IMPLEMENTASI ALGORITMA NAÏVE BAYES UNTUK MEMPREDIKSI LAMA MASA STUDI DAN PREDIKAT KELULUSAN MAHASISWA

Windy[1)], Daniel Rudiaman Sijabat[2)], Febry Eka Purwiantono[3)]

[1, 2] Program Studi Teknik Informatika, STIKI Malang
[3] Program Studi Manajemen Informatika, STIKI Malang
Email: windychua1996@gmail.com, daniel223@stiki.ac.id, febry@stiki.ac.id

Abstract

Informatics is one of the study programs at STIKI Malang which has quite large student data both data on active students and students who have graduated. Every year students who graduate on time are far less than new students, which will affect the quality of students and also influence to the accreditation of STIKI Malang. Therefore, the purpose of this study is to make an application to predict the duration of study and the predicate of student graduation by applying supervised learning techniques. Output predictions for the period of study are on time, late, or not graduated and output for predicate are summa cum laude, cum laude, very satisfying, or satisfying. The application method used in supervised learning for prediction is a Naïve Bayes algorithm. This is used to analyze data, especially in the pattern recognition process, predicting the period of study and the predicate of graduation. Before entering the calculation phase of the Naive Bayes algorithm, the Correlation Feature Selection is used to select relevant features that function to improve the accuracy of the system. The results of the study based on 10 Fold Cross Validation testing indicate that the application can be used to assist informatics study programs in order to find strategic information related to the duration of study and the predicate of student graduation with an accuracy of 77.19% and 87.65%.

Keywords: *Prediction, Naïve Bayes, Study Periods, Correlation Feature Selection, 10 Fold Cross Validation*

1. PENDAHULUAN

Lama masa studi yang ditempuh oleh mahasiswa merupakan salah satu standar yang termasuk ke dalam standar penilaian pada Standar Nasional Pendidikan Tinggi atau SN-DIKTI untuk mengukur akreditasi perguruan tinggi tersebut yakni Mahasiswa dan Lulusan berdasarkan PERMENRISTEKDIKTI Nomor 44 Tahun 2015. Masa studi untuk program sarjana maksimal 7 (tujuh) tahun akademik dengan beban belajar mahasiswa paling sedikit 144 (seratus empat puluh empat) SKS. Mahasiswa sarjana memiliki masa studi antara 8 semester hingga 14 semester. Mahasiswa dengan masa studi lebih dari 8 semester sudah tergolong mahasiswa yang menempuh masa studi yang lama.

Pada institusi pendidikan perguruan tinggi seperti STIKI Malang, data mahasiswa dan data jumlah kelulusan mahasiswa dapat menghasilkan informasi yang berlimpah berupa jumlah kelulusan mahasiswa setiap tahunnya, profil, dan hasil akademik mahasiswa selama menempuh proses kegiatan belajar mengajar di STIKI Malang. Adanya informasi mengenai lama masa studi mahasiswa tentunya akan menjadi pendukung pengambilan keputusan yang tepat bagi manajemen STIKI Malang dalam mengambil langkah berikutnya.

Salah satu permasalahan yang sering terjadi di perguruan tinggi khususnya STIKI Malang yaitu ketidakseimbangan antara mahasiswa yang masuk dan yang lulus. Mahasiswa

yang masuk dalam jumlah banyak, namun jumlah yang lulus tepat waktu jauh lebih sedikit daripada mahasiswa yang masuk ke STIKI Malang yang akan berpengaruh terhadap mutu mahasiswa itu sendiri dan akreditasi STIKI Malang. Untuk itu diperlukan suatu sistem untuk memprediksi. Selama ini STIKI Malang belum memiliki model prediksi lama masa studi mahasiswa yang dapat digunakan untuk memprediksi mahasiswa yang lulus tepat waktu, padahal data mahasiswa yang tersedia sudah sangat berlimpah, hanya saja data-data tersebut belum dimanfaatkan untuk dianalisis lebih jauh.

Algoritma *Naïve Bayes* dapat digunakan untuk mengklasifikasi dengan waktu komputasi yang cepat, menghapus fitur yang tidak relevan akan meningkatkan kinerja klasifikasi dan dapat menghasilkan akurasi yang akurat [1].

Pada penelitian ini, algoritma *Naïve Bayes* akan digunakan untuk memprediksi masa studi mahasiswa STIKI Malang apakah tergolong tepat waktu, terlambat atau tidak lulus dan memprediksi predikat kelulusan berupa *summa cumlaude*, *cumlaude*, sangat memuaskan atau memuaskan.

2. KAJIAN LITERATUR

Pada penelitian yang berjudul Drop out Estimation Students based on the Study Period: Comparisonbetween Naïve Bayes and Support Vector Machines Algorithm Methods dilakukan perbandingan metode klasifikasi untuk mengestimasi *drop out* mahasiswa dengan menggunakan metode *Naïve Bayes* dan *Support Vector Machine* [2]. Atribut atau variabel yang digunakan adalah *Gender, High School, Majoring, NEM, Father's Education, Mother's Education, Father's Occupation, Mother's Occupation, GPA of the fourth semesters*. Hasil dari penelitian ini adalah perbandingan estimasi *drop out* mahasiswa dengan 2 algoritma. Algoritma yang digunakan yaitu algoritma *Naïve Bayes* dengan *Support Vector Machine*, dan menunjukkan bahwa *Naïve Bayes* dengan *Support Vector Machine* dapat diterapkan. Dengan akurasi masing-masing 80,67% dan 60%.

Pada penelitian yang berjudul Students performance prediction using KNN and Naïve Bayesian dilakukan perbandingan metode klasifikasi untuk memprediksi kinerja mahasiswa baru dengan menggunakan metode *K-Nearest Neighbor* dan *Naïve Bayes* [1]. Atribut atau variabel yang digunakan pada penelitian ini adalah *Gender, DOB, Specialization, City, Secondary school name, status, Father's job, student status.* Hasil penelitian ini berupa perbandingan prediksi kinerja mahasiswa baru dengan tujuan untuk membantu kementerian pendidikan dalam meningkatkan pembelajaran siswa di Gaza, dan menunjukkan *Naïve Bayes* dan *KNN* dapat diterapkan dalam kasus ini.

Sedangkan pada penelitian yang berjudul Analysis on students performance using naÏve Bayes classifier dibuat sistem prediksi untuk klasifikasi kinerja siswa pada Mata Pelajaran tertentu seperti *English, Islamic Education subject, History, Mathematics, Additional Mathematic, Physics, Chemistry* untuk kategori atau *grade* yang telah ditentukan seperti *Excelent, Good, Average, Poor* [3]. Data yang digunakan adalah data tahun 2011 hingga 2014 dengan total 488 siswa. Hasil penelitian ini adalah prediksi klasifikasi kinerja siswa, dan menunjukkan *Naïve Bayes Classifier* dapat diterapkan akurasi *Naïve Bayes* mencapai 73,4% pada kasus ini.

Selain itu, ada pula penelitian yang berjudul Aplikasi Pemrediksi Masa Studi dan Predikat Kelulusan Mahasiswa Informatika Universitas Muhammadiyah Surakarta Menggunakan Metode Naive Bayes dilakukan prediksi untuk masa studi dan predikat kelulusan pada jurusan Informatika di Universitas Muhammadiyah Surakarta dengan menggunakan metode *Naïve Bayes* [4]. Atribut yang dipakai untuk perhitungannya yaitu jurusan asal sekolah, *gender*, daerah asal, asal

sekolah, dan asisten. Hasil dari penelitian ini yaitu prediksi masa studi berupa Tepat waktu dan Terlambat serta predikat kelulusan (IPK) berupa *Cumlaude*, Sangat Memuaskan dan Memuaskan.

Machine Learning adalah bagian dari *Artificial Inteligence* di mana mesin dilatih untuk belajar dari pengalaman masa lalu. Pengalaman masa lalu dikembangkan melalui data yang dikumpulkan. Dalam *machine learning* sendiri terbagi menjadi 2 jenis yaitu *supervised learning* dan *unsupervised learning.* [5].

Supervised Learning adalah suatu teknik pembelajaran yang digunakan di mana sudah terdapat data yang dilatih, dan terdapat variabel yang ditargetkan. *Supervised Learning* biasanya digunakan untuk tujuan prediksi dengan klasifikasi yaitu mengelompokkan suatu data ke kelas-kelas data yang sudah ada (label). Algoritma dalam *supervised learning* antara lain *Support Vector Machine, Nearest Neighbor*, *Artificial Neural Network, Naïve Bayes*, *Decision Tree* dan *Fuzzy K-Nearest Neighbor*.

Unsupervised learning berbeda dengan *supervised learning,* dimana *unsupervised learning* tidak memiliki data kelas atau label sebelumnya. Data tersebut akan dikelompokkan ke dalam beberapa kelompok berdasarkan kriteria-kriterianya masing-masing yang biasa disebut juga dengan *clustering*. Algoritma dalam *unsupervised learning* antara lain *K-Means*, *Hierachical Clustering, DBSCAN, Fuzzy C-Means* dan *Self-Organizing Map*.

Naïve bayes classifier adalah suatu implementasi dari algoritma *Naïve Bayes* yang dipergunakan untuk memprediksi dengan mengklasifikasikan data. *Naïve bayes* sendiri berasal dari teorema *Bayes* (aturan *Bayes*). Aturan *Bayes* merupakan suatu teknik prediksi yang menggunakan probabilitas sederhana berdasarkan pengalaman di masa lalu. Disebut *Naïve Bayes* karena asumsi independensi yang kuat pada setiap atribut. *Naïve Bayes Classifier* diasumsikan ada atau tidaknya ciri tertentu dari satu kelas tidak berhubungan dengan ciri dari kelas yang lain [6]. Adapun rumus perhitungan *Naïve Bayes* yaitu :

$$p(Ck|xi) = \frac{p(xi|Ck)p(Ck)}{p(x)} \quad \textbf{(2.1)}$$

Keterangan

Ck	:	Kategori lama studi mahasiswa yang berupa tepat waktu, terlambat atau tidak lulus dan kategori predikat kelulusan yang berupa *summa cumlaude*, *cumlaude*, sangat memuaskan dan memuaskan.
P(Xi|Ck)	:	Probabilitas xi pada kategori Ck
Xi	:	Atribut yang mempengaruhi lama masa studi mahasiswa dan predikat kelulusan.
P(Ck)	:	Probabilitas dari kategori Ck.

Laplace Smoothing adalah sebuah metode *smoothing* yang paling umum dan disebut juga dengan *default smoothing* serta merupakan *smoothing* tertua yang pernah diimplementasikan pada *Naïve Bayes Classifier*. *Laplace Smoothing* juga disebut dengan *add one smoothing*, karena metode ini menambahkan angka 1 pada setiap perhitungan yang didapat [7].

Menurut [8], istilah Sarjana merupakan gelar strata satu yang dicapai oleh seseorang yang telah menamatkan pendidikan tingkat terakhir di perguruan tinggi. Masa studi paling lama 7 (tujuh) tahun akademik untuk program sarjana, program diploma empat/ sarjana terapan, dengan beban belajar mahasiswa paling sedikit 144 (seratus empat puluh empat) SKS. Mahasiswa yang lulus dengan masa studi melebihi 8 semester sudah termasuk mahasiswa dinyatakan terlambat dan mahasiswa yang telah menempuh masa studi melebihi 14 semester dinyatakan tidak lulus [9].

Berdasarkan Peraturan Akademik STIKI Malang pasal 30 tahun 2009, Predikat kelulusan untuk mahasiswa S1

yang lulus dengan masa studinya kurang dari atau sama dengan 9 semester (Lama Studi ≤ 9 semester) berlaku ketetapan sesuai dengan Tabel 2.1.

Tabel 2.1 Ketetapan Predikat Kelulusan STIKI Malang

Kelompok IPK	Predikat
IPK ≥ 3,75	*Summa Cum Laude*
3,5 ≤ IPK < 3,75	*Cum Laude*
2,75 ≤ IPK < 3,5	Sangat Memuaskan
2,20 ≤ IPK < 2,75	Memuaskan

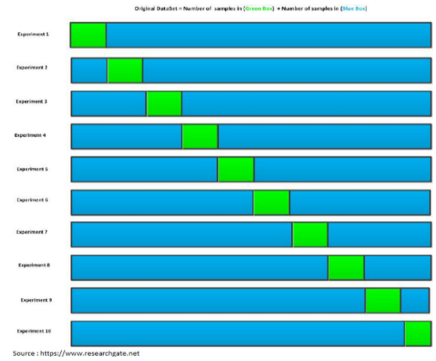
Correlation Feature Selection adalah suatu metode untuk menentukan fitur-fitur terbaik yang relevan berdasarkan korelasi terhadap suatu kelas. Seleksi berbasis korelasi jauh lebih berhasil dalam menghapus fitur yang *redundant* dan tidak relevan dari *dataset* sehingga dapat menurunkan waktu komputasi dan meningkatkan akurasi prediksi [10]. Adapun rumus perhitungan merit *Correlation Feature Selection* yaitu :

$$M_s = \frac{Kr_{cf}}{\sqrt{k+k\,(k-1)r_{ff}}} \qquad \textbf{(2.2)}$$

Keterangan

- Ms : *Heuristic Merit* dari *feature subset* s yang mengandung K fitur.
- K : Banyaknya fitur yang sedang dicari
- r_{cf} : *Mean* / rata-rata dari korelasi *feature-class*
- r_{ff} : *Mean* / rata-rata dari fitur-fitur *inter-corellation*.

10 *Fold Cross Validation* adalah salah satu K *Fold Cross Validation* yang direkomendasikan untuk pemilihan model terbaik, karena cenderung memberikan estimasi akurasi yang kurang bias. Dalam 10 *Fold Cross Validation*, data dibagi menjadi 10 *fold* berukuran kira-kira sama, sehingga akan memiliki 10 *subset* data untuk mengevaluasi kinerja model atau algoritma. Untuk masing-masing dari 10 *subset* data tersebut, *cross validation* akan menggunakan 9 *fold* untuk pelatihan dan 1 *fold* untuk pengujian, kemudian didapatkan tingkat akurasi dari setiap 10 *subset* yang nantinya akan dirata-rata [11]. Gambaran 10 *fold cross validation* dapat dilihat pada Gambar 2.1.



Gambar 2.1 Gambaran 10 *Fold Cross Validation*

Pengukuran kinerja akurasi menggunakan *Confusion matrix*, terdapat empat istilah sebagai representasi dari hasil proses prediksi yaitu *True Positive (TP), False Positive (FP), False Negative (FN),* dan *True Negative (TN)* [12].

Keterangan

- *TP* : *True Positive, TP* menyatakan jumlah data Kelas A yang diprediksi benar yaitu Kelas A
- *FN* : *False Negative, FN* menyatakan jumlah data Kelas A yang diprediksi salah yaitu Kelas B
- *FP* : *False Positive, FP* menyatakan jumlah data Kelas B namun diprediksi salah yaitu Kelas A
- *TN* : *True Negative, TN* menyatakan jumlah data Kelas B yang diprediksi benar yaitu Kelas B.

Nilai akurasi menggambarkan seberapa akurat sistem dapat memprediksi data secara benar [13]. Dengan kata lain, nilai akurasi merupakan perbandingan antara data yang terprediksi benar dengan jumlah keseluruhan data.

Menurut [12], rumus akurasi adalah sebagai berikut :

$$\text{Akurasi} = \frac{\sum_{i=1}^{l} \frac{TP_i + TN_i}{(TP_i + TN_i + FP_i + FN_i)}}{l} \text{ X } 100\% \quad \textbf{(2.3)}$$

Keterangan

TP_i : *True Positive* untuk kelas i
FN_i : *False Negative* untuk kelas i
FP_i : *False Positive* untuk kelas i
TN_i : *True Negative* untuk kelas i
L : Jumlah kelas

3. METODE PENELITIAN

3.1. Analisa Masalah

Siswa-siswa sekolah menengah atas (SMA/sederajat) yang melanjutkan studi ke STIKI Malang setiap tahun terus mengalami peningkatan, namun mahasiswa yang lulus atau yang telah menyelesaikan studi STIKI Malang setiap tahun tidak sebanding dengan mahasiswa barunya. Jika terus dibiarkan, akan mempengaruhi mutu dan kualitas dari mahasiswa itu sendiri dan akreditasi STIKI Malang. Ada berbagai penyebab mahasiswa tidak dapat menyelesaikan studi tepat waktu atau bahkan di *drop out*, di antara lain yaitu nilai mata kuliah mata kuliah yang tidak lulus, nilai indeks prestasi yang rendah, dan jumlah pengambilan SKS yang sedikit

3.2. Pemecahan Masalah

Berkaitan dengan analisa masalah di atas, maka didapat pemecahan masalahnya yaitu dengan sebuah sistem prediksi untuk memprediksi masa studi dan predikat kelulusan mahasiswa STIKI Malang. Sehubungan dengan hal tersebut, maka diambil data dari semester awal yaitu semester 1 dan semester 2.

Pada permasalahan yang akan diteliti, akan dilakukan pencarian dan perhitungan fitur terbaik berdasarkan tingkat korelasi (*Correlation Feature Selection)* terhadap kategori masa studi dan predikat kelulusan. Pemilihan fitur pada perhitungan CFS dengan menghitung merit terbesar berdasarkan rumus 2.1, dipilih dengan merit maksimal (terbesar). Jika pada perhitungan berikutnya (K+1) menghasilkan merit lebih kecil, maka pemilihan fitur berakhir dengan fitur-fitur pada perhitungan K.

Untuk memprediksi lama masa studi mahasiswa digunakan *Naïve Bayes Classifier* pada data mahasiswa. *Naïve Bayes Classification* merupakan suatu teknik prediksi dengan menggunakan probabilitas sederhana berdasarkan pengalaman masa lalu dengan asumsi independensi (ketidaktergantungan) yang kuat pada setiap atribut.

Dengan adanya data atau informasi mengenai lama masa studi mahasiswa seorang mahasiswa, tentunya akan menjadi pendukung pengambilan keputusan yang tepat bagi manajemen STIKI Malang dalam mengambil langkah berikutnya.

3.3. Pengumpulan Data

Pengumpulan data dilakukan dengan mengumpulkan data-data mahasiswa Teknik Informatika STIKI Malang angkatan 2008, 2009 dan 2010 dengan jumlah total 354 data untuk prediksi masa studi serta 197 data untuk prediksi predikat pada Badan Administrasi dan Akademik STIKI Malang. Pengumpulan data mahasiswa tersebut dilakukan dengan cara *query* pada *database* mahasiswa STIKI Malang oleh pihak Badan Administrasi dan Akademik, dan diketik kembali ke Microsoft Excel. Data mahasiswa yang telah terkumpul, dilabeli sesuai dengan kelas masa studi yaitu tepat waktu, terlambat dan tidak lulus. Data dengan kelas masa studi tepat waktu dan terlambat, dilabeli kembali sesuai dengan kelas predikat kelulusan yaitu predikat *summa cumlaude*, *cumlaude*, sangat memuaskan, dan memuaskan.

3.4. Data *Preprocessing*

Dalam tahap ini, data akan menjadi sangat penting untuk proses pembangunan pengetahuan dan proses prediksi. *Preprocessing* ini dibagi menjadi 2 tahapan, tahap-tahapnya adalah sebagai berikut :

- Data *cleaning*
  Pembersihan data dengan cara menghilangkan *noise* dan data yang tidak konsisten ataupun kosong dan tidak relevan. Data-data tersebut akan dihapus.

- Data *transformation*
  Isi data mahasiswa berupa jenis kelamin, IP Semester dan nilai mata kuliah akan diubah ke dalam bentuk yang telah ditentukan yang dapat dilihat pada Tabel 3.1, Tabel 3.2, dan Tabel 3.3.

Tabel 3.1 Data *Transformation* Jenis Kelamin

Atribut	Data Aktual	Data *Transformation* (Kode)
Jenis Kelamin	Laki-laki	1
	Perempuan	2

Tabel 3.2 Data *Transformation* Nilai IP Semester

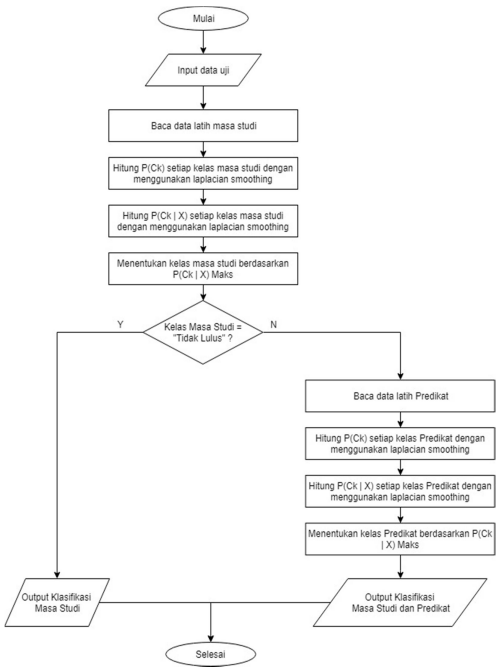
Atribut	*Range* Data Aktual	Data *Transformation* (Kode)
Nilai IP Semester	IP = 0	0
	0 < IP ≤ 1	1
	1 < IP ≤ 2	2
	2 < IP ≤ 2,5	2,5
	2,5 < IP ≤ 3	3
	3 < IP ≤ 3,5	3,5
	3,5 < IP ≤ 4	4

Tabel 3.3 Data *Transformation* Nilai Mata Kuliah

Atribut	Data Aktual	Data *Transformation* (Kode)
Nilai Mata Kuliah	E	0
	D	1
	C	2
	C+	2,5
	B	3
	B+	3,5
	A	4

3.5. Perancangan Algoritma *Naïve Bayes*

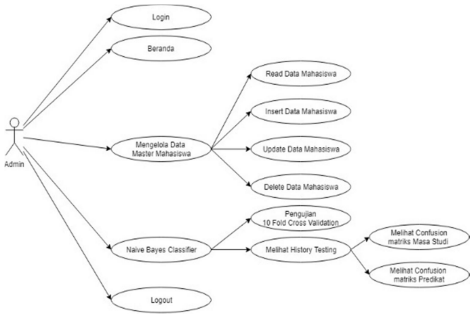
Pada proses ini, data uji akan melewati proses prediksi berdasarkan data latih. *Flowchart* untuk tahap prediksi dapat dilihat pada Gambar 3.1.



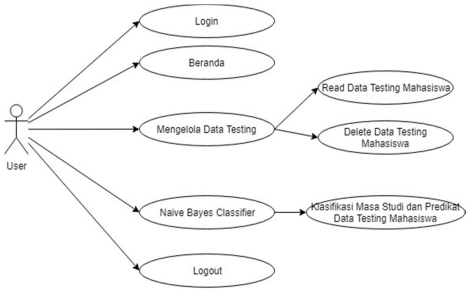
Gambar 3.1 *Flowchart* Perancangan Algoritma *Naïve Bayes*

3.6. *Use Case* Sistem Aplikasi

Diagram *use case* admin dan *user* dari sistem aplikasi prediksi lama masa studi dan predikat kelulusan mahasiswa dapat dilihat pada Gambar 3.3 dan Gambar 3.4.



Gambar 3.3 *Use Case Diagram* Admin



Gambar 3.4 ***Use Case Diagram User***

4. HASIL DAN PEMBAHASAN

4.1. Pencarian Fitur Dengan Perhitungan *Correlation Feature Selection*

Untuk mencari fitur-fitur terbaik yang relevan terhadap masa studi dan predikat kelulusan, akan dilakukan perhitungan dengan menggunakan *Correlation Feature Selection* terhadap 15 fitur data mahasiswa, yaitu nilai MK Logika Algoritma, nilai MK Aljabar Linier, nilai MK Bahasa Inggris 1, nilai MK Pemrograman Dasar 1, nilai MK Pemrograman Dasar 2, nilai IP semester 1, nilai IP semester 2, SKS lulus semester 1, SKS lulus semester 2, SKS lulus hingga semester 2, umur mahasiswa saat mendaftar, jenis kelamin mahasiswa, pekerjaan ayah, sekolah asal, dan jurusan asal.

4.1.1 Perhitungan *Correlation Feature Selection* Terhadap Masa Studi

Adapun perhitungan CFS dengan cara menghitung merit berdasarkan rata-rata dari *feature-class* dan rata-rata dari fitur-fitur *inter-correlation* terhadap masa studi berdasarkan korelasi semua fitur terhadap kelas masa studi.

Gambar 4.1 adalah perhitungan rumus pencarian *Correlation Feature Selection* pada rumus (2.2) dengan mencari merit terbesar pada fitur terbaik yang pertama, didapatkan fitur pertama yang paling baik untuk masa studi yaitu SKS Lulus Semester 2.

		k	Rcf	Rff	Merit		MAX
NumLogic	0,550606805	1	0,550606805	1	0,550606805		
NumLinier	0,59107409	1	0,59107409	1	0,59107409		
NumEnglish	0,459235129	1	0,459235129	1	0,459235129		
NumProgOne	0,596337346	1	0,596337346	1	0,596337346		
NumProgtwo	0,654951442	1	0,654951442	1	0,654951442		
RangeIPSOne	0,591445985	1	0,591445985	1	0,591445985		
RangeIPSTWO	0,691755439	1	0,691755439	1	0,691755439		
SKLONE	0,524396782	1	0,524396782	1	0,524396782		
SKTWO	0,705351761	1	0,705351761	1	0,705351761		0,705351761
KOMSKS	0,677995278	1	0,677995278	1	0,677995278		
UMUR	-0,116517201	1	-0,116517201	1	-0,116517201		
JK	0,248724084	1	0,248724084	1	0,248724084		
FATHERJOB	0,007310865	1	0,007310865	1	0,007310865		
SMA	-0,114604518	1	-0,114604518	1	-0,114604518		
JURUSANSMA	-0,013799897	1	-0,013799897	1	-0,013799897		

Gambar 4.1 Perhitungan CFS - 1 Terhadap Masa Studi

Selanjutnya dengan rumus yang sama pada perhitungan *Correlation Feature Selection* yaitu rumus (2.2), dicari fitur terbaik yang kedua setelah SKS Lulus Semester 2 dapat dilihat pada Gambar 4.3.

			k	Rcf	Rff	Merit		MAX
SKTWO	NumLogic	0,627979283	2	0,627979283	0,673138307	0,68658513		
SKTWO	NumLinier	0,648212926	2	0,648212926	0,779255592	0,687248048		
SKTWO	NumEnglish	0,582293445	2	0,582293445	0,574395942	0,656296243		
SKTWO	NumProgOne	0,650844554	2	0,650844554	0,727780937	0,700241642		
SKTWO	NumProgtwo	0,680151602	2	0,680151602	0,813743246	0,714221334		0,714221
SKTWO	RangeIPSOne	0,648398873	2	0,648398873	0,765491218	0,690119769		
SKTWO	RangeIPSTWO	0,6985536	2	0,6985536	0,940269467	0,709224465		
SKTWO	SKLONE	0,614874272	2	0,614874272	0,733987894	0,660356244		
SKTWO	KOMSKS	0,69167352	2	0,69167352	0,957180817	0,699198796		
SKTWO	UMUR	0,29441728	2	0,29441728	-0,100796276	0,439085652		
SKTWO	JK	0,477037922	2	0,477037922	0,150263455	0,629026934		
SKTWO	FATHERJOB	0,356331313	2	0,356331313	-0,054899132	0,518358119		
SKTWO	SMA	0,295373621	2	0,295373621	-0,128399866	0,447433051		
SKTWO	JURUSANSMA	0,345775932	2	0,345775932	-0,048064614	0,501194184		

Gambar 4.2 Perhitungan CFS - 2 Terhadap Masa Studi

Berikutnya dengan rumus yang sama, dicari fitur terbaik hingga merit maksimal k+1 lebih kecil dari merit k maksimal. Didapatkan fitur-fitur terbaik setelah perhitungan 1 dan 2 pada Gambar 4.1 dan Gambar 4.2, yaitu nilai IP semester 2, Jenis Kelamin, nilai MK Pemrograman Dasar 1, nilai MK Aljabar Linier, dan nilai MK Logika Algoritma.

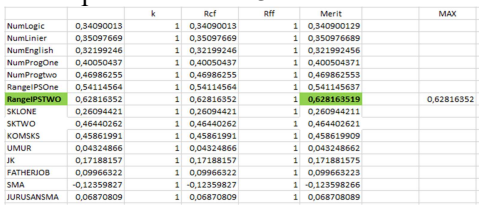
Berdasarkan perhitungan diatas diperoleh informasi bahwa fitur-fitur terbaik untuk kelas masa studi yaitu nilai MK Pemrograman Dasar 2, nilai IP semester 2, jenis kelamin, nilai MK Pemrograman Dasar 1, nilai MK Aljabar Linier, dan nilai MK Logika Algoritma.

4.1.2 Perhitungan *Correlation Feature Selection* Terhadap Predikat

Perhitungan CFS dengan cara menghitung merit maksimal berdasarkan rata-rata dari *feature-class* dan rata-rata dari fitur-fitur *inter-correlation* terhadap predikat berdasarkan korelasi semua fitur terhadap kelas predikat.

Berdasarkan rumus pencarian CFS pada rumus (2.2) dengan mencari merit terbesar pada fitur terbaik yang pertama,

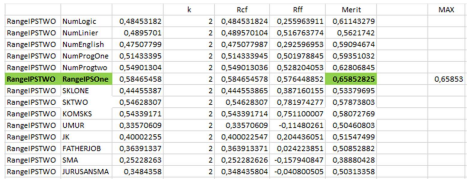
didapatkan fitur pertama yang terbaik untuk predikat kelulusan yaitu nilai Indeks Prestasi Semester 2 yang dapat dilihat pada Gambar 4.3.



		k	Rcf	Rff	Merit		MAX
NumLogic	0,34090013	1	0,34090013	1	0,340900129		
NumLinier	0,35097669	1	0,35097669	1	0,350976689		
NumEnglish	0,32199246	1	0,32199246	1	0,321992456		
NumProgOne	0,40050437	1	0,40050437	1	0,400504371		
NumProgtwo	0,46986255	1	0,46986255	1	0,469862553		
RangeIPSOne	0,54114564	1	0,54114564	1	0,541145637		
RangeIPSTWO	0,62816352	1	0,62816352	1	0,628163519		0,62816352
SKLONE	0,26094421	1	0,26094421	1	0,260944211		
SKTWO	0,46440262	1	0,46440262	1	0,464402621		
KOMSKS	0,45861991	1	0,45861991	1	0,458619909		
UMUR	0,04324866	1	0,04324866	1	0,043248662		
JK	0,17188157	1	0,17188157	1	0,171881575		
FATHERJOB	0,09966322	1	0,09966322	1	0,099663223		
SMA	-0,12359827	1	-0,12359827	1	-0,123598266		
JURUSANSMA	0,06870809	1	0,06870809	1	0,068708089		

Gambar 4.3 Perhitungan CFS – 1 Terhadap Predikat

Selanjutnya pada Gambar 4.4 menerangkan bahwa perhitungan CFS dengan rumus (2.2), dicari fitur terbaik yang kedua setelah Indeks Prestasi Semester 2.



			k	Rcf	Rff	Merit		MAX
RangeIPSTWO	NumLogic	0,48453182	2	0,484531824	0,259963911	0,61143279		
RangeIPSTWO	NumLinier	0,4895701	2	0,489570104	0,516763774	0,5621742		
RangeIPSTWO	NumEnglish	0,47507799	2	0,475077987	0,292596953	0,59094674		
RangeIPSTWO	NumProgOne	0,51433395	2	0,514333945	0,501978845	0,59351032		
RangeIPSTWO	NumProgtwo	0,54901304	2	0,549013036	0,528204053	0,62806845		
RangeIPSTWO	RangeIPSOne	0,58465458	2	0,584654578	0,576448852	0,65852825		0,65853
RangeIPSTWO	SKLONE	0,44455387	2	0,444553865	0,387160155	0,53379695		
RangeIPSTWO	SKTWO	0,54628307	2	0,54628307	0,781974277	0,57873803		
RangeIPSTWO	KOMSKS	0,54339171	2	0,543391714	0,751100007	0,58072769		
RangeIPSTWO	UMUR	0,33570609	2	0,33570609	-0,11480261	0,50460803		
RangeIPSTWO	JK	0,40002255	2	0,400022547	0,204436051	0,51547499		
RangeIPSTWO	FATHERJOB	0,36391337	2	0,363913371	0,024223851	0,50852882		
RangeIPSTWO	SMA	0,2522826	2	0,252282626	-0,157940847	0,38880428		
RangeIPSTWO	JURUSANSMA	0,3484358	2	0,348435804	-0,040800505	0,50313358		

Gambar 4.4 Perhitungan CFS – 2 Terhadap Predikat

Dengan rumus yang sama, dicari fitur terbaik lainnya untuk kelas predikat berupa nilai MK Pemrograman Dasar 2.

Dari hasil perhitungan CFS untuk kelas predikat diperoleh informasi bahwa fitur-fitur terbaik untuk kelas predikat yaitu IP Semester 1, dan nilai MK Pemrograman Dasar 2.

4.2. Penerapan Pengujian 10 *Fold Cross Validation* Pada Sistem

Untuk mengukur apakah algoritma *Naïve* Bayes dapat diterapkan pada studi kasus prediksi masa studi dan predikat kelulusan yang ada di STIKI Malang, dapat dilihat dari besarnya akurasi prediksi. Pencarian akurasi dapat diuji dengan menggunakan perhitungan 10 *Fold Cross Validation*. Pengujian 10 *Fold Cross Validation* akan menghasilkan akurasi yang kurang bias, karena data *testing* dan data *training* yang digunakan akan berbeda-beda serta seluruh data pernah dijadikan *testing* dan *training*.

Berdasarkan hasil pengujian akurasi dengan 10 *Fold Cross Validation* yang ditampilkan ke dalam tabel *confusion* matriks, akurasi rata-rata prediksi untuk masa studi dengan jumlah data sebanyak 320 data mencapai 77,19% dapat dilihat pada Tabel 4.1.

Tabel 4.1 Pengujian Akurasi 10 *Fold Cross Validation* Masa Studi

Fold	Akurasi
1	78,125 %
2	78,125 %
3	87,5 %
4	71,875 %
5	81,25 %
6	78,125 %
7	81,25 %
8	71,875 %
9	75 %
10	68,75 %
Rata-rata	77,19 %

Pengujian akurasi dengan 10 *Fold Cross Validation* untuk predikat kelulusan dengan data sebanyak 170 data, akurasi rata-rata mencapai 87,65% yang dapat dilihat pada Tabel 4.2.

Tabel 4.2 Pengujian Akurasi 10 *Fold Cross Validation* Predikat Kelulusan

Fold	Akurasi
1	76,471 %
2	88,235 %
3	88,235 %
4	94,118 %
5	88,235 %
6	88,235 %
7	94,118 %
8	70,588 %
9	94,118 %
10	94,118 %
Rata-rata	87,65 %

5. KESIMPULAN

Setelah dilakukan analisa, perancangan, implementasi, pelatihan algoritma, dan pengujian pada aplikasi prediksi masa studi dan predikat kelulusan dengan menggunakan metode *Naïve Bayes,* maka dapat disimpulkan bahwa aplikasi yang dibuat mampu untuk

memprediksi masa studi berdasarkan kategorinya yaitu tepat waktu, terlambat atau memuaskan dan memprediksi predikat kelulusan berdasarkan kategorinya yaitu *summa cumlaude*, *cumlaude*, sangat memuaskan dan memuaskan secara otomatis.

Atribut data yang digunakan sebagai fitur dalam prediksi masa studi dan predikat kelulusan telah dilakukan perhitungan pencarian fitur terbaik dengan menggunakan *Correlation Feature Selection*, di mana didapatkan bahwa untuk memprediksi masa studi, fitur terbaiknya berupa SKS lulus semester 2, nilai MK Pemrograman Dasar 2, nilai IP semester 2, jenis kelamin, nilai MK Pemrograman Dasar 1, nilai MK Aljabar Linier, dan nilai MK Logika Algoritma serta untuk memprediksi predikat kelulusan, fitur terbaiknya berupa nilai IP semester 2, nilai IP Semester 1, dan nilai MK Pemrograman Dasar 2.

Dari hasil pengujian 10 *Fold Cross Validation* yang telah dilakukan saat uji pelatihan algoritma, algoritma *Naïve Bayes* dalam prediksi masa studi akurasi rata-rata yang didapatkan sebesar 77,19 % dan akurasi rata-rata prediksi predikat kelulusan mencapai 87,65 %.

Berdasarkan penelitian yang telah dilakukan, sistem yang dibuat masih memiliki beberapa kekurangan. Beberapa saran yang dapat dijadikan acuan untuk pengembangan aplikasi atau sistem lebih lanjut antara lain penambahan fitur-fitur non-akademis yang berpengaruh terhadap masa studi atau predikat kelulusan yang dapat meningkatkan akurasi sistem, *output* masa studi berupa angka (semester), dan meningkatkan akurasi pada penelitian ini menggunakan metode *Supervised Learning* lainnya.

6. REFERENSI

[1] I. A. A. Amra and A. Y. A. Maghari, "Students performance prediction using KNN and Naïve Bayesian," in *ICIT 2017 - 8th International Conference on Information Technology, Proceedings*, 2017.

[2] Harwati, R. I. Virdyanawaty, and A. Mansur, "Drop out Estimation Students based on the Study Period: Comparisonbetween Naïve Bayes and Support Vector Machines Algorithm Methods," in *IOP Conference Series: Materials Science and Engineering*, 2016.

[3] M. Makhtar, H. Nawang, and S. N. W. Shamsuddin, "Analysis on students performance using naĬve Bayes classifier," *J. Theor. Appl. Inf. Technol.*, 2017.

[4] M. A. Nurrohmat, "Aplikasi Pemrediksi Masa Studi dan Predikat Kelulusan Mahasiswa Informatika Universitas Muhammadiyah Surakarta Menggunakan Metode Naive Bayes," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, 2015.

[5] A. Chandra, "PERBEDAAN SUPERVISED AND UNSUPERVISED LEARNING," *https://www.datascience.or.id*, 2017. [Online]. Available: https://www.datascience.or.id/article/Perbedaan-Supervised-and-Unsupervised-Learning-5a8fa6e6. [Accessed: 30-May-2018].

[6] Bustami, "Penerapan Algoritma Naive Bayes," *J. Inform.*, 2014.

[7] Z. H. Kilimci and M. C. Ganiz, "Evaluation of classification models for language processing," in *INISTA 2015 - 2015 International Symposium on Innovations in Intelligent SysTems and Applications, Proceedings*, 2015.

[8] Kemdikbud, "KBBI Daring," *KBBI Daring*, 2016. [Online]. Available: https://kbbi.kemdikbud.go.id/entri/sarjana.

[9] E. Poerbaningtyas, L. Isyriyah, and S. Widodo, *BUKU PEDOMAN AKADEMIK*. Malang: STIKI Malang, 2015.

[10] M. Doshi, "Correlation Based Feature Selection (Cfs) Technique To Predict Student Perfromance," *Int. J. Comput. Networks*

Commun., 2014.
[11] A. Wibowo, "10 Fold Cross Validation," *https://mti.binus.ac.id*, 2017. [Online]. Available: https://mti.binus.ac.id/2017/11/24/10-fold-cross-validation/.
[12] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, 2009.
[13] F. E. Purwiantono and A. Tjahyanto, "Classification model based on url and content feature approach for detection phishing website in Indonesia," *J. Theor. Appl. Inf. Technol.*, 2017.