Implementasi Algoritma *Naïve Bayes* Menggunakan *Feature Forward Selection* dan SMOTE Untuk Memprediksi Ketepatan Masa Studi Mahasiswa Sarjana

Dede Kurniadi
Jurusan Ilmu Komputer
Institut Teknologi Garut
Garut, Indonesia
dede.kurniadi@itg.ac.id

Fitri Nuraeni
Jurusan Ilmu Komputer
Institut Teknologi Garut
Garut, Indonesia
fitri.nuraeni@itg.ac.id

Sri Mulyani Lestari
Program Studi Teknik Informatika
Institut Teknologi Garut
Garut, Indonesia
1806131@itg.ac.id

Abstract— The punctuality of students in completing their studies is an important aspect of the study program. Because there are still students who have not been able to complete their studies on time. The purpose of this study is to determine the factors that influence students in completing their studies by extracting student academic data to obtain a classification model that can be used to predict the accuracy of the study period. The classification method for predicting the accuracy of the student's study period uses the Naive Bayes algorithm using the Feature Forward Selection and SMOTE. The method for data processing in this study uses CRISP-DM. The results of this study are in the form of a classification model to predict the accuracy of the study period of students who obtain a fairly high accuracy value of 87.13%, a recall value of 83.82%, and a precision value of 89.76%, and an AUC value of 0.92. included in the category of Excellent Classification. The use of SMOTE has succeeded in handling Imbalanced Class on the data, and the application of Feature Forward Selection resulted in 5 factors that most influence the accuracy of the student's study period, namely the attributes of Gender, School Category, Year of Entry, Study Program and Grade Point Average for the third semester. The prediction model generated using the Naïve Bayes algorithm, Feature Forward Selection, and SMOTE is expected to help study programs to find out earlier the possibility of students completing their studies on time or not on time.

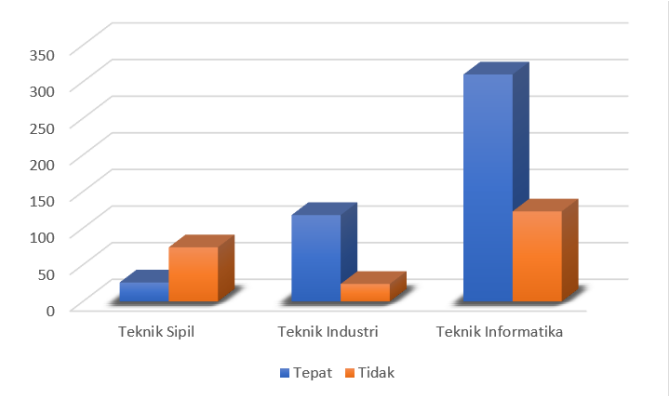
Keywords— Prediction, Algorithm, Naïve Bayes, Feature Forward Selection, SMOTE

I. PENDAHULUAN

Kualitas pengelolaan pendidikan suatu perguruan tinggi dan program studi, dapat dilihat dari peringkat akreditasinya. Pada proses akreditasi, salah satu aspek penting yang dinilai adalah mahasiswa, mulai dari bagaimana proses mahasiswa masuk ke perguruan tinggi, memilih program studi sampai pada ketepatan masa studi mahasiswa dari program studi tersebut. Masa studi mahasiswa telah diatur pada Permendikbud No.49 Tahun 2014 mengenai Standar Nasional Pendidikan Tinggi (SN-PT) yaitu beban belajar minimal mahasiswa pada jenjang pendidikan S1 adalah 144 SKS dan untuk menuntaskan seluruh beban SKS mahasiswa S1 diberi batas waktu 4-5 tahun (8-10 semester) [1]. Namun pada berlangsungnya proses akademik pada suatu program studi, tidak semua mahasiswa dapat menyelesaikan studi sesuai jangka waktu yang telah ditentukan. Sering kali ditemukan sejumlah mahasiswa yang menempuh studi S1 melebihi batas waktu maksimal dan beberapa diantaranya terancam *drop out*.

Gambaran permasalahan tersebut, masih ditemukan pada Institut Teknologi Garut (ITG) yang merupakan perguruan tinggi swasta di Kabupaten Garut yang sudah berdiri sejak Tahun 1990. ITG memiliki 5 program studi 2 diantaranya merupakan prodi baru yaitu Sistem Informasi dan Arsitek sedangkan 3 lainnya yaitu Teknik Informatika, Teknik Sipil dan Teknik Industri ketiganya sudah mendapat nilai akreditasi B (Baik Sekali) [2]. Pada masing-masing program studi tersebut, masih terdapat mahasiswa yang lulus tidak tepat waktu setiap tahunnya. Jika hal ini dibiarkan, dikhawatirkan akan berdampak pada

nilai akreditasi program studi yang telah diperoleh. Sehingga memerlukan suatu upaya agar mengurangi jumlah mahasiswa yang lulus tidak tepat waktu.



Gambar 1. Grafik Mahasiswa Lulus Tepat dan Tidak Tepat Waktu

Gambar 1 menunjukan bahwa pada tiap program studi di Institut Teknologi Garut, masih terdapat banyak mahasiswa yang terlambat dalam menyelesaikan studi. Oleh karena itu untuk mengurangi jumlah mahasiswa yang lulus tidak tepat waktu, maka pihak program studi perlu adanya pencegahan dan arahan dengan melakukan proses prediksi ketepatan masa studi mahasiswa lebih awal. Prediksi merupakan suatu kegiatan yang dilakukan secara sistematis dan bertujuan untuk memperkirakan suatu hal yang mungkin terjadi di masa depan. Proses prediksi dapat diketahui berdasarkan informasi di masa lalu, sehingga ditemukan perbedaan antara sesuatu yang telah terjadi dan hasil yang diharapkan dapat diminimalisir [3]. Hubungan antara masa studi mahasiswa dengan kelulusan dapat diprediksi menggunakan data riwayat akademik mahasiswa di masa lalu [1]. Dengan adanya prediksi ketepatan masa studi mahasiswa lebih awal yaitu ketika semester 6, pihak program studi dapat melakukan perencanaan, pengawalan studi dan bimbingan yang lebih intensif terhadap mahasiswa yang terindikasi lulus tidak tepat waktu dan terancam *drop out.*

Pada proses memprediksi ketepatan masa studi mahasiswa dengan menerapkan algoritma pohon keputusan C4.5 pada penelitian [4] memperoleh nilai akurasi sebesar 73,99%. Pada tahun berikutnya penelitian [5] dalam mengimplmentasikan data mining untuk klasifikasi masa studi mahasiswa menggunakan algoritma *K-Nearest Neighbor* diperoleh hasil performa akurasi 75.15%, *precision* 93.15% dan *recall* 77.65%. Kemudian penelitian [6] yang membahas sistem rekomendasi akademik menggunakan *Backpropagation Neural Network* untuk memprediksi kemampuan mahasiswa dalam menyelesaikan studi menghasilkan nilai akurasi sebesar 80,95%, presisi 83,20%, dan *recall* 83,53%. Penelitian selanjutnya yang dilaporkan oleh [7] penggunaan data induk mahasiswa sebagai prediktor ketepatan waktu lulus dan algoritma *CART* untuk pengklasifikasiannya diperoleh nilai *accuracy:* 63.19%, *specificity:* 54.48% dan *sensitivity:* 69.94%. Sedangkan penelitian [3] menggunakan data akademik dan non akademik dalam memprediksi ketepatan kelulusan mahasiswa menggunakan metode *K-Means* memperoleh nilai dengan tingkat akurasi mencapai 90% benar. Selanjutnya penelitian untuk mencari model prediksi ketepatan masa studi mahasiswa [8] dengan menerapkan fitur *Forward Selection* pada algoritma *Naïve Bayes* menghasilkan nilai akurasi yang cukup tinggi yaitu 92,94%.

Dari berbagai metode dan algoritma berdasarkan penelitian sebelumnya untuk memprediksi ketepatan masa studi mahasiswa, perlu dibangun model klasifikasi. Pemilihan algoritma *Naïve Bayes* dianggap menjadi salah satu proses klasifikasi yang menghasilkan nilai akurasi terbaik [8]. Selain itu, untuk mengetahui atribut/faktor yang

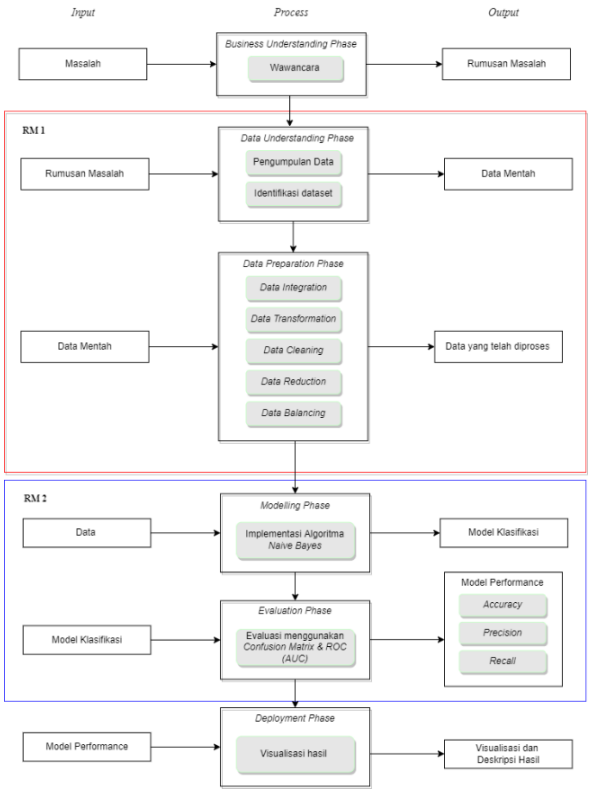
paling mempengaruhi terhadap ketepatan masa studi mahasiswa, diperlukan fitur seleksi atribut dengan menggunakan *Forward Selection*. Fitur ini mampu meningkatkan akurasi dengan membuang beberapa fitur yang kurang relevan terhadap proses klasifikasi [9]. Data yang digunakan yaitu data mahasiswa lulusan tahun 2015-2019 dari 3 program studi yaitu Teknik Informatika, Teknik Sipil dan Teknik Industri ITG. Berdasarkan data yang diperoleh terjadi *imbalanced class* diantara mahasiswa yang menyelesaikan studi tepat waktu dan tidak tepat waktu, maka diperlukan teknik SMOTE. Penggunaan teknik SMOTE ini mampu menyeimbangkan *class imbalance* pada data [10]. Berdasarkan latar belakang dan kajian penelitian sebelumnya, maka penelitian ini fokus pada Implementasi Algoritma *Naïve Bayes* Untuk Memprediksi Ketepatan Masa Studi Mahasiswa Sarjana. Hasil yang diharapkan dari penelitian ini, dapat menjadi acuan bagi program studi untuk memberikan arahan kepada mahasiswa agar meningkatkan jumlah mahasiswa yang lulus tepat waktu.

II. METODE PENELITIAN

Metode pada proses prediksi ketepatan masa studi mahasiswa menggunakan metode tahapan CRISP-DM (*Cross-Industry Standard Practice for Data Mining*) metode pendekatan ini memperkenalkan standar proses data mining sebagai strategi untuk memecahkan masalah umum untuk mengatasi proses bisnis maupun penelitian[4]. .

A. Kerangka Penelitian

Penerapan metode tahapan CRISP-DM diterapkan pada kerangka penelitian sebagai diagram yang menjelaskan secara garis besar dari alur berjalannya sebuah penelitian yang disajikan pada Gambar 2 berikut:



Gambar 2. Kerangka Penelitian

Keterangan :
RM 1 : "Bagaimana menentukan faktor/atribut yang mempengaruhi terhadap ketepatan masa studi mahasiswa?" Yang akan terjawab pada tahapan *Data Understanding Phase & Data Preparation Phase.*
RM 2 : "Bagaimana menerapkan algoritma *Naïve Bayes*, *Feature Forward Selection* dan SMOTE untuk mencari model klasifikasi untuk memprediksi ketepatan masa studi mahasiswa dengan tingkat akurasi yang baik?" Yang akan terjawab pada tahapan *Modelling Phase & Evaluation Phase.*

Penjelasan mengenai kerangka penelitian menggunakan metode tahapan CRISP-DM terdapat 3 tahapan yaitu *input*, *process* dan *output*. *Input* merupakan sumber yang diperlukan untuk menghasilkan *output*. Process merupakan kegiatan mengolah *input* agar mendapatkan *output*. Sedangkan *output* merupakan hasil yang diperoleh dari *input* yang telah di *process*. Pada tahapan *process* terdiri dari 6 tahapan yaitu *Business Understanding Phase*, *Data Understanding Phase*, *Data Preparation Phase*, *Modelling Phase*, *Evaluation Phase dan Deployment Phase.*

B. Sumber Data

Sumber data yang digunakan untuk memenuhi kebutuhan analisis data dan pemodelan pada penelitian ini, diperoleh dari database sistem informasi akademik Institut Teknologi Garut. Data yang digunakan merupakan data mahasiswa lulusan tahun 2015–2019 dari 3 program studi yaitu Teknik Informatika, Teknik Sipil dan Teknik Industri yang disajikan pada Gambar 3 berikut:

NIM	JK	Kategori Sekolah	Usia	Program Studi	Tahun Masuk	IPs 1	IPs 2	IPs 3	IPs 4	IPs 5	SKS 1	SKS 2	SKS 3	SKS 4	SKS 5	Status Kelulusan	Row No.
1211002	L	SMA	18	Teknik Sipil (S1)	2012	3.090	2.940	2.650	2.950	3	18	21	21	19	20	Tidak Tepat Waktu	29
1211003	P	SMA	18	Teknik Sipil (S1)	2012	3.360	2.940	2.400	2.860	2.900	18	21	21	19	20	Tidak Tepat Waktu	30
1211004	L	SMK	19	Teknik Sipil (S1)	2012	3.050	3.110	2.700	3	2.710	18	21	21	22	20	Tidak Tepat Waktu	31
1211005	L	SMK	19	Teknik Sipil (S1)	2012	3.140	3.110	2.700	3.090	2.810	18	21	21	19	20	Tidak Tepat Waktu	32
1211007	L	SMK	20	Teknik Sipil (S1)	2012	3.050	3	2.700	3.360	2.810	18	21	21	19	20	Tidak Tepat Waktu	33
1211008	L	SMK	37	Teknik Sipil (S1)	2012	3.320	3.210	3.160	2.950	2.740	18	18	19	15	20	Tidak Tepat Waktu	34
1211013	L	SMA	22	Teknik Sipil (S1)	2012	3.230	3	2.650	2.730	2.900	18	21	21	19	20	Tidak Tepat Waktu	35
1211015	P	SMK	28	Teknik Sipil (S1)	2012	3.320	3.280	3.450	3.230	3.570	18	21	21	19	20	Tidak Tepat Waktu	36
1211016	L	SMA	17	Teknik Sipil (S1)	2012	3.880	3.280	3.200	3.320	3.430	18	21	21	21	20	Tidak Tepat Waktu	37
1211022	L	SMK	21	Teknik Sipil (S1)	2012	3.140	2.890	2.800	3.090	3	18	21	21	19	20	Tidak Tepat Waktu	38
1211023	L	SMA	21	Teknik Sipil (S1)	2012	2.770	3.110	2.550	2.860	2.570	18	21	21	19	20	Tidak Tepat Waktu	39
1211024	L	SMK	18	Teknik Sipil (S1)	2012	2.680	3.110	2.800	3	2.860	18	21	21	19	20	Tidak Tepat Waktu	40
1211026	L	SMK	20	Teknik Sipil (S1)	2012	2.770	3	2.700	2.730	2.900	18	21	21	20	23	Tidak Tepat Waktu	41
1211028	L	SMK	18	Teknik Sipil (S1)	2012	2.840	2.790	2.740	2.630	2.680	18	18	19	15	20	Tidak Tepat Waktu	42
1211031	P	SMA	18	Teknik Sipil (S1)	2012	3.790	3.670	3.550	3.770	3.710	18	21	21	21	20	Tidak Tepat Waktu	43
1211035	P	SMK	17	Teknik Sipil (S1)	2012	3.630	3.390	3.550	3.550	3.430	18	21	21	21	20	Tidak Tepat Waktu	44

ExampleSet (681 examples, 0 special attributes, 17 regular attributes)

Gambar 3. Data yang terkumpul

Berdasarkan Gambar 3 merupakan data yang terkumpul dan masih berupa data mentah yang belum diolah. Terdiri dari 17 atribut dan 681 mahasiswa Institut Teknologi Garut. Berikut rincian data yang diperoleh, beserta atribut yang digunakan meliputi:

1) Data Mahasiswa (NIM, Jenis Kelamin, Kategori Sekolah, Usia)
2) Data Lulusan (Program Studi, Tahun Masuk, IPs 1, IPs 2, IPs 3, IPs 4, IPs 5, SKS 1, SKS 2, SKS 3, SKS 4, SKS 5 dan Status Kelulusan.
3) Menggunakan 17 atribut dan jumlah dataset yang terkumpul sebanyak 681 mahasiswa yang terdiri dari 435 mahasiswa Teknik Informatika, 143 mahasiswa Teknik Industri dan 103 mahasiswa Teknik Sipil.

III. HASIL DAN PEMBAHASAN

Hasil dan pembahasan dari penelitian ini merupakan proses dari implementasi algoritma *Naive Bayes* menggunakan *Feature Forward Selection* dan SMOTE. Proses

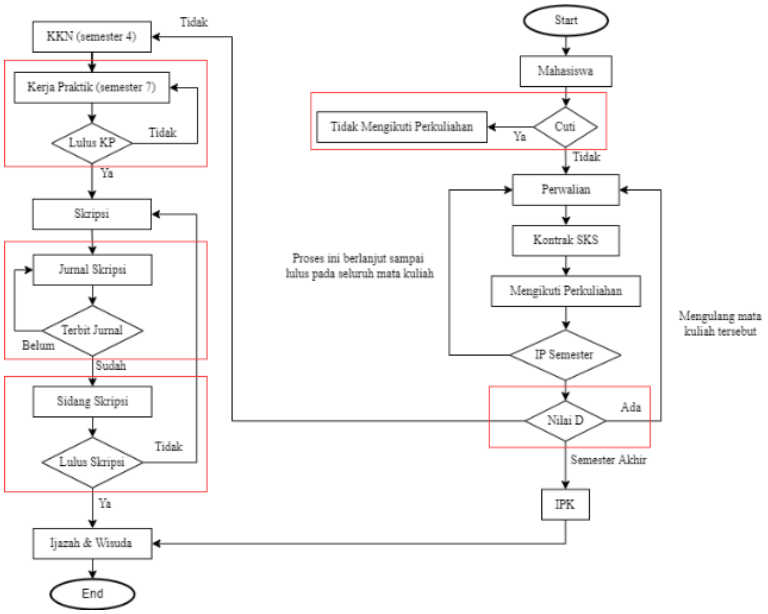
pemodelan ini dibantu dengan menggunakan *tools Rapidminer*. Untuk hasil penelitian dan pembahasan hasil yang diperoleh, disajikan pada sub bahasan berikut:

A. Hasil Penelitian

Penelitian ini menggunakan metode pendekatan CRISP-DM. Adapun hasil dari penelitian ini akan dijelaskan berdasarkan tahapan CRISP-DM. Untuk lebih jelasnya dapat dilihat pada sub poin berikut:

1) ***Business Understanding Phase***

Berdasarkan objek penelitian yang telah ditentukan, maka tahapan pertama dari penelitian ini yaitu mengidentifikasi masalah yang terjadi di ITG dengan melakukan wawancara dengan pihak kampus, yaitu ketua prodi Teknik Sipil, Teknik Informatika dan Teknik Industri mengenai permasalahan mahasiswa dalam menyelesaikan studi. Berikut merupakan hasil wawancara yang diperoleh bagaimana alur mahasiswa dalam menyelesaikan studi terdapat pada Gambar 4.



Gambar 4. Flowchart Studi Mahasiswa

Berdasarkan permasalahan tersebut ketepatan masa studi mahasiswa menjadi terhambat. Dikarenakan proses pengulangan/pengambilan semester untuk perbaikan yang mengharuskan mahasiswa menambah masa studi. Maka dibutuhkan solusi untuk menangani permasalahan tersebut agar dapat menyelesaikan studi tepat waktu. Proses yang dilakukan untuk mencegah/menangani permasalahan tersebut yaitu dengan melakukan perencanaan, pengawalan studi, dan bimbingan yang lebih intensif terhadap mahasiswa yang terindikasi lulus tidak tepat waktu ataupun yang terancam drop out jika lebih dari 14 semester.

2) ***Data Understanding Phase***

Setelah merumuskan masalah, selanjutnya adalah tahapan pemahaman data. Langkah-langkah yang harus dilakukan adalah sebagai berikut:

a) Tahap pengumpulan data yaitu proses mencari data dari pihak kampus ITG. Data yang diperoleh merupakan data mahasiswa lulusan tahun 2015-2019 dari 3 program studi yaitu Teknik Informatika, Teknik Sipil dan Teknik Industri yang disajikan pada Tabel 1.

Tabel 1. Data Mentah

No	NIMHS	NMMHS	…	Yudisium
1	1003014	Ridwan Trihasa	…	Tidak Berpredikat
2	1003018	Ardi Rustiardi	…	Tidak Berpredikat
3	1103003	Anggi Sutardi	…	Tidak Berpredikat
…	…	…	…	…
681	1706115	Agung A. A.	…	Memuaskan

Data pada Tabel 1 merupakan file data yang masih terpisah, karena diperoleh dari program studi yang berbeda yaitu Teknik Informatika, Teknik Industri dan Teknik Sipil sebanyak 681 *record* dan terdiri dari 24 atribut yaitu NIM, Nama Mahasiswa, Jenis Kelamin, Asal Sekolah, Tanggal Lahir, Tempat Lahir, Program Studi, Tahun Masuk, Skorusm, IPs 1 sampai IPs 5, SKS 1 sampai SKS 5, IPK, Tahun Lulus, Status Kelulusan, Lama Studi dan Yudisium.

b) Tahap mengidentifikasi jenis dataset dan menentukan atribut yang akan digunakan. Atribut yang digunakan pada penelitian ini meliputi NIM, jenis kelamin, asal sekolah, usia, program studi, tahun masuk, IPs 1, IPs 2, IPs 3, IPs 4, IPs 5, SKS 1, SKS 2, SKS 3, SKS 4, SKS 5 dan status kelulusan yang disajikan pada Tabel 2.

Tabel 2. Atribut dalam dataset

Atribut	Tipe Data	Indikator
NIM	Bilangan bulat	-
Jenis Kelamin	*Binomial*	Laki-laki & Perempuan
Kategori Sekolah	*Polynomial*	SMA, SMK, MA, SKB, Paket C & SPMA
Usia	Bilangan bulat	-
Program Studi	*Polynomial*	Teknik Sipil, Teknik Industri dan Teknik Informatika
Tahun Masuk	*Polynomial*	2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017
Indeks Prestasi semester 1	Bilangan desimal	-
Indeks Prestasi semester 2	Bilangan desimal	-
Indeks Prestasi semester 3	Bilangan desimal	-
Indeks Prestasi semester 4	Bilangan desimal	-
Indeks Prestasi semester 5	Bilangan desimal	-
Satuan Kredit Semester 1	Bilangan bulat	-
Satuan Kredit Semester 2	Bilangan bulat	-
Satuan Kredit Semester 3	Bilangan bulat	-
Satuan Kredit Semester 4	Bilangan bulat	-
Satuan Kredit Semester 5	Bilangan bulat	-
Status Kelulusan	*Binomial*	Tepat Waktu & Tidak Tepat Waktu.

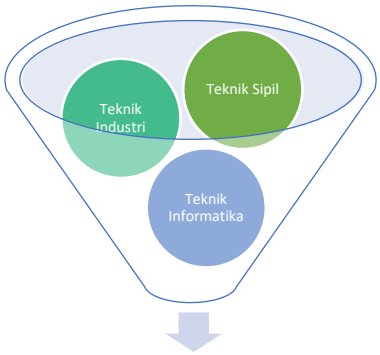
Berdasarkan Tabel 2 terdapat 17 atribut yang terdiri dari 7 tipe data bilangan bulat pada atribut NIM, Usia dan jumlah SKS (Satuan Kredit Semester). Terdapat 2 tipe data binomial pada atribut Jenis Kelamin dan Status Kelulusan. Terdapat 3 tipe data polynomial pada atribut Kategori Sekolah, Program Studi dan Tahun Masuk. Terdapat 5

tipe data bilangan desimal pada atribut IPs (Indeks Prestasi Semester). Setelah melakukan langkah-langkah pada tahapan *Data Understanding Phase*, maka diperoleh data mentah.

3) ***Data Preparation Phase***

Data mentah yang diperoleh pada tahapan sebelumnya perlu dilakukan *Data Preparation Phase.* Proses yang harus dilakukan dalam pengolahan data pada penelitian ini adalah sebagai berikut:

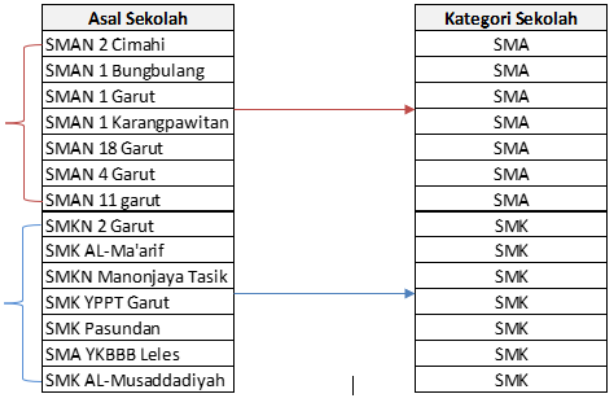
a) Proses *Data Integration* yaitu proses penggabungan data, dimana data yang diperoleh dari 3 program studi akan dikombinasikan atau dibuat menjadi satu yang diilustrasikan pada Gambar 5 berikut ini:



Gambar 5. Ilustrasi Proses *Data Integration*

Berdasarkan Gambar 5 ilustrasi proses penggabungan data yang terdiri dari 435 *record* mahasiswa Teknik Informatika, 143 *record* mahasiswa Teknik Industri dan 103 *record* mahasiswa Teknik Sipil dengan total jumlah data keseluruhan menjadi 681 *record.*

b) Proses *Data Transformation* digunakan untuk mengubah data ke dalam bentuk yang dibutuhkan sesuai proses data mining, dengan mengganti variabel menjadi kode. Berikut merupakan proses *Data Transformation seperti* pada Gambar 6.



Gambar 6. Proses Transformasi Data

Berdasarkan Gambar 6 proses transformasi yang dilakukan tidak hanya pada atribut KS (Kategori Sekolah). Proses ini dilakukan juga pada atribut JK (Jenis Kelamin) dengan mengubah keterangan Laki-laki dan Perempuan menjadi L dan P. Sedangkan untuk atribut Prodi (Program Studi) menggunakan kode, yaitu

Teknik Informatika menjadi IF, Teknik Industri menjadi TI dan Teknik Sipil menjadi TS.

c) Proses *Data Cleaning* yaitu proses pembersihan data mahasiswa lulusan tahun 2015-2019 mulai dari yang tidak valid, kosong, dan duplikat menjadi data yang siap untuk diolah. Data yang dihasilkan setelah melakukan proses *data cleaning* terdapat pada Tabel 3.

Tabel 3. Data Mahasiswa Lulusan Tahun 2015-2019

No	NIM	JK	KS	…	Status Kelulusan
1	0811010	L	SMA	…	Tidak Tepat Waktu
2	0811015	L	SMK	…	Tidak Tepat Waktu
3	0811038	P	SMK	…	Tidak Tepat Waktu
…	…	…	…	…	…
675	1706115	L	SMA	…	Tepat Waktu

Data pada Tabel 3 merupakan data yang telah melewati proses *Data Integration & Data Transformation*. Setelah dilakukan proses *data cleaning*, jumlah data yang awalnya 681 *record* sekarang menjadi 675 *record.*

d) Proses *Data Reduction* melalui proses seleksi untuk mengetahui atribut/faktor yang paling mempengaruhi terhadap ketepatan masa studi mahasiswa menggunakan *tools Rapidminer*. *Feature Forward Selection* merupakan salah satu seleksi fitur yang dilakukan sebelum proses klasifikasi. Proses ini terbukti efektif untuk menentukan atribut yang relevan dalam suatu data. Hal ini mempengaruhi terhadap hasil klasifikasi dan mengurangi dimensi data untuk meningkatkan akurasi pada proses klasifikasi [9]. Seleksi fitur ini bertujuan untuk mengetahui atribut yang mempengaruhi terhadap tingkat akurasi. Berikut langkah-langkah dalam menerapkan *Forward Selection* meliputi:

1. Membangun model dengan meregresikan variabel respons Y pada setiap prediktor. Kemudian pilih model dengan nilai R^2 tertinggi.
2. Regresi respons Y menggunakan prediktor X_a dan prediktor lain selain prediktor X_a. Kemudian pilih model dengan nilai R^2 tertinggi. Nilai $F_{sekuensial}$ X_b juga dapat diperoleh dengan mengkuadratkan statistik uji-T prediktor X_b.
3. Proses ini dilakukan berulang sampai $F_{sequential} > F_{in}$. Karena nilai $F_{in} = F(1, v,\alpha in)$, dan model terbaik yang dipilih adalah model tanpa prediktor $F_{sequential} < F_{in}$.

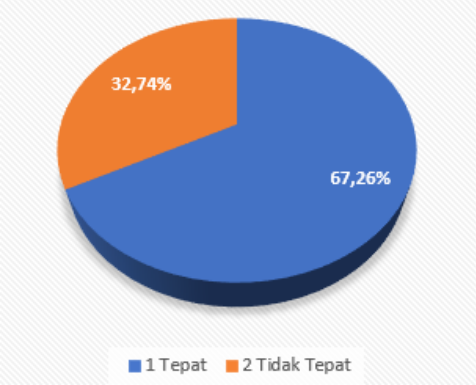
Berikut merupakan atribut hasil dari penerapan Feature Forward Selection disajikan pada Tabel 4.

Tabel 4. Atribut Hasil Forward Selection

No	*Attribute*	*Weight*
1	Jenis Kelamin	1
2	Kategori Sekolah	1
3	Usia	0
4	Program Studi	1
5	Tahun Masuk	1
6	Indeks Prestasi semester 1	0
7	Indeks Prestasi semester 2	0
8	Indeks Prestasi semester 3	1
9	Indeks Prestasi semester 4	0
10	Indeks Prestasi semester 5	0
11	Satuan Kredit Semester 1	0
12	Satuan Kredit Semester 2	0
13	Satuan Kredit Semester 3	0
14	Satuan Kredit Semester 4	0
15	Satuan Kredit Semester 5	0

Penggunaan *tools Rapidminer* membantu dalam proses perhitungan untuk mencari nilai pada setiap atribut. Hasil yang diperoleh dari penggunaan operator *Forward Selection* pada *Rapidminer*, menghasilkan nilai pada setiap atribut yang dinyatakan dalam bentuk biner 0 dan 1. Berdasarkan Tabel 4 terdapat 5 atribut yang memiliki bobot nilai 1. Maka dapat disimpulkan atribut yang paling mempengaruhi terhadap ketepatan masa studi mahasiswa yaitu Jenis Kelamin, Kategori Sekolah, Program Studi, Tahun Masuk dan Indeks Prestasi semester 3.

e) Proses *Data Balancing* ini dilakukan karena terdapat *class imbalanced* pada data mahasiswa yang tepat dan tidak tepat waktu. *Synthetic Minority Over Sampling Technique* (SMOTE) adalah suatu teknik untuk menyamakan distribusi data sampel pada kelas minoritas dengan cara memilih data sampel sampai jumlah sampel data sama dengan jumlah sampel pada kelas mayoritas. SMOTE memungkinkan untuk mengatasi *overfitting. Overfitting* dapat terjadi karena duplikasi data di kelas minoritas, memungkinkan data pelatihan yang sama digunakan [10]. Untuk proses menyeimbangkan data pada atribut Status Kelulusan. Atribut ini merupakan label klasifikasi pada penelitian yang disajikan pada diagram berikut ini:



Gambar 7. *Diagram Class Imbalanced*

Proses sampling ulang dilakukan pada jumlah data mahasiswa tepat waktu sebanyak 32,74% menyesuaikan dengan jumlah data mahasiswa tidak tepat waktu. Proses pengambilan data sampel dilakukan secara random. Sehingga dibutuhkan *tools* yang membantu dalam proses menyeimbangkan data. *Tools* yang digunakan yaitu *Rapidminer* dengan menerapkan operator SMOTE Upsampling pada saat proses pemodelan.

4) ***Modelling Phase***

Setelah memiliki data yang siap untuk diproses, tahapan selanjutnya yaitu melakukan pemilihan teknik pemodelan yang sesuai untuk mengoptimalkan hasil. Teknik Pemodelan data mining yang dipilih pada penelitian ini, dalam menemukan model klasifikasi yaitu dengan menerapkan algoritma *Naïve Bayes* menggunakan *Feature Forward Selection* dan SMOTE. Data mining memiliki banyak fungsi yang dapat digunakan. Fungsi data mining dapat diterapkan pada kasus tertentu untuk menyelesaikan permasalahan yang ada [11].

Algoritma *Naïve Bayes* adalah algoritma untuk mengklasifikasikan data, dengan cara yang sangat sederhana dalam mengasumsikan klasifikasi atribut. Algoritma ini sering digunakan dalam menyelesaikan masalah pada proses *machine learning*. Algoritma ini dikenal memiliki tingkat akurasi yang tinggi dengan perhitungan sederhana [12]. Berikut Teorema *Bayes* disajikan pada persamaan:

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

Keterangan:
X = Data dengan *class* yang belum diketahui
H = Hipotesis X merupakan suatu *class* spesifik
P(*H*|*X*) = Probabilitas hipotesis H berdasarkan kondisi x
P(*H*) = Probabilitas hipotesis H
P(*X*|*H*) = Probabilitas X berdasarkan kondisi tersebut
P(*X*) = Probabilitas dari X

Implementasi algoritma *Naïve Bayes* menggunakan *Feature Forward Selection* dan SMOTE untuk memprediksi ketepatan masa studi mahasiswa, diharapkan mampu memberikan hasil yang optimal dan memperoleh nilai akurasi yang baik. Untuk langkah pertama pada tahapan *Modelling Phase* yaitu mencari nilai probabilitas kelas yaitu atribut Status Kelulusan, Adapun proses untuk menghitung probabilitas menggunakan persamaan (1) sebagai berikut:

$$p(H|X) = \frac{p(X|H)\ p(H)}{p(X)}$$

$$p(\text{Tepat Waktu}) = \frac{454}{675} = 0{,}672592593$$

$$p(\text{Tidak Tepat Waktu}) = \frac{221}{675} = 0{,}327407407$$

Selanjutnya mencari nilai probabilitas pada setiap atribut. Berdasarkan perolehan hasil *Feature Forward Selection*, atribut yang paling mempengaruhi terhadap ketepatan masa studi mahasiswa yaitu Jenis Kelamin, Kategori Sekolah, Tahun Masuk, Program Studi dan IPS 3. Adapun proses untuk menghitung probabilitas atribut menggunakan persamaan (1) disajikan pada Tabel 5 sampai Tabel 9 sebagai berikut:

Tabel 5. Nilai Probabilitas Atribut Jenis Kelamin

Jenis Kelamin	***Class***	
	Tidak Tepat Waktu	Tepat Waktu
L	0,848	0,735
P	0,152	0,265

Berdasarkan Tabel 5 merupakan perolehan nilai dari hasil perhitungan menggunakan persamaan (1), dalam mencari nilai probabilitas atribut jenis kelamin. Nilai yang diperoleh yaitu 0,848 mahasiswa laki-laki dan 0,152 mahasiswa perempuan dari jumlah total mahasiswa tidak tepat waktu sebanyak 221 orang. Sedangkan nilai mahasiswa yang tidak tepat waktu yaitu 0,735 mahasiswa laki-laki dan 0,265 mahasiswa perempuan dari jumlah total mahasiswa tepat waktu sebanyak 454 orang.

Tabel 6. Nilai Probabilitas Atribut Kategori Sekolah

Kategori Sekolah	***Class***	
	Tidak Tepat Waktu	Tepat Waktu
SMA	0,167	0,006
SMK	0,136	0,206
MA	0,015	0,051
Paket C	0,015	0

SPMA	0	0,002
SKB	0	0,002
	0,666	0,647

Berdasarkan Tabel 6 merupakan perolehan nilai dari hasil perhitungan menggunakan persamaan (1), dalam mencari nilai probabilitas atribut kategori sekolah. Nilai yang diperoleh yaitu 0,167 mahasiswa yang berasal dari SMA. 0,136 mahasiswa yang berasal dari SMK dan 0,015 mahasiswa yang berasal dari MA dan Paket C. Untuk mahasiswa yang berasal dari SPMA dan SKB bernilai 0 atau dinyatakan tidak terdapat mahasiswa yang tidak tepat waktu dan terdapat data *missing* pada atribut Kategori Sekolah dengan perolehan nilai 0,666 mahasiswa yang tidak tepat waktu dari jumlah total 221 orang.

Selain itu, mahasiswa yang tepat waktu memperoleh nilai 0,096 mahasiswa yang berasal dari SMA. 0,206 mahasiswa yang berasal dari SMK. 0,051 mahasiswa yang berasal dari MA. 0,002 mahasiswa yang berasal dari SPMA dan SKB. Sedangkan mahasiswa dari Paket C bernilai 0 atau dinyatakan tidak terdapat mahasiswa yang tepat waktu. Untuk data yang *missing* pada mahasiswa tepat waktu memperoleh nilai 0,647 dari jumlah total 454 orang.

Tabel 7. Nilai Probabilitas Atribut Program Studi

Program Studi	***Class***	
	Tidak Tepat Waktu	Tepat Waktu
TS	0,288	0,044
TI	0,121	0,228
IF	0,591	0,728

Berdasarkan Tabel 7 merupakan perolehan nilai dari hasil perhitungan menggunakan persamaan (1) yang, dalam mencari nilai probabilitas atribut program studi. Nilai yang diperoleh yaitu 0,288 mahasiswa Program Studi Teknik Sipil, 0,121 mahasiswa Program Studi Teknik Industri dan 0,591 mahasiswa Program Studi Teknik Informatika dari jumlah total mahasiswa yang tidak tepat waktu sebanyak 221 orang.

Sedangkan mahasiswa yang tidak tepat waktu memperoleh nilai 0,44 mahasiswa Program Studi Teknik Sipil, 0,228 mahasiswa Program Studi Teknik Industri dan 0,728 mahasiswa Program Studi Teknik Informatika dari jumlah total mahasiswa sebanyak 454 orang.

Tabel 8. Nilai Probabilitas Atribut Tahun Masuk

Tahun Masuk	***Class***	
	Tidak Tepat Waktu	Tepat Waktu
2008	0,015	0
2009	0,030	0
2010	0,106	0
2011	0,212	0,103
2012	0,197	0,132
2013	0,091	0,235
2014	0,333	0,176
2015	0,015	0,345
2016	0	0,007
2017	0	0,004

Berdasarkan Tabel 8 merupakan perolehan nilai dari hasil perhitungan menggunakan persamaan (1), dalam mencari nilai probabilitas atribut tahun masuk. Nilai yang diperoleh yaitu 0,015 pada mahasiswa Angkatan 2008 dan 2015. 0,030 pada mahasiswa Angkatan 2009. 0,106 pada mahasiswa Angkatan 2010. 0,212 pada mahasiswa Angkatan 2011. 0,197 pada mahasiswa Angkatan 2012. 0,091 pada mahasiswa Angkatan 2013. 0,333 pada mahasiswa Angkatan 2014 dan untuk mahasiswa Angkatan 2016 dan 2017 bernilai 0 atau dinyatakan tidak terdapat mahasiswa yang tidak tepat waktu dari jumlah total 221 orang. Lalu mahasiswa yang tepat waktu memperoleh nilai 0,103 pada mahasiswa

Angkatan 2011. 0,132 pada mahasiswa Angkatan 2012. 0,235 pada mahasiswa Angkatan 2013. 0,176 pada mahasiswa Angkatan 2014. 0,345 pada mahasiswa Angkatan 2015. 0,007 pada mahasiswa Angkatan 2016. 0,004 pada mahasiswa Angkatan 2017 dan untuk mahasiswa Angkatan 2008, 2009 dan 2010 bernilai 0 atau dinyatakan tidak terdapat mahasiswa yang tepat waktu dari jumlah total 454 orang.

Tabel 9. Nilai Probabilitas Atribut IPs 3

IPs 3	***Class***	
	Tidak Tepat Waktu	Tepat Waktu
2,1	0,009	0
2,2	0,004	0
2,3	0,013	0,002
2,4	0,018	0,002
2,5	0,054	0,019
2,6	0,058	0,024
2,7	0,095	0,026
2,8	0,122	0,048
2,9	0,063	0,061
3	0,104	0,103
3,1	0,113	0,151
3,2	0,126	0,110
3,3	0,045	0,110
3,4	0,045	0,116
3,5	0,036	0,085
3,6	0,049	0,044
3,7	0,022	0,055
3,8	0,009	0,015
3,9	0,009	0,011
4	0	0,011

Berdasarkan Tabel 9 merupakan perolehan nilai dari hasil perhitungan menggunakan persamaan (1), dalam mencari nilai probabilitas atribut Indeks Prestasi semester 3. Perolehan nilai tertinggi pada mahasiswa tidak tepat waktu yaitu 0,126 yang terdapat pada indikator nilai IPs 3,2 dan nilai terendah pada mahasiswa tidak tepat waktu yaitu bernilai 0 atau dinyatakan tidak terdapat mahasiswa yang tidak tepat waktu pada indikator nilai IPs 4 dari jumlah total 221 orang. Sedangkan Perolehan nilai tertinggi pada mahasiswa tepat waktu yaitu 0,151 yang terdapat pada indikator nilai IPs 3,1 dan nilai terendah pada mahasiswa tepat waktu yaitu bernilai 0 atau dinyatakan tidak terdapat mahasiswa yang tepat waktu pada indikator nilai IPs 2,1 dan 2,2 dari jumlah total 454 orang.

Selanjutnya yaitu proses klasifikasi dengan menghitung semua nilai probabilitas atribut. Adapun proses untuk menghitung nilai label yaitu dengan mengalikan nilai seluruh atribut label. Setelah itu cari nilai tertinggi pada label, maka itu merupakan hasil dari klasifikasi tersebut. Untuk contoh proses perhitungan klasifikasi terdapat pada Tabel 10 berikut:

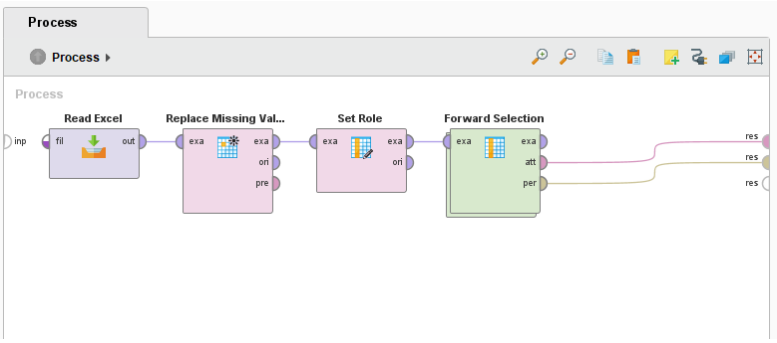
Tabel 10. Proses Perhitungan Klasifikasi

NIM	**JK**	**KS**	**Prodi**	**TM**	**IPs 3**	**Prediksi**
1506123	L	SMK	IF	2015	2,8	Tepat
Tepat	0,735	0,206	0,345	0,048	0,048	**0,00148**
Tidak	0,848	0,136	0,015	0,122	0,122	0,00015

Berdasarkan Tabel 10 proses perhitungan klasifikasi dilakukan dengan cara mengalikan seluruh nilai indikator atribut pada setiap *class* yaitu Tepat dan Tidak. Setelah menemukan hasilnya, lakukan perbandingan diantara kedua *class* tersebut. Berdasarkan hasil yang diperoleh, nilai tertinggi didapat dari *Class* Tepat. Maka ilustrasi penerapan

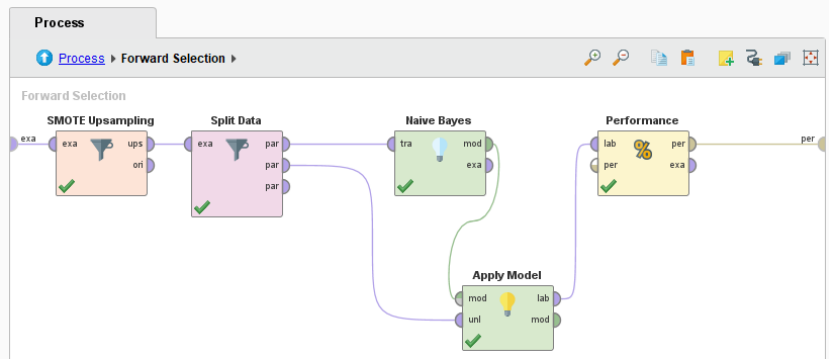
model klasifikasi untuk memprediksi ketepatan masa studi mahasiswa dinyatakan benar atau sesuai berdasarkan data yang dimiliki.

Untuk mengetahui performa hasil dari pemodelan, diperlukan *tools* untuk mengetahui nilai akurasi yang diperoleh. *Tools* yang digunakan pada penelitian ini yaitu *Rapidminer*. Untuk mengetahui tahapan dan operator-operator yang digunakan pada penelitian ini, yaitu terdapat pada Gambar 8 dan Gambar 9 sebagai berikut:



Gambar 8. *Data Preprocessing*

Berdasarkan Gambar 8 tahapan penggunaan *Rapidminer* untuk melakukan data *preprocessing* yaitu memilih operator *Read Excel* yang berfungsi untuk membaca jenis data yang digunakan. Selanjutnya operator *Replace Missing Values*, operator ini berfungsi untuk menormalisasikan data yang missing. Lalu operator *Set Role* berfungsi untuk mengatur label atribut yang akan digunakan pada proses pemodelan dan operator *Forward Selection*, operator ini berfungsi untuk mengetahui atribut pada data yang paling mempengaruhi terhadap ketepatan masa studi mahasiswa.



Gambar 9. *Modelling Phase*

Berdasarkan Gambar 9 tahapan penggunaan Rapidminer untuk melakukan proses *modelling* yaitu menggunakan operator SMOTE Upsampling yang berfungsi untuk menyeimbangkan *Imbalanced Data*. Lalu operator *Split Data*, operator ini berfungsi untuk membagi 2 data menjadi *Data Testing & Data Training*. Nilai yang digunakan untuk membagi *data testing* dan *data training* yaitu 0,7 dan 0,3. Pemilihan nilai perbandingan tersebut merupakan model yang paling optimal dalam memperoleh tingkat akurasi dibandingkan nilai perbandingan lainnya yang terdapat pada tabel Tabel 11 berikut:

Tabel 11. Perbandingan nilai *Split Data*

No.	Nilai perbandingan	*Accuracy*
1	0,4 : 0,6	83,46%
2	0,6 : 0,4	84,34%
3	0,3 :0,7	82,70%
4	**0,7 : 0,3**	**87,13%**
5	0,2 : 0,8	83,20%
6	0,8 : 0,2	84,62%
7	0,1 : 0,9	81,91%
8	0,9 : 0,1	85,56%

Selanjutnya operator *Naïve Bayes* digunakan sebagai algoritma untuk proses pemodelan dan operator *Apply Model* yang berfungsi membantu saat proses pemodelan. Setelah itu operator *Performance* untuk mengetahui performa model dan tingkat akurasi yang diperoleh.

5) ***Evaluation Phase***

Proses pada tahapan ini yaitu melakukan evaluasi model dari hasil pemodelan klasifikasi yang diperoleh pada tahapan sebelumnya. Proses yang dilakukan pada tahapan ini adalah melakukan uji akurasi menggunakan *Confusion Matrix*[13] dan disajikan dalam bentuk kurva ROC untuk mengetahui kategori klasifikasi yang diperoleh[14]. Pengujian ini menghasilkan performa model dalam berbagai nilai, berikut merupakan persamaan dalam memperoleh nilai dari hasil pengujian:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Keterangan:

P : Condition Positive
N : Condition Negative
TP : True Positive
TN : True Negative
FP : False Positve
FN : False Negative

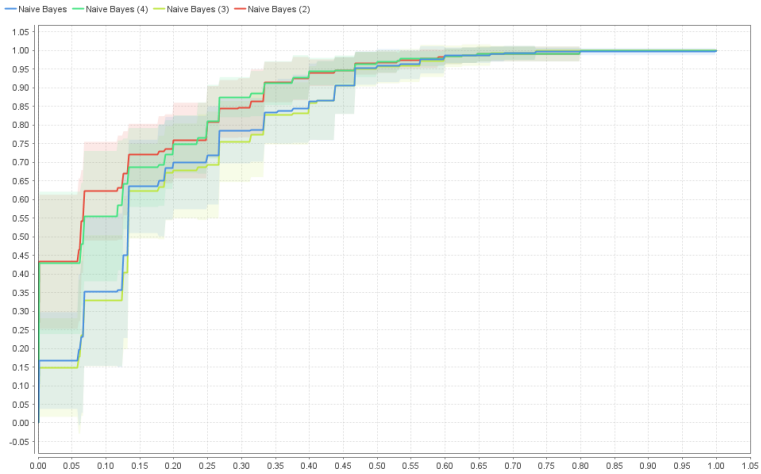
Hasil yang diperoleh dari penggunaan *tools Rapidminer* dibuat menjadi beberapa model. Proses pemodelan dengan menerapkan algoritma *Naïve Bayes* dilakukan beberapa kali untuk mengetahui model mana yang menghasilkan tingkat akurasi paling baik. Penyajian tabel perbandingan dalam penerapan algoritma *Naïve Bayes* terdapat pada Tabel 12 sebagai berikut:

Tabel 12. Perbandingan Model

Jenis Model	*Precision*	*Recall*	*Accuracy*	*AUC*
1	85,71%	79,41%	77,23%	0,82
2	81,42%	67,65%	76,10%	0,88
3	86,33%	**88,24%**	82,67%	0,87
4	**89,76%**	83,82%	**87,13%**	**0,92**

Keterangan:
Model 1: *Naive Bayes*
Model 2 : *Naive Bayes* & SMOTE
Model 3 : *Naive Bayes* & *Feature Forward Selection*
Model 4: *Naive Bayes* dengan *Feature Forward Selection* & SMOTE

Berdasarkan Tabel 12 dari beberapa model yang digunakan pada penelitian ini perolehan nilai *Precision* tertinggi didapat dari model 4 yaitu penerapan algoritma *Naïve Bayes* dengan *Forward Selection* & SMOTE. Untuk perolehan *Recall* tertinggi didapat dari model 3 yaitu penerapan algoritma *Naïve Bayes* dan *Forward Selection.* Sedangkan perolehan *Accuracy* tertinggi didapat dari model 4 yaitu penerapan algoritma *Naïve Bayes* dengan *Forward Selection* & SMOTE. Terakhir, perolehan nilai AUC tertinggi didapat dari model 4 yaitu penerapan algoritma *Naïve Bayes* dengan *Forward Selection* & SMOTE. Maka dapat disimpulkan bahwa model terbaik dalam mengimplementasikan algoritma *Naïve Bayes* yaitu terdapat pada model 4. Hasil perbandingan implementasi algoritma *Naïve Bayes* disajikan dalam kurva ROC *Compare* yang terdapat pada Gambar 10.



Gambar 10. Kurva Perbandingan Algoritma *Naïve Bayes*

Berdasarkan Gambar 10 terdapat 4 garis kurva dengan warna yang berbeda. Kurva tersebut berada diantara 2 sumbu, yaitu sumbu X dan sumbu Y. Sumbu X merupakan *False Positive Rate* sebagai *Specitivity* sedangkan sumbu Y merupakan *True Positive Rate* sebagai *Sensitivity*. Rentang nilai pada kedua sumbu dimulai dari 0,00 hingga 1,05. Jika garis kurva mendekati titik (1,1) maka proses klasifikasi yang diperoleh sangat baik [14].

Kurva perbandingan penerapan algoritma *Naïve Bayes* pada gambar terdiri dari 4 model. Pertama garis kurva yang berwarna biru merupakan hasil dari penerapan algoritma *Naïve Bayes*. Garis kurva yang berwarna kuning merupakan hasil dari penerapan algoritma *Naïve Bayes* dan SMOTE. Garis kurva yang berwarna merah merupakan hasil dari penerapan algoritma *Naïve Bayes* dan *Feature Forward Selection*. Sedangkan garis kurva yang berwarna hijau merupakan hasil dari penerapan algoritma *Naïve Bayes* dengan *Feature Forward Selection* dan SMOTE.

Selanjutnya pada tahap evaluasi terhadap model yang dipilih yaitu Implementasi algoritma *Naïve Bayes* dengan *Feature Forward Selection* & SMOTE menggunakan 2 metode pengujian yaitu *Confusion Matrix* dan Kurva ROC(AUC). Penggunaan *Confusion Matrix* ini bertujuan untuk menghitung performa model. Adapun nilai yang diperoleh terdapat pada Tabel 13.

Tabel 13. Performa Model

Predicted Condition	***True Condition***	
	Positive	*Negative*
Positive	123	22
Negative	13	114

Proses perhitungan *Confusion Matrix* ini bertujuan untuk mengetahui performa model. Nilai tersebut diperoleh dari *data testing* yang digunakan pada proses evaluasi. Berikut ini merupakan perolehan nilai untuk jumlah *True Positive* sebanyak 123, jumlah *True Negative* sebanyak 114, jumlah *False Positive* sebanyak 22 dan jumlah *False Negative* sebanyak 13. Berdasarkan nilai yang diperoleh maka proses perhitungan performa untuk mengetahui nilai *Accuracy, Precision* dan *Recall* adalah sebagai berikut:

Proses untuk mengetahui nilai *Accuracy* dilakukan dengan menggunakan persamaan (2). Adapun proses perhitungannya adalah sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$= \frac{123+114}{123+114+22+13} = \frac{237}{272} = 0{,}8713 / 87{,}13\%$$

Sedangkan proses untuk menghitung nilai *Recall* dilakukan dengan menggunakan persamaan (3). Adapun proses perhitungannya adalah sebagai berikut:

$$Recall = \frac{TP}{TP+FN}$$

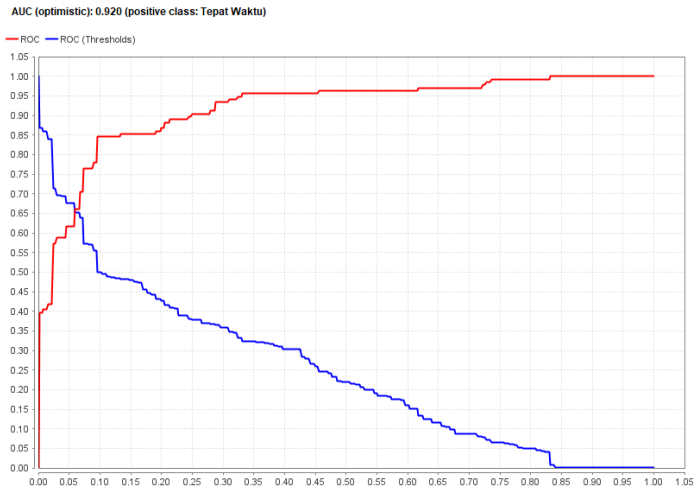
$$= \frac{123}{123+13} = \frac{123}{136} = 0{,}9044 / 90{,}44\%$$

Kemudian proses untuk menghitung nilai *Precision* dilakukan dengan menggunakan persamaan (4). Adapun proses perhitungannya adalah sebagai berikut:

$$Precision = \frac{TP}{TP+FP}$$

$$= \frac{123}{123+22} = \frac{123}{145} = 0{,}8482 / 84{,}82\%$$

Setelah diketahui performa model dengan perolehan nilai *Accuracy* 87,13%, *Precision* 84,82% dan *Recall* 90,44%. Maka prediksi ketepatan masa studi mahasiswa melalui proses klasifikasi dengan menerapkan algoritma *Naïve Bayes*, *Feature Forward Selection* dan *Synthetic Minority Over Sampling Technique* dinyatakan baik dan cukup akurat dikarenakan memiliki tingkat akurasi yang tinggi[12]. Proses evaluasi selanjutnya yaitu ditampilkan pada kurva ROC yang bertujuan untuk mengetahui kategori dan nilai dari hasil penelitian ini terdapat pada Gambar 12 sebagai berikut:



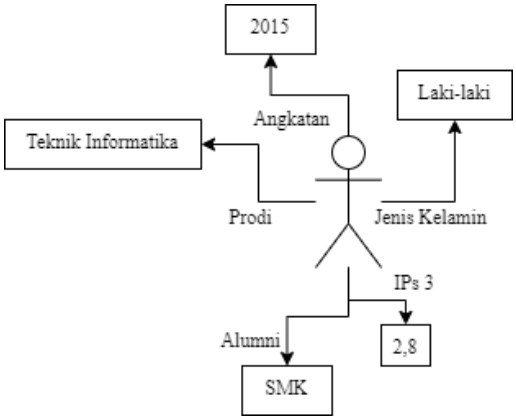
Gambar 12. Hasil Kurva ROC

Berdasarkan Gambar 12 dari penerapan algoritma *Naïve Bayes* dengan *Feature Forward Selection* dan SMOTE terdapat 2 garis kurva, yaitu garis kurva berwarna merah dan garis kurva berwarna biru. Garis kurva merah merupakan representasi data *training* sedangkan garis kurva biru merupakan representasi data *testing*. Kurva tersebut berada diantara 2 sumbu, yaitu sumbu X dan sumbu Y. Sumbu X merupakan *False Positive Rate* sebagai *Specitivity* sedangkan sumbu Y merupakan *True Positive Rate* sebagai *Sensitivity*. Rentang nilai pada kedua sumbu dimulai dari 0,00 hingga 1,05. Jika garis kurva mendekati titik (1,1) maka proses klasifikasi yang diperoleh sangat baik.

Garis kurva merah bergerak mulai dari titik (0,0) dengan arah tujuan kurva keatas menuju titik (1,1). Sedangkan garis kurva biru bergerak mulai dari titik (0,1) dengan arah tujuan kebawah menuju titik (1,0). Sebagaimana terlihat pada gambar, garis kurva membentuk cermin. Maka dapat dinyatakan jumlah kelas data telah seimbang, antara data *training* dan data *testing*. Perolehan nilai *Accuracy* sebesar 87,13%, menghasilkan nilai AUC sebesar 0,92. Berdasarkan perolehan nilai AUC tersebut, maka dapat disimpulkan bahwa AUC tersebut termasuk pada kategori *Excellent Classification* [14].

6) ***Deployment Phase***

Fase penyebaran ini menjadi gambaran sebagaimana hasil dari pemodelan ini akan digunakan oleh program studi, dalam memberikan arahan dan bimbingan terhadap mahasiswa berdasarkan hasil prediksi ketepatan masa studi mahasiswa yang dilakukan pada semester 6. Jika terdapat mahasiswa yang terprediksi akan terlambat dalam menyelesaikan studi. Maka program studi akan memberikan perhatian lebih terhadap mahasiswa tersebut agar dapat menyelesaikan studi tepat waktu. Untuk contoh penerapan model klasifikasi untuk memprediksi ketepatan masa studi mahasiswa pada Gambar 13:



Gambar 13. Ilustrasi Penerapan Model Klasifikasi

Berdasarkan Gambar 13 ilustrasi penerapan model klasifikasi untuk memprediksi ketepatan masa studi mahasiswa di Institut Teknologi Garut. Data yang tertera pada gambar merupakan indikator atribut yang dimiliki mahasiswa tersebut. Untuk mengetahui mahasiswa tersebut akan menyelesaikan studi tepat waktu atau tidak tepat waktu, diperlukan proses klasifikasi. Proses klasifikasi yang dilakukan yaitu dengan cara menghitung nilai pada setiap atribut yang sudah diperoleh pada tabel distribusi dan membagi 2 kelas yaitu Tepat Waktu dan Tidak Tepat waktu. Untuk proses perhitungan klasifikasi dalam memprediksi ketepatan masa studi mahasiswa terdapat pada Tabel 14:

Tabel 14. Proses Prediksi Model Klasifikasi

NIM	JK	KS	Prodi	TM	IPs 3	Prediksi
1506123	L	SMK	IF	2015	2,8	Tepat
Tepat	0,735	0,206	0,345	0,048	0,048	**0,00148**
Tidak	0,848	0,136	0,015	0,122	0,122	0,00015

Berdasarkan Tabel 14 proses perhitungan klasifikasi dilakukan dengan cara mengalikan seluruh nilai indikator atribut pada setiap *class*. Setelah menemukan hasilnya, lakukan perbandingan diantara kedua *class* tersebut. Berdasarkan hasil yang diperoleh, nilai tertinggi didapat dari *Class* Tepat Waktu. Maka ilustrasi penerapan model klasifikasi yang terdapat pada *Phase Modelling* dalam memprediksi ketepatan masa studi mahasiswa dapat digunakan karena hasil klasifikasi yang diperoleh benar atau sesuai berdasarkan data yang dimiliki.

B. Pembahasan Hasil

Pada latar belakang telah disinggung bahwa masa studi mahasiswa telah diatur pada Permendikbud No. 49 Tahun 2014 mengenai Standar Nasional Pendidikan Tinggi (SN-PT) yaitu beban belajar minimal mahasiswa pada jenjang S1 diberi batas waktu 4-5 tahun (8-10 semester)[1]. Namun pada proses berlangsungnya akademik, tidak semua mahasiswa dapat menyelesaikan studi sesuai jangka waktu yang telah ditentukan. Dari hasil yang diperoleh pada penelitian ini memberikan dampak positif atau manfaat, baik secara teoritis maupun praktis.

Pembahasan mengenai keselarasan penelitian merupakan kejelasan apakah hasil dari penelitian ini sesuai/selaras berdasarkan latar belakang penelitian, rumusan masalah, dan hasil dari penelitian yang diperoleh. Dari hasil penelitian yang diperoleh, setiap Program Studi di Institut Teknologi Garut harus menangani permasalahan mahasiswa dalam menyelesaikan studi. Terdapat 5 permasalahan yang menghambat mahasiswa dalam menyelesaikan studi antara lain: Mahasiswa mengambil cuti terlalu lama, Terdapat nilai D pada mata kuliah yang telah diambil, Tidak Lulus Kerja Praktik, Penerbitan Jurnal Skripsi yang memakan waktu dan Tidak Lulus Sidang Skripsi. Untuk menangani

permasalahan tersebut, maka diperlukan suatu proses prediksi agar dapat mencegah/ menangani mahasiswa yang terindikasi akan terlambat dalam menyelesaikan studi. Permasalahan dalam menentukan faktor/atribut yang mempengaruhi terhadap ketepatan masa studi mahasiswa, dilakukan penyeleksian atribut. Proses penyeleksian ini dilakukan untuk mengetahui atribut manakah yang paling mempengaruhi terhadap ketepatan masa studi mahasiswa. Penerapan *Feature Forward Selection* membantu dalam proses penyeleksian atribut. Hasil dari penerapan fitur ini yang terdapat pada tahapan *Data Preparation Phase*, pada proses *Data Reduction* memperoleh 5 atribut yang paling mempengaruhi terhadap ketepatan masa studi mahasiswa. Kelima atribut tersebut terdiri dari Jenis Kelamin, Kategori Sekolah, Tahun Masuk, Program Studi dan Indeks Prestasi Semester 3.

Berdasarkan penelitian yang diperoleh, proses klasifikasi untuk memprediksi ketepatan masa studi mahasiswa menggunakan algoritma *Naïve Bayes* dengan *Feature Forward Selection* dan SMOTE, menghasilkan nilai akurasi yang tinggi. Hal ini selaras dengan penelitian Nuraeni, dkk [8] bahwa penggunaan *Feature Forward Selection* mempengaruhi terhadap tingkat akurasi. Selain itu penerapan teknik SMOTE dalam mengatasi *Imbalance Class* memberikan dampak negatif terhadap perolehan nilai akurasi, jika tanpa menggunakan fitur seleksi atribut. Hal ini selaras dengan penelitian Kasana, dkk [10] bahwa penerapan teknik SMOTE harus disertai penggunaan fitur seleksi atribut seperti *Feature Forward Selection*.

Berdasarkan hasil prediksi yang diperoleh dari proses klasifikasi yang telah dilakukan, memperoleh hasil yang sesuai dengan data yang dimiliki. Maka keselarasan hasil dari penelitian ini, dapat disimpulkan selaras dengan apa yang dibutuhkan pada setiap Program Studi di Institut Teknologi Garut dalam menangani permasalahan mengenai ketepatan masa studi mahasiswa sarjana. Selaras dengan penelitian Kurniadi, dkk [6] pada objek penelitian yang sama yaitu di Institut Teknologi Garut, namun perbedaaanya dari *output* yang dihasilkan yang bersifat *multi-class*. yang mana penelitian tersebut dalam menangani permasalahan mengenai ketepatan masa studi mahasiswa dilakukan dengan menghasilkan model yang mampu memprediksi kinerja akademik mahasiswa dengan merepresentasikan *output* secara *multiclass* yang bersifat *multi-level representation*.

IV. KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan yang diperoleh, maka dapat ditarik kesimpulan berdasarkan nilai akurasi yang diperoleh bahwa, penerapan *Feature Forward Selection* dan SMOTE mempengaruhi terhadap perolehan tingkat akurasi. Hasil pengujian model prediksi ketepatan masa studi mahasiswa menggunakan *Confusion Matrix*, menunjukkan efektivitas model tertinggi diperoleh dari penerapan algoritma *Naïve Bayes* dengan *Feature Forward Selection* dan SMOTE, dibandingkan implementasi algoritma *Naïve Bayes* tanpa *Feature Forward Selection* dan SMOTE. Perolehan hasil performa model ini sangat baik yaitu nilai *Accuracy* sebesar 87,13 %, nilai *Recall* sebesar 83,82% dan nilai *Precision* sebesar 89,76%. Perolehan nilai AUC pada visualisasi kurva ROC menghasilkan kinerja model sebesar 0,92 yang termasuk pada kategori *Excellent Classification.* Maka model prediksi ketepatan masa studi mahasiswa ini dapat diterapkan pada setiap program studi sebagai bahan perencanaan, pengawalan studi serta membimbing mahasiswa dalam menangani permasalahan mengenai ketepatan masa studi mahasiswa Sarjana. Adapun saran untuk penelitian berikutnya, yaitu dengan menambahkan atribut data mengenai permasalahan pribadi, diluar lingkup akademik yang dapat menghambat mahasiswa dalam menyelesaikan studi untuk meningkatkan kualitas data agar lebih aktual.

DAFTAR PUSTAKA

[1] E. P. K. Orpa, E. F. Ripanti, and Tursina, "Model Prediksi Awal Masa Studi Mahasiswa Menggunakan Algoritma Decision tree c4.5," vol. 7, no. 4, pp. 272–278, 2019.

[2] ITG, "Pedoman Akademik Institut Teknologi Garut," *Angew. Chemie Int. Ed. 6(11), 951–952.*, pp. 10–27, 2021.

[3] M. R. A. Fernanda, P. Sokibi, and R. Fahrudin, "Sistem Prediksi Ketepatan Kelulusan Mahasiswa Berdasarkan Data Akademik Dan Non Akademik Menggunakan Metode K-Means (Studi Kasus : Universitas Catur Insan Cendekia)," *J. Digit*, vol. 11, no. 1, p. 89, 2021.

[4] S. Yunianita, N. Setiani, and S. Mulyati, "Prediksi Ketepatan Masa Studi Mahasiswa dengan Algoritma Pohon Keputusan C45," pp. 22–28, 2018.

[5] I. A. Nikmatun and I. Waspada, "Implementasi Data Mining Untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor," vol. 10, no. 2, pp. 421–432, 2019.

[6] D. Kurniadi, E. Abdurachman, H. Leslie, H. Spits, and W. Suparta, "Predicting Student Performance With Multi-Level Representation In An Intelligent Academic Recommender System Using Backpropagation Neural Network," no. August, 2021.

[7] A. Jananto, Sulastri, E. Nur Wahyudi, and Sunardi, "Data Induk Mahasiswa sebagai Prediktor Ketepatan Waktu Lulus Menggunakan Algoritma CART Klasifikasi Data Mining," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 10, no. 1, pp. 71–78, 2021.

[8] F. Nuraeni, Y. H. Agustin, S. Rahayu, D. Kurniadi, Y. Septiana, and S. M. Lestari, "Student Study Timeline Prediction Model Using Naïve Bayes Based Forward Selection Feature," *8th Int. Conf. ICT Smart Soc. Digit. Twin Smart Soc. ICISS 2021 - Proceeding*, pp. 1–5, 2021.

[9] Y. D. Atma and A. Setyanto, "Perbandingan Algoritma C4.5 dan K-NN Dalam Identifikasi Mahasiswa Berpotensi Drop Out," vol. 2, no. 2, 2018.

[10] A. N. Kasanah, Muladi, and U. Pujianto, "Penerapan Teknik SMOTE Untuk Mengatasi Imbalance Class Dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," vol. 1, no. 10, 2019.

[11] I. G. I. Suwardika, I. G. N. Suariana, I. B. P. Bhiantara, and N. Y. Arso, "Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naive Bayes: Studi Kasus Fakultas Ekonomi dan Bisnis Universitas Pendidikan Nasional," vol. 4, no. 2, 2019.

[12] R. Y. Hayuningtyas, "Penerapan Algoritma Naïve Bayes untuk Rekomendasi Pakaian Wanita," vol. 6, no. 1, pp. 18–22, 2019.

[13] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," vol. 5, no. November 2019, pp. 697–711, 2021.

[14] T. Arifin and S. Syalwah, "Prediksi Keberhasilan Immunotherapy Pada Penyakit Kutil Dengan Menggunakan Algoritma Naïve Bayes," vol. 2, no. 1, pp. 38–43, 2020.