa dengan algoritma K-Means clustering yang hasilnya setelah proses clustering mahasiswa akan dikelompokkan ke dalam tiga kategori, yaitu kategori mahasiswa yang memiliki kinerja tinggi dengan hasil 46,67%, sedang 45%, dan rendah 8,33% sehinga mempengaruhi ketepatan waktu lulus mahasiswa.

Hijazi dan Naqvi[7] dalam penelitian prediksi kinerja mahasiswa mengambil sampel data mahasiswa sebanyak 300 mahasiswa yang terdiri dari 225 laki-laki, 75 perempuan menghasilkan bahwa kehadiran mahasiswa, jam dihabiskan untuk belajar setiap hari setelah kuliah, pendapatan keluarga, usia ibu dan pendidikan ibu secara signifikan mempengaruhi kinerja mahasiswa. Hasil analisis dengan menggunakan *linear regression* sederhana, ditemukan bahwa faktor-faktor seperti pendidikan ibu dan pendapatan keluarga mahasiswa sangat berkorelasi dengan prestasi akademik mahasiswa sebesar 75%.

Menurut Peneliti Jonh Fredrik Ulysses[8] memecahkan masalah bagaimana mengkasifikasikan dan memprediksi lama masa studi mahasiswa berdasarkan jalur penerimaan mahasiswa metode yang digunakan *Naïve Bayes*. Hasil penelitian ini adalah

Dengan mengunakan *Naïve Bayes* mahasiswa yang masuk melalui jalur khusus memiliki kecenderungan untuk lulus lebih cepat dibanding mahasiswa SPMB, dari percobaannya mahasiswa yang menggunakan jalur khusus diprediksi hampir mencapai destiny 1 atau 90% lulus dengan waktu 5 semester sedangkan mahasiswa jalur SPMB 54% lulus dengan 6 semester, 31% lulus dengan waktu 7 semester dan dibawah 15% lulusan dengan 8 semester.

Melihat kemampuan data mining dengan algoritma *naïve bayes*, dalam melakukan klasifikasi, maka penelitian ini memfokuskan pada data mining algoritma untuk prediksi mahasiswa lulus tepat waktu. Penelitian yang sudah dilakukan menitik beratakan pada faktor eksternal mahasiswa contohnya pada variable IPK atau pun IPS mahasiswa. Penelitian ini menggunakan data pendidikan berupa data alumni D-3 STMIK YMI Tegal, Jurusan Manajemen Informatikadari tahun masuk 1999sampai 2014. Sedangkan metode yang digunakan adalah metode *Naive Bayes* yang merupakan salah teknik pengklasifikasian dalam data mining, menganalisis untuk memperoleh informasi terhadap kasus lama masa studi mahasiswa berdasarkan faktor internal mahasiswa NIM, IPS1, IPS2, IPS3, IPS4, tahunmasuk, tahunlulus dan jeniskelamin serta memasukan faktor eksternalyaitu status pekerjaan danstatus pernikahan.

2. TINJAUAN PUSTAKA

2.1. Algoritma Klasifikasi Naive Bayes

Klasifikasi Bayesian didasarkan pada teorema Bayes.Studi yang membandingkan algoritma-algoritma klasifikasi telah menemukan sebuah klasifikasi Bayes yang sederhana yang dikenal

sebagai klasifikasi Naive Bayes yang dapat dibandingkan *performance*-nya dengan klasifikasi keputusan dan jaringan syarat tiruan.

Klasifikasi Bayes juga telah memperlihatkan keakurasian yang tinggi dan kecepatan yang baik ketika di jalankan pada database yang besar[16].

a. Bayesian Classification

P(H | X) Kemungkinan H benar jika X. X adalah kumpulah atribut.
P(H) Kemungkinan H di data, independen terhadap X
P (Single|muka sayu,baju berantakan, jalan sendiri) = nilainya besar
P ("Non Single"|"muka ceria", "baju rapi", "jalan selalu berdua") = nilainya besar
P ("Single") = jumlah single / jumlah mahasiswa
P(H | X) = posterior
P(H) = a priori
P (X | H) probabilitas X, jika kita ketahui bahwa H benar = data training
Kegiatan klasifikasi: kegiatan mencari P (H | X) yang paling maksimal
Teorema Bayes:

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

Keterangan:

X	:data dengan class belum di ketahui
H	:hipotesis data X merupakan suatu class spesifikasi
P(H|X)	:probabilitas hipotesis H berdasar kondisi X (posteriori probability)
P(H)	:probabilitas hipotesis H (posteriori probability)
P(X|H)	:probabilitas X berdasar kondisi H
P(X)	:probabilitas dari X

b. Klasifikasi

X = (muka cerah, jalan sendiri,baju rapi) Kelasnya Single atau Non Single?
Cari: P(H|X) yang paling besar:
(Single | muka cerah,jalan sendiri, baju rapi)
("Non Single" | "muka cerah", "jalan sendiri", "baju rapi")

c. Naïve Bayes

Jika bentuknya kategori , P(xk|Ci) = jumlah kelas Ci yang memiliki xk dibagi | Ci | (jumlah anggota kelas Ci di data contoh) Jika bentuknya *continous* dapat menggunakan distribusi Gaussian.

Rumus Naïve Bayes:

P(X|H)= P(H|X)P(X) (2)

Keterangan:

X	: data dengan class yang belum diketahui
H	: hipotesis data X, merupakan suatu class yang spesifik
P(H|X)	: probalitas hipotesis H berdasarkan kondisi X (*posteriori probability*)
P(H)	: probabilitas hipotesis (*posteriori probability*)
P(X|H)	: probabilitas X berdasarkan kondisi pada hipotesis H
P(X)	: probabilitas dari X

2.2. Evaluasi Algoritma Klasifikasi Data Mining

2.2.1. Evaluasi *Confusion Matrix*

Untuk melakukan evaluasi terhadap model klasifikasi berdasarkan perhitungan objek testing yang mana diprediksi benar dan tidak benar.*Confusion Matrix* berisi tentang aktual (*actual*) dan prediksi (*predicted*) pada system klasifikasi. Kinerja system seperti ini biasanya dievaluasi dengan menggunakan data pada matriks. Perhitungan ini ditabulasikan kedalam table yang disebut *confusion matrix*[17]. Bentuk *confusion matrix* dapat dilihat pada table berikut ini:

Tabel 1 Confusion Matrix[17]

Classification	Predicted class		
Observed class		Class=yes	Class=no
	Class=yes	A *(true positive-tp*	B *(false negative-fn)*
	Class=no	C *(false positive-fp)*	D *(true negative-tn)*

Setelah dilakukan confusion matrix berikutnya akan dihitung *accuracy, sensitivity, specificity, PPV, NPV. Sensitivity* digunakan untuk membandingkan jumlah true positives terhadap jumlah tupel yang *positives* sedangkan *specificity* adalah perbandingan jumlah true negatives terhadap jumlah tupel yang negatives. Sedangkan untuk PPV (*Positives Predictive Value* atau nilai prediktif positif) adalah proposi khusus dengan hasil diagnose positif, NPV (*Negatives Predictive Value*) adalah proposi khusus dengan hasil diagnose negative[17]. Berikut perhitungaanya:

Keakuratan *(Accuracy)* adalah proporsi jumlah prediksi yang benar.Hal ini ditentukan dengan menggunakan rumus accuracy.

$$Accuracy = \frac{a + b}{a + b + c + d} \quad \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Sensitivity = \frac{Number\ of\ true\ possitive}{Number\ of\ true\ positive + number\ of\ false\ negative} \quad (4)$$

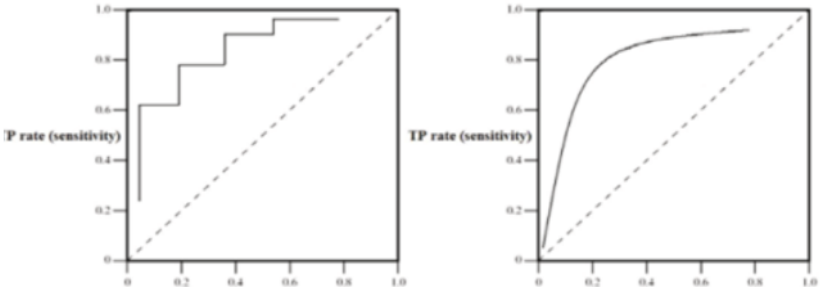
$$Specificity = \frac{Number\ of\ true\ possitive}{Number\ of\ true\ negative + number\ of\ false\ negative} \quad (5)$$

$$PPV = \frac{Number\ of\ true\ possitive}{Number\ of\ true\ possitive + number\ of\ false\ possitive} \quad (6)$$

$$NPV = \frac{Number\ of\ true\ negative}{Number\ of\ true\ negative + number\ of\ false\ negative} \quad (7)$$

2.3. Evaluasi ROC Curve

Kurve ROC *(Receiver Operating Characteristic)* adalah ilustrasi grafis dari kemampuan diskriminan dan biasanya diterapkan untuk masalah klasifikasi biner. Secara teknik kurva ROC juga disebut grafik ROC, dua dimensi grafik yaitu TP rate diletakan pada sumbu Y, sedangkan FP rate diletakan pada sumbu X. grafik ROC menggambarkan trade-off antara manfaat *(true positives)(true positives)* dan biaya *('false positives').* Berikut tampilan dua jenis kurva ROC *(discrete* dan *continuous)*



Gambar 1 Grafik ROC (discrete/continuous case) [17]

Dari gambar diatas ada beberapa hal yang perlu diperhatikan, untuk kordinat titk kiri bawah (0,0) yaitu diatara nilai TP dan FP, titik (1,1) merupakan klasifikasi positif. Titik (0,1) merupakan klasifikasi sempurna (yaitu tidak ada FN dan tidak ada FP) yang bener-benar acak akan memberikan titik sepanjang garis diagonal dari kiri bawah ke suduk kanan atas.

Dapat disimpulkan bahwa satu point pada kurva ROC adalah lebih baik dari pada yang lainnya jika arah garis melintang dari kiri bawah ke kanan atas didalam grafik. Tingkat AUC dapa di diagnose sebegai berikut[17]:

0.90-1.00 *Excellent classification* (paling baik)
0.80-0.90 *Good classification* (Baik)
0.70-0.80 *Fair classification* (sama)
0.60-0.70 *Poor classification* (Rendah)
0.50-0.60 *Failure* (Gagal)

3. METODE PENELITIAN

Dalam penelitian ini metode yang digunakan untuk prediksi kelulusan tepat waktu mahasiswa pada objek penelitian Mahasiswa STMIK-YMI Tegal menggunakan Algoritma Naïve Bayes dengan langkah-langkah penyelesaian sebagai berikut:

1. Tahap 1 menghitung prior *probability* dari setiap kelas/label

2. Tahap 2 menghitung probabilitas hipotesis (*posteriori probability*)
   Dalam tahap 2 ini perlu dicari standar deviasi dari masing masing clas variabel yang bernilai numeric menggunakan persamaan sebagai berikut:

$$Standar\ Deviasi = \sqrt{\frac{\sum y^2 - \frac{(\sum y)^2}{n}}{n-1}} \quad (8)$$

   Akan tetapi jika atribut ke i bersifat diskret, maka $P(x_i|C)$ diestimasi sebagai frekuensi relatif dari sampel yang miliki nilai $x_i$sebagai atribut i dalam class C dan jika atribut ke-i bersifat kontinu, maka $P(x_i|C)$ diestimasi dengan fungsi densitas gauss sebagai berikut:

$$f(x|y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (9)$$

   Keterangan:

   σ = Deviasi Standar
   π = Mean , Menyatakan rata − rata dari seluruh atribut
   x = nilai atribut i
   e = 2,7183
   f = Peluang
   y = Kelas yang dicari

3. Membandingkan Hasil class

4. HASIL DAN PEMBAHASAN

4.1. Hasil Analisa Data

Proses pengumpulan data yang masih asli di dapat dari BAAK diperoleh data mahasiswa lulus dari tahun 1999 sampai 2014 pada D3 Manajemen Informatika dan Komputer STMIK YMI sebanyak 510 record.

4.2. Tahap modelling

Tahap ini juga dapat disebut tahap learning karena pada tahap ini data diklasifikasikan oleh model yang kemudian menghasilkan aturan aturan.Berikut ini adalah penjelasan lebih terperinci langkah-langkah prediksi kelulusan menggunakan algoritmaNaive Bayes dengan menggunkan sampel data sebanyak 510 record data lulusan.

a. Menghitung Prior Probabilitas dari dari setiap kelas yang ada, jumlah kasus untuk Jenis Kelamin, Status Kerjaan, Status Perkawainan untuk sedangkan kelas yang ada adalah lulus tepat waktu dan mahasiswa lulus tidak tepat waktu dengan mengacu pada data yang digunakan, maka bisa di dapatkan probabilitas dari atribut atribut tersebut sebagai berikut:

Tabel 2 Probabilitas Jenis Kelamin, Status Kerja dan Status Perkawinan untuk setiap kategori keterangan Lulus

Jenis Kelamin			Status Kerja			Status Perkawinan		
	Tepat Waktu	Tidak Tepat		Tepat Waktu	Tidak Tepat		Tepat Waktu	Tidak Tepat
L	211	20	Kerja	50	2	Nikah	11	1
P	263	16	Tidak Kerja	425	33	Belum Nikah	464	34
Jumlah	474	36		475	35		475	35
Nilai Probabilitas								
L	212/475	19/35	Kerja	50/475	2/35	Nikah	11/475	1/35
P	263/475	16/35	Tidak Kerja	425/475	33/35	Belum Nikah	464/475	34/35

b. Menghitung standar Deviasi dan Maen dari masing masing variabel terikat
Mean dan standar deviasi variabel yang bernilai kontinu antara lain IPS1, IPS2, IPS3 dan IPS4 pada setiap kategori. Berikut proses perhitungan yang menunjukan nilai mean dan standar deviasi menggunakan persamaan(8):

Dengan menggunakan rumus persaan diatas berikut perhitungan untuk Standar Deviasi IPS1 Kategori Tepat Waktu:

$$S = \sqrt{\frac{3782{,}76 - \frac{(6204{,}71)^2}{475}}{475-1}} \; S = \sqrt{\frac{3782{,}76 - \frac{175}{475}}{475-1}} = \sqrt{\frac{3782{,}76 - 3682{,}84}{474}} = \sqrt{\frac{99{,}92}{474}} = 0.46$$

Dengan menggunakan rumus persaan yang sama diatas berikut perhitungan untuk Standar Deviasi IPS1 Kategori Tidak Tepat Waktu:

$$S = \sqrt{\frac{182{,}81 - \frac{(78{,}77)^2}{35}}{74-1}} \; S = \sqrt{\frac{182{,}81 - \frac{6204{,}71}{35}}{35-1}} = \sqrt{\frac{182{,}81 - 177{,}278}{34}} = \sqrt{\frac{5{,}53}{34}} = 0{,}40$$

Sedangkan hasil penerapat rumus persamaan (8) untuk standar deviasi IPS2, IPS3 dan IPS4 dengan kategori tepat waktu dan tidak tepat waktu di dapat sebagai berikut:

Tabel 3 Mean dan Standar Deviasi untuk setiap IPS dengan masing-masing kategori

Standar Deviasi dan Mean setiap IPS dengan kategori Lulusan								
	IPS1		IPS2		IPS3		IPS4	
	Tepat	Tidak Tepat	Tepat	Tidak Tepat	Tepat	Tidak Tepat	Tepat	Tidak Tepat
Mean	2,79	2,24	2,67	2,00	2,76	2,00	2,78	1,64
StanDev	0,46	0,40	0,49	0,61	0,50	0,70	0,57	0,84

Akan tetapi jika atribut ke i bersifat diskret, maka $P(x_i|\mathrm{C})$ diestimasi sebagai frekuensi relatif dari sampel yang miliki nilai $x_i$sebagai atribut i dalam class C dan jika atribut ke-i bersifat kontinu, maka $P(x_i|\mathrm{C})$ diestimasi dengan fungsi densitas gauss sebagai berikut menggunakan persamaan (9) Dengan aturan aturan tersebut jika diberikan data baru yang terdapat pada tabel dibawah ini maka prediksi lulusan dengan katergori dapat ditentukan menggunakan langkah sebagai berikut:

Tabel 4 Data baru yang belum di ketahui keterangan lulusnya

NIM	Sex	Status Kerja	Status Pernikah	IPS1	IPS2	IPS3	IPS4	Keterangan
11237012	L	Kerja	Belum Nikah	2,5	2,3	3	1,89	?

c. Dari kasus diatas makan akan dilakukan proses menentukan Nilai Probabilitas dari variabel yang bersifat kontinu yaitu IPS1 s/d IPS4 berikut proses perhitungan probabilitas masing masing IPS:

1. Probabilitas IPS1 untuk Tepat Waktu
   Diketahui:
   Standar deviasi $\sigma = 0{,}46$
   Mean $\mu = 2{,}79$

$$f(ips1 = 2{,}50|TepatWaktu) = \frac{1}{\sqrt{2\pi(0{,}46)}} e^{-\frac{(2{,}50-2{,}79)^2}{2(0{,}46)^2}}$$

$$= \frac{1}{\sqrt{2{,}8888}} 2{,}7183^{-\frac{(3{,}50-2{,}79)^2}{2(0{,}46)^2}}$$

$$= \frac{1}{\sqrt{2{,}888}} 2{,}7183^{-0{,}199}$$

$$= \frac{1}{1{,}6996} 2{,}7183^{-0{,}199}$$

= 0,588357457* 0,819775025
= 0,48

IPS1 untuk Tidak Tepat Waktu
Diketahui:
Standar deviasi $\sigma = 0{,}40$
Mean $\mu = 2{,}24$

$$f(ips1 = 2{,}50|TidakTepatWaktu) = \frac{1}{\sqrt{2\pi(0{,}40)}} e^{\frac{(2{,}50-2{,}24)^2}{2(0{,}40)^2}}$$

$$= \frac{1}{\sqrt{1{,}58}} 2{,}7183^{-\frac{(2{,}50-2{,}24)^2}{2(0{,}40)^2}}$$

$$= \frac{1}{\sqrt{1{,}58}} 2{,}7183^{-0{,}211}$$

$$= \frac{1}{0{,}63} 2{,}7183^{-0{,}211}$$

= 0,630943081* 0,809570505
= 0,510793

Dengan cara yang sama di atas maka didapat nilai probailitas dari IPS2 sampai 4 sebagai berikut:

- IPS2 untuk tepat waktu sebesar 0,428655592, Probabilitas IPS2 untuk Tidak Tepat Waktu sebesar 0,452723882,
- IPS3 untuk tepat waktu sebesar 0,502926034, Probabilitas IPS3 untuk Tidak Tepat Waktu sebesar 0,171913729
- Probabilitas IPS4 untuk Tepat Waktu sebesar 0,197394591,Probabilitas IPS4 untuk Tidak Tepat Waktu0,401147959

d. Membandingkan Hasil class Tepat waktu dan Tidak Tepat waktu:

Likelihood Tepat Waktu=P(L|TempatWaktu) * P(Kerja|TepatWaktu) *P(single|TepatWaktu)* P(IPS1|TepatWaktu) * P(IPS2|TepatWaktu) *P(IPS3|TepatWaktu) * P(IPS4|TepatWaktu)
Jadi: 212/475*50/475*463/475*0,48*0,428655592*0,502926034*0,197394591
= 0,000935395

Likelihood Tidak Tepat Waktu= P(L|TidakTempatWaktu) * P(Kerja|TidakTepatWaktu) * P(single|TidakTepatWaktu) * P(IPS1|TidakTepatWaktu) * (IPS2|TidakTepatWaktu) * P(IPS3|TidakTepatWaktu) * P(IPS4|TidakTepatWaktu)
Jadi: 19/35*3/35*34/35*0,510793*0,452723882*0,171913729*0,401147959
= 0,000720847

Probabilitas of Tepat Waktu: $\frac{0,000935395}{0,000935395 + 0,000720847}$=56,48%

Probabilitas Tidak Tepat Waktu:$\frac{0,000720847}{0,000720847+0,000935395}$= 43,52%

Dari contoh kasus di atas bisa di prediksi bahwa mahasiswa tersebut Hasilnya= Tepat Waktu.

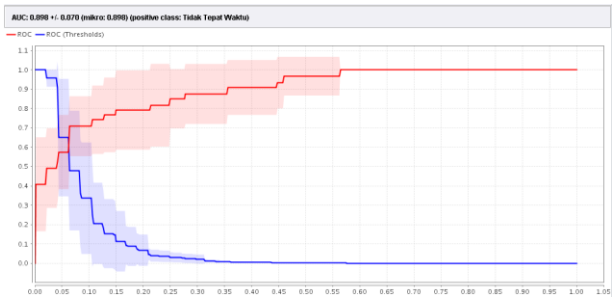
k-fold cross validation atau *n-fold cross validation* merupakan salah satu metode yang digunakan untuk mengetahui rata-rata keberhasilan dari suatu sistem dengan cara melakukan perulangan dengan mengacak atribut masukan sehingga sistem tersebut teruji untuk beberapa atribut input yang acak. *k-foldcross validation* diawali dengan membagi data sejumlah *k-fold* yang diinginkan[8]Dari pengujian 10 fold validation di dapat rata-rata sebagai berikut:

Tabel 5 Hasil Klasifikasi *k-Fold Cross Validation*

n-fold cross validation	accuracy	recall	Specificity	precision
1-fold	80,83	79,08	91,04	98,10
2-fold	94,34	97,44	50,00	96,54
3-fold	91,50	94,02	45,83	96,92
4-fold	93,25	98,13	25,81	94,81
5-fold	92,59	100,00	0,00	92,59
6-fold	92,59	100,00	0,00	92,59
7-fold	92,59	100,00	0,00	92,59
8-fold	92,59	100,00	0,00	92,59
9-fold	92,59	100,00	0,00	92,59
10-fold	89,98	90,95	75,00	98,25
Rata-Rata Hitung	91,29	95,96	28,77	94,76

Dari hasil pengujian menggunakan 10-fold Cross Validation di dapatkan rata-rata nilai *accuracy, recall, Specificity* dan *precision* untuk semua interaksi dengan nilai *accuracy, recall, Specificity* dan *precision* masing-masing 91,29%, 95.96%, 28,77% dan 94,76%. Nilai *accuracy, recall, Specificity* dan *precision* tertinggi dari semua percobaan yaitu 94,34%, 100%, 91.04% dan 98.25%. Sedangkan *nilai precision, recall*, dan *accuracy* terendah dari semua percobaan yaitu 80,83, 79.08%, 0% dan 92.55%. Sedangkan nilai *accuracy* tertinggi dihasilkan dari percobaan 10-*fold cross validation* dengan accuracy 94,34%.
Kurva ROC *(Receiver Operating Characteristic)*.

Kurva ROC dibagi dalam dua dimensi, dimana tingkat TP diplot pada sumbu Y dan tingkat FP diplot pada sumbu X. Tetapi untuk merepresentasikan grafis yang menentukan klasifikasi mana yang lebih baik, digunakan metode yang menghitung luas daerah dibawah kurva ROC yang disebut AUC *(Area Under the ROC Curve)* yang diartikan sebagai probabilitas [9]. Dari hasil perhitungan divisualisasikan dengan menggunakan kurva ROC yang bisa di lihat pada gambar dibawah ini:



Gambar 2 Graphic AUC Lulusan Tepat Waktu Naïve Bayes

Tabel 6 Hasil Perfomence Algoritma Naïve Bayes

Accuracy	91.37 %
AUC	0.898

Tabel pengujian data mahasiswa menggunakan RepidMiner dihasilkan bahwa accuracy dari algoritma naïve bayes sebesar 92.37% sedangkan pada pengujian ROC Curve menunjukan nilai AUC 0.898 rule hasil klasifikasi untuk memperdiksi lulusan tepat waktu,hasil evaluasi dengan *Confusion Matrix* dan Kurva ROC Menjelaskan bahwa garis yang berwarna merah merupakan kurva ROC dengan nilai sebesar 0.898 dengan demikian hasil klasifikasi dengan nilai *Good classification*sedangkan untuk warna biru merupakan ambang(Thresholds).

5. KESIMPULAN

Dari hasil penelitian yang dilakukan dari tahap awal sampai dengan tahap pengujian penerapan metode *naïve bayes* untuk proses prediksi kelulusan mahasiswa, didapatkan kesimpulan bahwa:

1. Metode Naïve Bayes digunakan untuk menghitung probabilitas dengan kemungkinan tepat waktu atau tidak tepat waktu dalam menentukan prediksi kelulusan mahasiswa yang dievaluasi dengan *confusion matrix*dengan menggunakan 10-*fold cross validation* dengan melibatkan faktor internal dan faktor eksternal menghasilkan akurasi klasifikasi mahasiswa lulus tepat waktu sebesar 91.37%, sedangkan evaluasi pengujian menggunakan *10-fold cross validation* menghasilkan nilai tertinggi dengan accuracy 94,34% dengan rata rata accuracy sebesar 91.29 sedangkan evaluasi dengan kurva *ROC* dengan metode *AUC* sebesar 0.898 dan ini berarti termasuk dalam *Good classification*.
2. Metode Naïve Bayes digunakan untuk menghitung probabilitas dengan kemungkinan tepat waktu atau terlambat dalam menentukan prediksi kelulusan mahasiswa menunjukan peningkatan dibanding penelitian sebelumnya yang berada pada penelitian terkait.

6. SARAN

Beberapa saran yang dapat dijadikan pertimbangan untuk penelitian selanjutnya yaitu:

1. Menambahkan beberapa atribut dalam melakukan prediksi mahasiswa lulus tepat waktu selain Indek Pretasi Semester 1 sampai dengan Indek Presasi Semester 4 yang dalam hal ini menggunakan Indek Prestasi Komulatif dengan penghasilan orang tua, domisili mahasiswa, dan faktor exsternal dan internal lainnya.
2. Menerapkan teknik penyeleksian atribut yang paling berpengaruh dengan *chi-square* sehingga tingkat akurasi bisa lebih tinggi.
3. Melakukan komparasi dari beberapa algoritma dalam klasifikasi untuk memperoleh algoritma dengan tingkat akurasi yang paling tinggi dalam memprediksi mahasiswa lulus tepat waktu.

DAFTAR PUSTAKA

[1] Peraturan Pemerintah No 66, 2010Tentang "Peraturan Pemerintah Republik Indoneisa Tentang Pengolahan Dan Penyelenggara Pendidikan".

[2] BAN. PT, Badan Akreditasi Nasional Perguruan Tinggi, 2010, Akreditasi Institusi Perguruan Tinggi - Buku III Pedoman Penyusunan Borang, pp 4.

[3] Abu Tair Mohammed M., And El-Halees Alaa M. "Mining Educational Data to Improve Students' Performance: A Case Study 2012," International Journal of Information and Communication Technology Research, Vol. 2, No. 2, pp 140-146.

[4] Ridwan Mujib, Suyono Hadi, And Sarosa, M, 2013 "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," Journal EECCIS, Vol. 7, No. 1, pp. 59-64.

[5] Salim Yeffriansjah, 2012 "Penerapan Algoritma Naïve Bayes untuk Penentuan Status Turn-Over Pegawai," Journal Media Sains, STIMIK Indonesia Banjarmasin, Vol. 4, No. 2, pp 196-205.

[6] Shovon Islam, H And Haque Mahfuza, 2012 "Prediction Of Student Academic Performance By An Application Of K-Means Clustering Algorithm," International Journal of Advanced Research in Computer Science and Software Engeneering, Vol. 2, No. 7, pp. 353–355.

[7] Hijazi Tahir, S. and Naqvi Raza, S. M. M, 2006 "Factors Affecting Students ' Performance," Bangladesh e-Journal of Sociology, Vol. 3, No. 1, pp 90-100.

[8] Ulysses Fredrik, J, 2008 "Data Mining Classification Untuk Prediksi Lama Masa Studi Mahasiswa Berdasarkan Jalur Penerimaan Dengan Metode Naive Bayes," Journal Kampus Atma Jaya Yogyakarta, Vol. 2,No.1, pp. 1–8.

[9] Hadjarawatie Lillyan, 2010 "Prediksi Dan Pemetaan Data Mahasiswa Fakultas Teknik Universitas Negeri Gorontalo Menggunakan Pendekatan Data Mining," Tesis, Universitas Negeri Gorontalo.

[10] Jananto Arief, 2013 "Algoritma Naive Bayes Untuk Mencari Perkiraan Waktu Studi Mahasiswa," Journal Teknologi Informasi DINAMIK Vol. 18, No. 1, pp. 9–16.

[11] Jindal, A, and Singh, W, 2014 "Data Mining In Education For Students Academic Performance: A Systematic Review," International Journal of Computers and Technology, Vol. 13, No. 9, pp. 5020–5028.

[12] B. Neel Mehta, 2011 "Predictive Data Mining And Discovering Hidden Values Of Data Warehouse," ARPN Journal of Systems and Software, Vol. 1, No. 1, pp. 1–5.

[13] Hamzah Amir, 2012 "Klasifikasi Teks Dengan Naïve Bayes Classifier (NBC) Untuk Pengelompokan Teks Berita Dan Abstract Akademis," Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III Yogyakarta, No.ISSN 1979-911X, pp. 269–277.

[14] Erdogan Zafer, S and Timor Mehpare, 2005 "A Data Mining Application In A Student Database," Journal of Aeronautics and Space Technologies Vol. 2, No. 2, pp. 53–57.

[15] Kulkarni, G.S, Rampure, C.G And Yadav Bhagwat, 2005 "Understanding Educational Data Mining (EDM)," International Journal of Electronics and Computer Science Engineering, ISSN 2277-1959/V2N2,pp. 773–777.

[16] Nagendra, V, K and Rajendra, C, 2012 "Customer Behaviour Analysis Using Cba (Data Mining Approach)," National Conference on Research Trends in Computer Science and Technology, Vol. 3, No. 1, pp. 65–68.

[17] Gorunesce Florin, 2012 "*Data Mining Concepts, Model And Techniques*" Intelligent Systems Reference Library, Springer-Verlag Berlin Heidelberg, Vol. 12.

[18] Tim Akdemik, 2012 "Buku Panduan Kuliah STMIK-YMI", Edisi III, Tegal,.

[19] M Ridwan, H Suyono, M Sarosa "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Naïve Bayes", Jurnal EECCISS, Vol.2.No.1.2013.