Naïve Bayes *dan Filtering Feature Selection Information Gain* untuk Prediksi Ketepatan Kelulusan Mahasiswa

Ade Ricky Rozzaqi
Program Studi Informatika, Fakultas Teknik, Universitas PGRI Semarang
Gedung Utama Lantai 3, Kampus 1 Jl. Sidodadi Timur 24, Semarang
Email: nikil.arorr988@gmail.com

ABSTRACT *- Student graduation rates is very important for the prestige of the university, student graduation rates also affect the value of accreditation a college itself, because it's by research on graduation prediction becomes very interesting to study, in this study, researchers tried to compare the two algorithms namely Naïve Bayes classification algorithm and algorithm feature Selection Information gain to obtain the highest accuracy results and outcomes AUC values were high.*

Inthis research, the processing stage by using two methods : the method that only uses Naïve Bayesalgorithm, and amethodto compare the two algorithms namely Naïve Bayes algorithm and algorithm Feature Selection Information Gain.

The results showed that the highest accuracy is obtained with a method that combines Naïve Bayes algorithm and algorithm Feature Selection Information Gain to obtain the valueof up to 89,79 % forthe use of 3 attributes, and in creased AUC increased by 3 attributes.

Keyword: *Prediction graduation, naïve Bayes, Feature Selection Information Gain.*

ABSTRAK - Tingkat kelulusan mahasiswa merupakan hal sangat penting untuk prestise suatu perguruan tinggi, tingkat kelulusan mahasiswa juga berpengaruh terhadap nilai akreditasi suatu perguruan tinggi itu sendiri, oleh karna itu penelitian tentang prediksi kelulusan menjadi hal yang sangat menarik untuk diteliti, dalam penelitian ini peneliti mencoba mengkomparasikan 2 algoritma yaitu algoritma klasifikasi Naïve Bayes dan algoritma Fitur Selection Information Gain untuk memperoleh hasil akurasi nilai tertinggi dan hasil AUC yang tinggi.

Dalam penelitian ini dilakukan tahap pengolahan dengan menggunakan dua metode yaitu: metode yang hanya menggunakan algoritma *Naïve Bayes*,dan metode yang mengkomparasikan dua algoritma yaitu algoritma *Naïve Bayes* dan algoritma *Fitur Selection Information Gain*.

Hasil penelitian menunjukan bahwa nilai akurasi tertinggi diperoleh dengan metode yang menggabungkan antara algoritma Naïve Bayes dan algoritma Fitur Selection Information Gain dengan memperoleh nilai hingga 89,79 % untuk penggunaan 3 atribut, dan peningkatan AUC meningkat dengan 3 atribut.

Kata Kunci: *Prediksi kelulusan, naïve bayes, Fitur Selection Information Gain.*

PENDAHULUAN

Latar Belakang Masalah

Pendidikan adalah suatu aktivitas sosial yang memungkinkan masyarakat tetap ada dan berkembang[1]. Jenjang pendidikan perguruan tinggi menjadi salah satu persyaratan dasar dalam mencari pekerjaan, dimana perguruan tinggi akan mempersiapkan calon-calon sarjana yang berkualitas dan mempunyai keterampilan dibidangnya. Tentunya dalam pencapaian gelar kesarjanaan tersebut membutuhkan waktu normal selama 4 tahun. Akan tetapi dalam praktiknya banyak mahasiswa tidak selalu dapat menuntaskan studinya selama waktu normal yang telah ditentukan. Banyak faktor yang menyebabkan ketidaktepatan waktu kelulusan mahasiswa tersebut, faktor-faktor tersebut dapat bersumber dari faktor internal dan faktor eksternal

Masa proses mahasiswa merupakan masa proses yang penting kaitanya dalam proses pengembangan intelektual personal untuk menghadapi tantangan dunia luar, mahasiswa juga merupakan komponen penting dalam sebuah Negara mengingat mahasiswa sebagai unsur intelektual dalam suatu Negara.

Adapun pengertian mahasiswa dalam hal ini yaitu Mahasiswa dalam peraturan pemerintah RI No.30 tahun 1990 adalah "peserta didik yang terdaftar dan belajar di perguruan tinggi tertentu"[2]. Perguruan tinggi merupakan satuan pendidikan yang menjadi terminal terakhir bagi seseorang yang berpeluang belajar setinggi-tingginya melalui jalur pendidikan sekolah[3]. Untuk tujuan memikat peminat perguruan tinggi tersebut, perguruan tinggi sanggat dituntut kualitasnya berdasarkan SDM maupun Fasilitasnya. Adapun SDM yang dimaksudkan disini adalah tenaga pendidik,kebijakan kebijakan yang diluarkanoleh pejabat structural Perguruan tinggi dll, adapun fasilitas yang di magsudkan disini adalah fasilitas yang kaitannya dengan fasilitas penunjang kegiatan belajar, fasilitas penunjang kemudahan dalam mengexpresikan kegiatan mahasiswa. Hal ini dianggap penting kepentingan untuk persaingan menarik minat calon mahasiswa baru, apalagi kaitanya dengan persentasi kelulusan mahasiswa.

Sebuah perguruan tinggi berada dalam lingkungan kompetitif yang sangat tinggi dan bertujuang untuk menghasilkan keuntungan yang lebih kompetitif melalui persaingan kompetisi bisnis lainnya.Dimana semua perguruan tinggi harus meningkatkan kualitas layanan mereka untukmendapatkan pengakuan pelayanan yang baik dimata masayarakat khususnya calon mahasiswa baru.Dimana mereka menganggap mahasiswa dan dosen sebagai asset utama mereka dan mereka ingin terus meningkatkan indikator-indikator kunci mereka dengan menggunakan asset secara efektif dan efisien[4]. Dalam Sistem pendidikan mahasiswa adalah asset penting bagi sebuah institusi pendidikan untuk itu perlu diperhatikan tingkat kelulusan mahasiswa tepat pada waktunya.

Untuk meningkatkan tingkat kelulusan untuk berbagai alasan dikemukakan oleh pengurus/pejabat sekolah,dari mulai misi masing-masing sekolah untuk mendidik siswa (yaitu menghasilkan lulusan) yang menjadi anggota produktif masyarakat dan berkontribusi terhadap kesejahteraan ekonomi bangsa. bahkan, masing-masing sekolah tahu bahwa jumlah siswa yang putus diterjemahkan sebagai hilangnya pendapatan bagi lembaga[5]. Penilaian publik terhadap kredibilitas sekolah atau institusi pendidikan sangat erat kaitannya dengan ketepatan kelulusan siswanya, sehingga berbagai upaya dilakukan sebuah sekolahan/kampus untuk mendapatkan hasil yang maksimal kaitanya

dengan ketepatan kelukusan siswa. Ketepatan kelulusan siswa juga sangat berpengaruh pada perbandingan rasio siswa/mahasiswa dengan guru/dosen.

Kelulusan tepat waktu menjadi hal yang sangat penting dan hal ini menjadi isu kebijakan yang diprioritaskan, bahkan menurut Qurdi"Tingkat kelulusan dianggap sebagai salah satu efektivitas kelembagaan" [6]. Tingkat penurunan kelulusan mahasiswa yang signifikan dan terus berkembang merupakan sebuah masalah yang ada pada perguruan tinggi.Maka dari itu pemantauan atau evaluasi terhadap kecenderungan mahasiswa lulus tepat waktu atau tidak menjadi sangat sangat vital dan hal ini menjadi tugas semua bagi semua pegawai suatu perguruan tinggi, sehingga pemaksimalan kinerja harus dilakukan. Salah satu pemaksimalan kinerja yang harus dilakukan adalah pemantauan kinerja yang melibatkan penilaian yang melayani peran penting dalam menyediakan informasi yang diarahkan untuk membantu siswa atau mahasiswa, guru atau dosen, administrator, dan pembuat kebijakan mengambil keputusan [6].

Dari uraian di atas, sangat jelas bahwa melakukan prediksi kelulusan merupakan hal yang penting bagi institusi dan berpotensi besar bagi institusi untuk menentukan kebijaksaan strategis bagi institusinya. Oleh karena itu mengidentifikasi mahasiswa mejadi jalan keluar untuk mengatasi permasalahan ini, Setelah mengidentifikasi mahasiswa yang berpotensi beresiko ketepatan waktu kelulusannya, maka intitusi bisamenggunakan mekanisme pendukung seperti orientasi, menasihati, *monitoring*, dan lain-lain untuk meningkatkan kekurangan kekurangan dari hasil indentifikasi yang dilakukan untuk bias meningkatkan ketepatan waktu lama studi. Tugas prediksi dapat dianggap sebagai menjadi dua kelas yaitu "sukses" yakni mahasiswa yang lulus tepat waktu dan "gagal" bagi mahasiswa yang lulus terlambat.

Dalam hal pengelolahan data siswa atau mahasiswa untuk mempediksi, telah diselesaikan dengan metode yang berbeda-beda yaitu menggunakan metode *neural network*[5], *decision tree*[4], *naïve bayes*[7], dan masih bayak lagi.

TujuanPenelitian

Tujuan Penelitian ini adalah mengetahui Bagaimana tingkat akurasi dan efiensi penelitian yang hanya menggunakan metode dengan algoritma *Naïve Bayes* dan penelitian yang menggunakan metode *naive bayes* dengan*Feature Selection Information Gain*

LANDASAN TEORI

Kelulusan Mahasiswa

Mahasiswa merupakan masyarakat kalangan elite dimana mahasiswa mempunyai ciri intektualitas yang lebih komplek dibandingkan kelompok seusia mereka yang bukan mahasiswa, ataupun kelompok usia dibawah dan diatas mereka. "Ciri intelektualitas tersebut adalah kemampuan mahasiswa menghadapi, memahami dan mencari cara pemecahan masalah secara lebih sistematis" [8]. Dalam setiap fakultas ataupun jurusan pada suatu universitas sangat jarang sekali bahkan tidak pernah terjadi dimana jumlah mahasiwa yang masuk dan terdaftar sama dengan jumlah dimana nantinya mahasiswa harus lulus (ketepatan waktu lama studi).

Ketepatan masa studi mahasiwa adalah hal yang sangat penting untuk diperhatikan, hal ini dikeranakan penurunan jumlah kelulusan akan menghilangkan jumlah pendapatan institusi dan mempengaruhi penilaian pemerintah serta memperngaruhi status akreditasi institusi [5]. Menurut

Suhartinah & Ernastuti ada Beberapa faktor yang dapat mempengaruhi kelulusan mahasiswa antara lain adalah nilai akhir SMA, Indeks Prestasi Semester (IPS), gaji orang tua dan pekerjaan orang tua[7].

Suatu perguruan tinggi biasanya menggunakan indeks prestasi sebagai penilaian akademik, banyak universitas memberi standar minimum yang sulit di peroleh mahasiswa[10]. Banyak variabel yang dapat digunakan dalam prediksi kelulusan mahasiswa seperti umur, status pernikahan, jumlah saudara.[11] Pada penenlitian ini parameter yang digunakan adalah jenis kelamin, Program Studi, SKS semester 1 (satu) sampai SKS semester satu 6 (Enam), jenis kelamin, IP (Indeks Prestasi)semester satu sampai IP (Indeks Prestasi)semester 6 (Enam).

Data mining

Menurut Witten Data Mining dapat difenisikan sebagai "proses penemuan pola dalam data". Dan bila menurut Daryl Pregibons dalam [12] "Data mining adalah perpaduan dari ilmu statistik, kecerdasan buatan, dan penelitian bidang database". Nama data mining berawal dari kemiripan antara pencarian informasi yang bernilai dari database yang besar dengan menambang sebuah gunung untuk sesuatu yang bernilai[13]. Diman keduanya memerlukan filtering melalui sejumlah besar atribut, atau melakukan penyelidiki dengan cerdas untuk mencari keberadaan sesuatu yang bernilai. Istilah lain dari data menurut Han[14] yaitu "knowledge mining from databases, knowledge extraction, data/pattern analysis, data archeology, dan data dredging". Banyak yang menggunakan data mining sebagai istilah popular dari KDD.

Algoritma naïve bayes

Klasifikasi Bayesian merupakan teknik prediksi berbasis probabilistic sederhana yang berdasar pada teorema Bayes (aturan bayes) dengan asumsi independensi (tidak ketergantungan). Yang kuat (naïf) dengan kata lain naïve bayes merupkan model yang menggunakan "model feature independen"

Dalam naïve bayes, hal yang dimagsudkan dari independensi yang kuat pada feature adalah bahwa sebuah fitur sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama, contoh pada kasus klasifikasi pada hewan dengan atribut, daun telinga, melahirkan, berat dan menyusui. Dalam kenyataannya hewan yang berdaun telinga dan menyusui biasanya berkembang biak dengan beraanak seperti monyet, babi, kambing , kuda dll, sebaliknya hewan yang tdk berdaun telinga dan tidak menyusui biasanya berkembang biak dengan bertelur seperti ular, burung, kadal dll. Disini ada ketergantungan pada Atribut menyusui , berdaun telinga biasanya melahirkan sebaliknya juga sama. Dalam bayes, hal tersebut tidak dipandang sehingga masing-masing fitur seperti tidak mempunyai hubungan.

Feature SelectionInformation Gain

Pada bagian ini algoritma yang dipakai dalam seleksi fitur dibahas secara singkat. Seleksi fitur, kita bias deskripsikan dengan cara formal sebagai berikut: suatu masalah dengan banyak fitur $fi \in n$ *dengan* $F=\{f1,f2,..,fk\}$, bila fitur bernilai riil (R) bisa dinyatakan sebagai satu himpunan contoh subset $V=\{v1,v2,..vn\}$ dengan $n < k$ merupakan subset kelas C dengan klasifier $K: R^k \rightarrow C$didefinisikan sebagai:

$$\forall vi \in Vi, j \in (1,..,k), vi,j \in fj \ldots\ldots\ldots\ldots\ldots (1)$$

Information gain adalah ukuran simetris, yaitu jumlah informasi yang diperoleh Y setelah mengamatai X adalah

sama dengan jumlah informasi yang diperoleh X setelah mengamati Y. Simetrical adalah properti yang diinginkan untuk mengukur feature - feature yang saling berkorelasi. Rumus untuk menghitung koefisien symetrical uncertainty

Symetrical uncertainty

$2x[\frac{\text{gain}}{H(y)+H(X)}]$… …… …… (2)

Information gain (IG) mengukur jumlah informasi dalam bit tentang prediksi kelas, jika satu-satunya informasi yang tersedia adalah adanya feature dan distribusi kelasnya. Konkretnya, mengukur pengurangan diharapkan dalam entropi (ketidakpastian yang berhubungan dengan feature yang acak)[29] mengingat S_x untuk set sampel uji $S_{Xi} = x_v|/|S_X|$ adalah contoh fungsi dari S_x dimana i^{th} yang mempunyaio nilai variabel

v: $IG(S_{X,Xi}) =$

$H(S_X) -$

$\sum_{v=values\ (X_i}^{\frac{|S_{X1}=v|}{|S_X|}} H(S_{Xi=v})$ with entropy..3)..... 3

Pengujian 10-Fold Cross Validation

Cross validation adalah metode pengamatan dalam kumpulan data yang akan digunakan untuk pengujian dan pelatihan, tetapi tidak pada waktu yang sama. Sebagai contoh, cross validation 5% berarti bahwa, untuk setiap model, 5% dari kumpulan data akan disisihkan untuk pengujian dan 95% sisanya akan digunakan untuk membangun model [16]. Secara umum pengujian nilai k dilakukan sebanyak 10 kali untuk memperkirakan akurasi estimasi.Dalam penelitian ini nilai k yang digunakan berjumlah 10 atau 10-fold Cross Validation.

Tabel 3 10-fold cross validation

DATASET									
Split 1	Split 2	Split 3	Split 4	Split 5	Split 6	Split 7	Split 8	Split 9	Split 10
TEST	TRAINING								
	TEST	TRAINING							
TRAINING		TEST	TRAINING						
TRAINING			TEST	TRAINING					
TRAINING				TEST	TRAINING				
TRAINING					TEST	TRAINING			
TRAINING						TEST	TRAINING		
TRAINING							TEST	TRAINING	
TRAINING								TEST	
TRAINING									TEST

Kurva ROC

Untuk dapat melihat akurasi secara manual dilakukan perbandingan klasifikasi menggunakan kurva ROC hasil eksperesi dari confusion matrix. Kurva ROC menghasilkan dua garis dengan bentuk true positives sebagai garis vertical dan false positives sebagai garis horizontal. Kurva ROC adalah grafik antara sensitifitas (true positives rate) pada sumbu Y dengan 1-spesifisitas pada sumbu X (false positives rate), kurva ROC ini seakan-akan menggambarkan tawar menawar antara sumbu Y atau sensitifitas dengan sumbu X atau spesifisitas. Nilai dari kurva ROC ini diharapkan mempunyai nilai akurat dalam uji kuantitas dalam sebuah pengujian antara kasus.

AUC memiliki nilai antara 0,0 dan 1,0 dengan tingkat keakuratan klasifikasi sebagai berikut:

0,90 – 1,00 = sangat baik
0,80 – 0,90 = baik
0,70 – 0,80 = sama
0,60 – 0,70 = rendah
0,50 – 0,60 = gagal

HASIL DAN PEMBAHASAN

Hasil

Tujuan dari penelitian ini adalah menerapkan Algoritma Naïve Bayes dengan

Algoritma Fitur Selection Information Gain untuk prediksi kelulusan mahasiswa, yang nantinya akan diteliti antara metode pertama (algoritma naïve bayes) dengan metode ke dua (Information Gain dan Naïve Bayes) untuk prediksi kelulusan sehingga nanti akan membandingkan dua metode ini untuk memperoleh akurasi dan AUC yang tertinggi.

Maka diharapkan dengan menggunakan algoritma fitur selection Information Gain dan algoritma Naïve bayes diharapkan akan bisa meningkatkan hasil akurasi dan Area Under Curve (AUC)

Pembahasan

Perhitungan Data Mining Algoritma Naïve Bayes

Berikut Penggunaan metode Naïve Bayes menggunakan data yang telah diacak yang memang disiapkan untuk dilakukan perhitungan manual, akan teteapi dalam perhitungan manual ini tidak bisa di jadikan acuan dalam penelitian ini hal ini dikarenakan hasil yang akan didapatkan akan berhubungan dengan jumlah total data yang akan dihitung.

Langkah pertama yang dilakukan adalah menghitung Probabilitas pada tabel data hitung manual, adapun pengertian probabilitas adalah suatu nilai untuk mengukur tingkat kemungkinan terjadinya suatu kejadian yang tidak pasti. (Johannes Supranto,2005), berikut adalah perhitungan yang dilakukan secara manual untuk menghitung probabilitas prior Untuk menghitung probabilitas suatu kejadian adalah dengan cara mencari banyaknya anggota kejadian, dibandingkan dengan banyaknya anggota ruang

Tabel 4 data training dan data testing di pilih secara acak (data hitung manual)

No	NPM	NAMA	Jurusan	SKS semstr 1	IP semstr 1	SKS semstr 2	IP semstr 2	SKS semstr 3	IP semstr 3	SKS semstr 4	IP semstr 4	SKS semstr 5	IP semstr 5	SKS semstr 6	IP semstr 6	Keterangan kelulusan
1	10210040	NUR CHIKMAH	PPKN	Paket	baik	Paket	Cukup	Paket	Cukup	kurang	baik	lebih	baik	lebih	Cukup	Tepat
2	10210041	SUSI SUSANTI	PPKN	Paket	baik	Paket	Cukup	Paket	Cukup	kurang	Cukup	kurang	baik	lebih	baik	Tepat
3	10210042	OKTAFIANA ENDAH KUSUMASTUTI	PPKN	Paket	baik	Paket	baik	lebih	baik	lebih	baik	kurang	baik	Paket	baik	Tepat
4	10210043	Didi Susanto	PPKN	Paket	baik	Paket	baik	lebih	baik	lebih	baik	Paket	baik	lebih	baik	Tepat
5	10210044	MEGA TRIANA	PPKN	Paket	baik	Paket	baik	lebih	Cukup	kurang	Cukup	kurang	sangat kurang	kurang	baik	Tepat
6	10210045	DURROTUN NAFIAH	PPKN	Paket	baik	Paket	baik	Paket	baik	lebih	baik	Paket	baik	kurang	baik	Tepat
7	10210046	AZIZ NUR ISNAINI	PPKN	Paket	baik	Paket	baik	lebih	baik	lebih	baik	Paket	baik	Paket	baik	Tepat
8	10210047	MASMUJIONO	PPKN	Paket	Cukup	Paket	Cukup	Paket	kurang	kurang	baik	lebih	baik	lebih	baik	Tepat
9	10210048	HANDY SURYA PRADANA	PPKN	Paket	kurang	kurang	Cukup	Paket	kurang	kurang	Cukup	Paket	kurang	kurang	kurang	[redacted]
10	10210049	LILA AMBAR WATIK	PPKN	Paket	baik	Paket	Cukup	Paket	Cukup	kurang	baik	Paket	baik	lebih	baik	Tepat
11	10210050	Yuyun Suciana	PPKN	Paket	baik	Paket	baik	lebih	Cukup	kurang	Cukup	Paket	baik	lebih	Cukup	Tepat
12	10210051	Hendiawan	PPKN	Paket	Cukup	Paket	baik	lebih	baik	lebih	baik	Paket	baik	Paket	baik	Tepat
13	10210052	MELIANA MAESAROH	PPKN	Paket	baik	Paket	baik	lebih	baik	Paket	kurang	kurang	baik	Paket	baik	Tepat
14	10210053	APIK DEWI APRIANTIKA	PPKN	Paket	baik	Paket	baik	lebih	baik	lebih	baik	kurang	baik	lebih	baik	Tepat
15	10210054	DWI RAMDHANI	PPKN	Paket	baik	Paket	Cukup	Paket	baik	lebih	baik	kurang	baik	lebih	Cukup	[redacted]
16	10210057	MEGA GERRIMALA	PPKN	Paket	Cukup	Paket	Cukup	Paket	kurang	kurang	Cukup	paket	baik	lebih	Cukup	Tepat
17	10210058	SONY ARIFIN	PPKN	Paket	baik	Paket	baik	lebih	Cukup	kurang	Cukup	paket	baik	lebih	Cukup	Tepat
18	10210059	UMI LATIFAH	PPKN	Paket	baik	Paket	baik	lebih	baik	lebih	baik	paket	baik	kurang	baik	Tepat
19	10430080	ENDAH NURHEDAYAH	Bhs, Jawa	Paket	baik	Paket	sangat baik	Paket	baik	Paket	baik	Paket	baik	Paket	baik	[redacted]
20	10430081	DWI THANIA NINGTYAS	Bhs, Jawa	Paket	baik	Paket	baik	Paket	baik	Paket	baik	Paket	baik	Paket	baik	Tepat

Menghitung jumlah kelas dari tahun lulus berdasarkan klasifikasi yang terbentuk (prior probability) :

1. C1 (Keterangan kelulusan = “Tepat”) = jumlah “Tepat” pada kolom Keterangan Tahun Lulus = 17/20 = 0,85

2. C2 (Keterangan kelulusan = "Terlambat") = jumlah "terlambat" pada kolom Keterangan Tahun Lulus = 3/20 = 0,15

Menghitung jumlah kasus yang sama pada setiap atribut dari keterangan (yes / no) berdasarkan data hitung.
Misal kolom jurusan :
1. C1 (PPKN = "Tepat") = jumlah "Tepat" pada kolom Keterangan Tahun Lulus = 18/18 = 1
2. C2 (PPKN = "Terlambat") = jumlah "terlambat" pada kolom Keterangan Tahun Lulus = 0/20 = 0

(Hitung Probabilitas dari Seluruh Atribut progdi, SKS semester 1 samapai 6, IPK semester 1 sampai semester 6).

Tabel 5 Probailitas keseluruhan data hitung

		jml DATA	TEPAT	TERLAMBAT	probabilitas tepat	terlambat
TOTAL		20	17	3	0.85	0.15
Jurusan	PPKN	18	18	0	1	0
	BHS, Jawa	2	1	1	0.5	0.5
sks smtr 1	Paket	20	17	3	0.85	0.15
Ip smtr 1	Baik	16	14	2	0.875	0.125
	Cukup	3	3	0	1	0
	Kurang	1	0	1	0	1
sks smtr 2	Paket	19	17	2	0.895	0.105
	Kurang	1	0	1	0	1
Ip smtr 2	sgt baik	1	0	1	0	1
	Baik	12	12	0	1	0
	Cukup	7	5	2	0.714	0.286
sks smtr 3	Paket	10	10	0	1	0
	Lebih	10	7	3	0.7	0.3
Ip smtr 3	Baik	11	9	2	0.818	0.182
	Cukup	6	6	0	1	0
	Kurang	3	2	1	0.667	0.333
sks smtr 4	Paket	3	2	1	0.667	0.333
	Lebih	8	7	1	0.875	0.125
	Kurang	9	8	1	0.889	0.111
Ip smtr 4	Baik	13	11	2	0.846	0.154
	Cukup	6	5	1	0.833	0.167
	Kurang	1	1	0	1	0
sks smtr 5	Paket	12	10	2	0.833	0.167
	Lebih	2	2	0	1	0
	Kurang	6	5	1	0.833	0.167
Ip smtr 5	Baik	18	17	2	0.944	0.111
	Kurang	2	1	1	0.5	0.5
sks smtr 6	Paket	6	5	1	0.833	0.167
	Lebih	10	9	1	0.9	0.1
	Kurang	4	3	1	0.75	0.25
Ip smtr 6	Baik	14	13	1	0.929	0.071
	Kurang	1	0	1	0	1
	Cukup	5	4	1	0.8	0.2

Langkah selanjutnya adalah kalikan semua hasil variabel
Untuk semua atribut berketerangan = "Tepat"
- P (X | Class Tahun Lulus = "yes")

= 1 X 0,5 X 0,85 X 0,875 X 1 X 0 X 0,895 X 0 X 0 X 1 X 0,714X 1 X 0,7 X 0,818 X 1 X 0,667 X 0,667 X 0,875 X 0,889 X 0,846 X 0,833 X 1 X 0,833 X 1 X 0,833 X 0,944 X 0,5 X 0,833X 0,9 X 0,75 X 0,929 X 0 X 0,8
= 0

- Untuk semua atribut berketerangan = "Terlambat"

P (X | Class Tahun Lulus = "no")
= 0 X 0,5 X 0,15 X 0,125 X 0 X 1 X 0,105 X 1 X 1 X 0 X 0,286 X 0 X 0,3 X 0,182 X 0 X 0,333X 0,333X 0,125 X 0,111X 0,154 X 0,167 X 0 X 0,167 X 0 X 0,167 X 0,111 X 0,5 X 0,167 X 0,1 X 0,25 X 0,071 X 1 X 0,2
= 0

- Perkalian prior probability dengan semua atribut Keterangan kelulusan = "Tepat"

P (Ci) | Keterangan kelulusan = "Tepat") x P(X| Keterangan kelulusan = "Tepat")
= 0,85 x 0
= 0

- Perkalian prior probability dengan semua atribut Class Tahun Lulus ="Terlambat"

P (Ci) | Keterangan kelulusan = "Terlambat") x P(X| Keterangan kelulusan = "Terlambat")
= 0,15 x 0
= 0

- Bandingkan hasil kelas

P (Ci) | Keterangan kelulusan = "Tepat") P(X| Keterangan kelulusan = "Tepat") = P (Ci) | Keterangan kelulusan = "Terlambat") P(X| Keterangan kelulusan = "Terlambat")
Kesimpulan :
(Perhitungan antara perkalian Keterangan kelulusan "Tepat" dengan Keterangan kelulusan "Terlambat" menunjukkan bahwa nilai Keterangan kelulusan = "Terlambat" sama besar dibandingkan Keterangan kelulusan "Tepat").

Implementasi dengan *RapidMiner*

Jumlah data yang sangat banyak menuntut peneliti untuk menggunakan tools pembantu yang nantinya akan mempermudah peneliti dalam menghitung seluruh data. Oleh karna itu penulis menggunggunakan tools *RapidMiner*.

Berikut adalah hasil pengolahan data dengan menggunakan naïve bayes pada *Rapid Miner* :

Tabel 6 Hasil akurasi dan AUC dari *RapidMiner* dengan *Naïve Bayes Naïve bayes*

acurasi	83.84%			
		true tepat	true terlambat	class prediction
	pred.tepat	1460	110	92.99%
	prediksi terlambat	227	289	56.01%
	class recall	86.54%	72.43%	
AUC	0.873			

Information Gain

Informasi gain adalah suatu algoritma fitur seleksi dimana algoritma ini nantinya yang akan menentukan jumlah atribut yang akan dipake, adapun perhitungan formula dari algoritma information gain

$entropy(s) = \sum_{i=n}^{n} -pi.log_2.pi$ ……*(4)*

jika diterapkan pada data yang telah di ambil secara acak sesuai perbandingan presentasi dari attribute jurusan yang terdapat bada hitungan manual seperti pada lampiran.

Sebelum sebelum menghitung algoritma Information Gain terlebeh dahulu peneliti harus mengetahui nilai entropy masing-masing, adapun formula entropy itu sendiri yaitu :

$$entropy(s) = \sum_{i=n}^{n} -pi.log_2.pi$$

$$entropy(s) = \sum_{i=n}^{n} -pi.log_2.pi$$

$$entropy(s) = \left(-\frac{78}{98} log2 \text{ X } \frac{78}{98}\right) + (-\frac{20}{98} log2 \text{ X } \frac{20}{98})$$

$$tropy(s) = 0{,}730017$$

Setelah hasil entropy keluar maka barulah memasukan nilai entropy dan data acak tersebut kedalam rumus formula information gain, yaitu:

$$Gain(S,A) = entropy(s) - \sum_{i=n}^{n} \frac{|Si|}{|s|} * Entropy\ (Si)$$

$$Gain(S,A) = 0.730017 - \{\left(\frac{6}{98}\text{X}\ 0{,}918296\right) + \left(\frac{4}{98}\text{X}0{,}811278\right) + \left(\frac{13}{98}\text{X}0{,}391244\right) + \left(\frac{15}{98}\text{X}0{,}353359\right) + \left(\frac{5}{98}\text{X}0{,}721928\right) + \left(\frac{8}{98}\text{X}0{,}543564\right) + \left(\frac{14}{98}\text{X}0{,}371232\right) + \left(\frac{4}{98}\text{X}0{,}811278\right) + \left(\frac{17}{98}\text{X}0{,}522559\right) + \left(\frac{12}{98}\text{X}0{,}811278\right)\}$$

$$Gain(S,A) = 0{,}730017 - 0{,}552661$$

$$Gain(S,A) = \ 0{,}177355$$

Tabel 7 Nilai entropy dan gain untuk menentukan simpul akar

AMPUL		DATA	TEPAT	TERLAMBAT	ENTROPY	I(s)	GAIN
TOTAL		98	78	20	0.730017		
JURUSAN	Ppkn	6	4	2	0.918296	0.552661	0.177355
	B.jawa	4	3	1	0.811278		
	B.Inggris	13	12	1	0.391244		
	B.Indo	15	14	1	0.353359		
	FISIKA	5	4	1	0.721928		
	biologi	8	7	1	0.543564		
	MTK	14	13	1	0.371232		
	Paud	4	3	1	0.811278		
	Pgsd	17	15	2	0.522559		
	BK	12	3	9	0.811278		
SKS SMT 1	Paket	98	78	20	0.730017	0	0.730017
	kurang	0	0	0	0.000000		
IP SMTR 1	Baik	50	40	10	0.721928	0.689979	0.040038
	Cukup	41	35	6	0.600609		
	Kurang	7	3	4	0.985228		
SKS SMT 2	Paket	65	69	5	0.193191	0.461983	0.268034
	Lebih	23	13	10	0.987693		
	kurang	10	5	5	1.000000		
IP SMTR 2	Baik	60	52	8	0.566510	0.626114	0.103905
	sanagt baik	2	2	1	0.500000		
	Cukup	33	25	8	0.799049		
	Kurang	3	0	3	0.000000		
SKS SMT 3	Paket	78	63	15	0.706274	0.653154	0.076865
	Lebih	11	11	0	0.000000		
	kurang	9	4	5	0.991076		
IP SMTR 3	sanagt baik	1	1	0	0.000000	0.708751	0.021266
	Baik	46	36	10	0.755375		
	Cukup	44	37	7	0.632130		
	Kurang	7	4	3	0.985228		
SKS SMT 4	Paket	50	38	12	0.795040	0.69925	0.030767
	Lebih	19	18	1	0.297472		
	kurang	29	22	7	0.797327		
IP SMTR 4	sanagt baik	39	27	12	0.890492	0.570636	0.159380
	Baik	2	2	0	0.000000		
	Cukup	33	29	4	0.532835		
	Kurang	5	1	4	0.721928		

SKS SMT 5	Paket	59	52	7	0.525451	0.665107	0.064910
	Lebih	28	19	9	0.905928		
	kurang	10	7	3	0.881291		
IP SMTR 5	sanagt baik	61	49	12	0.715322	0.722024	0.007992
	Baik	2	2	0	0.000000		
	Cukup	27	21	6	0.764205		
	Kurang	8	6	2	0.811278		
SKS SMT 6	Paket	51	45	6	0.522559	0.613204	0.116813
	Lebih	27	24	3	0.503258		
	kurang	20	9	11	0.992774		
IP SMTR 6	sanagt baik	64	53	11	0.661976	0.716765	0.013252
	Baik	1	1	0	0.000000		
	Cukup	25	18	7	0.855451		

Dari hasil perhitungan *entropy* dan *gain* yang didapat pada Tabel 4.1, terlihat bahwa atribut SKS Semester 1 yang mempunyai nilai *gain* tertinggi yaitu 0,730017. Oleh karena itu maka SKS Semester 1 merupakan simpul akar pada pohon keputusan.

Untuk pengujian data keseluruhan peneliti menggunakan tool *RapidMiner* hal ini dilakukan untuk mempermudah penelitian dikarenakan jumlah data set yang terbilang cukup besar yaitu 1920 data mahasiswa.

Hasil dari pengolahan *RapidMiner* diperoleh hasil sebagai berikut

Tabel 6 Perbandingan nilai akurasi dan AUC dua metode

“	METODE		acurasi		true tepat	true terlambat	class prediction	AUC
1	NAÏVE BAYES		83%	pred.tepat prediksi terlambat class recall	1349 221 85.92%	104 246 70.29%	92.84% 52.68%	0.864 (positive class: Terlambat)
	INFORMATION GAIN DAN NAÏVE BAYES	**K=2**	89%	pred.tepat prediksi terlambat class recall	1527 43 97.26%	166 184 52.57%	90.19% 81.06%	0.846 (positive class: Terlambat)
		K=3	89.79%	pred.tepat prediksi terlambat class recall	1516 54 96.56%	142 208 59.43%	91.44% 79.39%	0.875 (positive class: Terlambat)
		K=4	86.72%	pred.tepat prediksi terlambat class recall	1449 121 92.29%	134 216 61.71%	91.54% 64.09%	0.868 (positive class: Terlambat)
		K=5	85.05%	pred.tepat prediksi terlambat class recall	1408 162 89.68%	125 225 64.29%	91.85% 58.14%	0.865 (positive class: Terlambat)
		K=6	85%	pred.tepat prediksi terlambat class recall	1387 183 88.34%	108 242 69.14%	92.78% 56.94%	0.874 (positive class: Terlambat)
		K=7	84.79%	pred.tepat prediksi terlambat	1383 187	105 245	92.94% 56.71%	0.874 (positive class:

			class recall	88.09%	70.00%		Terlambat)
	K=8	84%	pred.tepat	1375	104	92.97%	0.865 (positive
			prediksi terlambat	195	246	55.78%	class:
			class recall	87.58%	70.29%		Terlambat)
	K=9	84.01%	pred.tepat	1364	101	93.11%	0.866 (positive
			prediksi terlambat	206	249	54.73%	class:
			class recall	86.88%	71.14%		Terlambat)
	K=10	83.49%	pred.tepat	1355	102	93.00%	0.865 (positive
			prediksi terlambat	215	248	53.56%	class:
			class recall	86.31%	70.86%		Terlambat)
	K=11	83.44%	pred.tepat	1355	103	92.94%	0.865 (positive
			prediksi terlambat	215	247	53.46%	class:
			class recall	86.31%	70.57%		Terlambat)
	K=12	83.07%	pred.tepat	1349	104	92.84%	0.864 (positive
			prediksi terlambat	221	246	52.68%	class:
			class recall	85.92%	70.29%		Terlambat)
	K=13	83.07%	pred.tepat	1349	104	92.84%	0.864 (positive
			prediksi terlambat	221	246	52.68%	class:
			class recall	85.92%	70.29%		Terlambat)

Dari hasil tabel diatas dari mulai tabel hasil pengolahan yang haya menggunakan algoritma Naïve Bayaes dan Naive Bayes yang menggunakan Fitur Selection Information Gain. Maka bisa dilihat perolehan akurasi yang tertinggi dan AUC yang tertinggi, seperti pada tabel berikut

PENUTUP

Kesimpulan

Setelah dilakukan penelitian ini yaitu membandingkan penggunaan yang hanya menggunakan algoritma Naïve Bayes dengan penggunaan algoritma information gain dan Naïve bayes untuk prediksi kelulusan maka diperoleh hasil akurasi tertinggi dengan menggunakan metode algoritma *information gain* dan *naïve bayes* sesuai pada *tabel 9* yaitu perbandingan nilai akurasi dan AUC, dengan nilai akuras tertinggi 89,79 % dengan menggunakan K=3, dan AUC tertinggi di peroleh hasil 0,875 dengan K=3

Dari hasil penelitian ini dapat di simpulkan bahwa algoritma *naïve* bayes dan metode *filtering feature selection information gain* berpengaruh pada akurasi dan AUC untuk prediksi kelulusan mahasiswa.

Saran

Dari hasil penelitian yang telah dilakukan maka muncul gagasan-gagasan yang dirangkum dalam usulan dan saran untuk penelitian yang berhubungan dengan prediksi kelulusan, antara lain:

1. Dalam Penelitian prediksi hendaknya pemilihan data dilihat nilai homogennya terlebih dahulu, karna dalam penelitian ini pengambilan data *traning* terlalu *complex*, hal ini nantinya akan sangat mempengaruhi akurasi
2. Dalam melakukan penelitian yang berkaitan dengan prediksi haruslah memilah algoritma yang sesuai dengan jenis data (algoritmayang menyesuaikan data).

DAFTAR PUSTAKA

[1] Brameld, T. 1999. Dasar Konsep Pendidikan Moral. ALFABETA: Jakarta.

[2] Peraturan pemerintah republik Indonesia no 66 tahun 2010 tentang perubahan atas peraturan pemerintah no 17 tahun 2010 tentang pengolahan dan penyelenggara pendidikan.

[3] Nawawi, H., & M, M. (1994). Kebijaksanaan Pendidikan di Indonesia di tinjau dari Sudut Hukum. Yogyakarta: Gajah Mada University Press.

[4] Qudri, M. N., & Kalyankar, N. V. (2010). Drop Out Feature of Student Data for Academic Performance Using Decision Tree techniques. Global Journal of Computer Science and Technology , 2-4.

[5] Karamouzis, T. S., & Vrettos, A. (2009). Sentivity Analysis of Neural Network for Identifying the Factors for Collage Students Success. World Congress on Computer Science and Information Engineering , 978-0-7695-3507-4.

[6] Ogor, E. N. (2007). Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques. Fourth Congress of Electronics, Robotics and Automotive Mechanics .

[7] Suhartinah, S. M., & Ernastuti. (2010). Graduation Prediction of Gunadarma University Students Using Algorithm and Naive Bayes C4.5 Algoritmh

[8] Azwar, S. (2004). *Penyusunan Skala Psikologi*. Yogyakarta: Pustaka Pelajar.

[9] Siregar, A. R. (2006). *Motivasi Belajar Mahasiswa ditinjau dari Pola Asuh*. Medan: Usu Repository.

[10] Oyelade, A. J., Oladipupo, O. O., & Obagbuwa, I. C. (2010). *Application of kmeans Clustering algorithm for predicting of Students Academic Performace*. International Journal of Computer Science and Information Security , 292-295.

[11] Yingkuachat, J., Praneetpolgrang, P., & Kijsirikul, B. (2007). An Application Probabilitic Model to the Prediction of Student Graduation Using Bayesian Belief Network. ECTI Transaction on Computer and Technology, 63-71

[12] Gorunescu, Florin (2011). Data Mining: Concepts, Models, and Techniques. Verlag Berlin Heidelberg: Springer

[13] Sumathi, & S., Sivanandam, S.N. (2006). *Introduction to Data Mining and its Applications*. Berlin Heidelberg New York: Springer

[14] Han, J.,&Kamber, M. (2006).*Data Mining Concept and Tehniques*.San Fransisco: Morgan Kauffman.

[15] Carlo Vercellis, Business Intelligence : Data Mining and Optimization for Decision Making. Milano, Italy: A John Wiley and Sons, Ltd., Publikation

[16] Myatt, A Practical Guide To Exploratory Data Analysis And Data Mining. New Jersey: John Wiley & Sons, 2007

[17] Vercellis C, Business *Intelligent: Data Mining and Optimization for Decision Making*. John Wiley & Sons, 2009