

# **SKRIPSI**

## **IMPLEMENTASI ALGORITMA C4.5 DAN *NAÏVE BAYES CLASSIFICATION* UNTUK MENGLASIFIKASI MAHASISWA BERPOTENSI *DROP OUT***



Sebagai Salah Satu Syarat untuk  
Memperoleh Gelar Sarjana Komputer  
Fakultas Sains dan Teknologi  
Universitas Islam Negeri Syarif Hidayatullah Jakarta

**DISUSUN OLEH:**

**HAMZAH AJI PRATAMA**  
**NIM. 11170930000091**

**PROGRAM STUDI SISTEM INFORMASI  
FAKULTAS SAINS DAN TEKNOLOGI  
UNIVERSITAS ISLAM NEGERI SYARIF HIDAYATULLAH  
JAKARTA  
2024 M / 1445 H**

# **SKRIPSI**

## **IMPLEMENTASI ALGORITMA C4.5 DAN *NAÏVE BAYES CLASSIFICATION***

### **UNTUK MENGLASIFIKASI MAHASISWA BERPOTENSI *DROP OUT***



Universitas Islam Negeri  
**SYARIF HIDAYATULLAH JAKARTA**

Sebagai Salah Satu Syarat untuk  
Memperoleh Gelar Sarjana Komputer  
Fakultas Sains dan Teknologi  
Universitas Islam Negeri Syarif Hidayatullah Jakarta

**DISUSUN OLEH:**

**HAMZAH AJI PRATAMA**  
**NIM. 11170930000091**

**PROGRAM STUDI SISTEM INFORMASI  
FAKULTAS SAINS DAN TEKNOLOGI  
UNIVERSITAS ISLAM NEGERI SYARIF HIDAYATULLAH  
JAKARTA  
2024 M / 1445 H**

## **PERNYATAAN**

DENGAN INI SAYA MENYATAKAN BAHWA SKRIPSI INI BENAR-BENARHASIL KARYA SENDIRI DAN BELUM PERNAH DIAJUKAN SEBAGAI SKRIPSI ATAU KARYA ILMIAH PADA PERGURUAN TINGGI ATAU LEMBAGA MANAPUN.

Tangerang Selatan, 08 Oktober 2023



**HAMZAH AJI PRATAMA**  
**NIM. 11170 930000091**

Universitas Islam Negeri  
**YARIF HIDAYATULLAH JAKARTA**

## ABSTRAK

**Hamzah Aji Pratama – 11170930000091**, Implementasi Algoritma C4.5 dan *Naïve Bayes Classification* untuk Mengklasifikasi Mahasiswa Berpotensi *Drop Out* di bawah bimbingan **Dr. Qurrotul Aini, M.T.** dan **Rinda Hesti Kusumaningtyas, MMSI**.

Salah satu faktor penentu kualitas sebuah perguruan tinggi adalah tingkat keberhasilan mahasiswa dalam menyelesaikan studi tepat waktu. Dalam konteks ini, data akademik dari UIN Syarif Hidayatullah Jakarta mengungkapkan bahwa terdapat peningkatan signifikan sebesar 7,95% dalam persentase mahasiswa Program Studi Sistem Informasi, Fakultas Sains dan Teknologi yang mengalami *drop out* dari perguruan tinggi sejak tahun 2012 hingga 2015. Penelitian ini bertujuan untuk mengidentifikasi kriteria kegagalan mahasiswa dalam menyelesaikan masa studi mereka. Selain itu, penelitian ini berupaya untuk mengimplementasikan algoritma C4.5 dan *Naïve Bayes Classification*, serta menganalisis hasil kinerjanya. Algoritma C4.5 memiliki keunggulan dalam menangani data tidak terstruktur, menghasilkan pohon keputusan yang mudah diinterpretasikan, serta efisien dalam penggunaan memori dan komputasi. *Naïve Bayes Classification* mudah diimplementasikan, efisien dalam komputasi, tahan terhadap *overfitting*, dan memberikan probabilitas kelas yang jelas untuk setiap *instance*. Dengan demikian, penelitian ini dapat memberikan hasil prediksi terhadap mahasiswa yang memiliki potensi untuk *drop out*. Oleh karena itu, penting bagi perguruan tinggi untuk memahami karakteristik mahasiswa yang mengalami *drop out* agar faktor-faktor penyebab kegagalan mereka dapat diidentifikasi. Dalam rangka melakukan penelitian ini, metode SEMMA (*Sample, Explore, Modify, Model, dan Assess*) digunakan sebagai alur penelitian yang komprehensif. Peneliti menggunakan *Google Colab* dan bahasa pemrograman Python untuk melakukan pengklasifikasian dan prediksi data. Dua algoritma yang digunakan adalah *Decision Tree* C4.5 dan *Naïve Bayes Classification*. Hasil dari penggunaan algoritma *Decision Tree* C4.5 menunjukkan tingkat akurasi sebesar 94,44%, sedangkan *Naïve Bayes Classification* mencapai akurasi sebesar 93%. Dari kedua algoritma tersebut, algoritma C4.5 menunjukkan performa terbaik, sehingga dapat diterapkan untuk mengklasifikasikan mahasiswa yang berpotensi mengalami *drop out*. Dengan demikian, penelitian ini memberikan kontribusi yang penting dalam mengidentifikasi faktor-faktor yang berperan dalam kegagalan mahasiswa dalam menyelesaikan studi mereka. Dengan menggunakan algoritma C4.5, perguruan tinggi dapat melakukan klasifikasi dan prediksi yang lebih efektif terhadap mahasiswa yang memiliki potensi untuk mengalami *drop out*. Terdapat tiga faktor utama dalam mengklasifikasi mahasiswa berpotensi *drop out*, yaitu kelulusan dalam semua matakuliah, penyelesaian laporan Praktik Kerja Lapangan (PKL), serta sedang dalam proses melakukan atau mengerjakan laporan skripsi. Penelitian ini memberikan landasan untuk penelitian selanjutnya yang melibatkan pengembangan aplikasi atau *dashboard* khusus untuk mengidentifikasi faktor-faktor yang memengaruhi mahasiswa berpotensi *drop out*, termasuk *variabel*

eksternal seperti keluarga, keuangan, pertemanan, dan pekerjaan. Selain itu, perlu dilakukan penelitian lebih lanjut untuk membandingkan efektivitas algoritma yang digunakan dalam penelitian ini dengan algoritma lain seperti *K-Nearest Neighbor* (KNN) atau *Support Vector Machine* (SVM) guna meningkatkan akurasi prediksi mahasiswa berpotensi *drop out*.

Kata Kunci: C4.5, *naïve bayes*, mahasiswa, *confusion matrix*, *cross validation*

Bab 1 - 5 + xiv Halaman + 70 Halaman + 19 Gambar + 15 Tabel + Lampiran

Pustaka Acuan (37, 2007-2021)



## KATA PENGANTAR

*Assalamu'alaikum Warahmatullaah Wabarakaatuh*

Puji syukur peneliti ucapkan kepada Allah SWT karena atas nikmat dan berkha-Nya peneliti dapat menyelesaikan skripsi ini sebagai syarat untuk mencapai gelar Sarjana Komputer Program Studi Sistem Informasi Fakultas Sains dan Teknologi Universitas Islam Negeri Syarif Hidayatullah Jakarta. Dalam prosesnya, penulisan skripsi ini tidak lepas dari berbagai bantuan, dukungan, saran, dan kritik yang peneliti dapatkan sehingga dalam kesempatan ini peneliti mengucapkan terima kasih kepada:

1. Bapak Husni Teja Sukmana, S.T., M.Sc, Ph.D selaku dekan Fakultas Sains dan Teknologi.
2. Ibu Dr. Qurrotul Aini, M.T selaku ketua Program Studi Sistem Informasi dan Bapak Ir. Eri Rustamaji, MBA selaku sekretaris Program Studi Sistem Informasi.
3. Ibu Dr. Qurrotul Aini, MT. selaku dosen pembimbing I dan Ibu Rinda Hesti Kusumaningtyas, M.M.S.I. selaku dosen pembimbing II yang secara kooperatif telah meluangkan waktu dan memberikan bimbingan, bantuan, semangat, dan motivasi kepada peneliti dalam menyelesaikan skripsi ini.
4. Seluruh dosen, staf karyawan Fakultas Sains dan Teknologi yang telah memberikan bantuan dan kerjasama yang terjalin dari awal perkuliahan.
5. Mama, Papa, Dek Syifa, Dek Haikal dan keluarga peneliti yang senantiasa mendoakan, dan mendukung peneliti dalam menyelesaikan skripsi.

6. Para pengurus dan anggota HMI, HIMSI, IMSII, Himakotas yang telah memberikan banyak pelajaran dan kesempatan untuk terus berproses didalamnya, walaupun terkadang membuat peneliti menjadi hilang fokus sehingga melupakan kewajibannya.
7. Kepada semua circle grup whatsapp Kaki Lima, Yu Kaya Yu, Kerkel Terus, Para Pengabdi, Warsis 17 terimakasih telah menjadi tempat peneliti sambat di kehidupan akhir zaman ini.
8. Adelya Faramita, sebagai partner peneliti, khusus nya dalam mengerjakan skripsi ini dan menjadi tempat berkeluh kesah.
9. Seluruh teman peneliti prodi Sistem Informasi Angkatan 2017 yang saling memberikan semangat serta saling mendoakan satu sama lain untuk menjadi orang sukses.
10. Seluruh pihak yang secara langsung maupun tidak langsung telah membantu, mendukung, serta mendoakan peneliti dalam menyelesaikan skripsi ini. Meski tidak tertulis namun tidak mengurangi rasa hormat dan terima kasih dari penulis.

Peneliti menyadari bahwa masih ada kekurangan dalam skripsi ini, oleh karena itu peneliti terbuka terhadap saran dan kritik yang membangun dari pembaca. Peneliti berharap dapat memberikan manfaat bagi pembaca.

Tangerang Selatan, 08 Oktober 2023

**HAMZAH AJI PRATAMA**  
NIM. 11170930000091

## DAFTAR ISI

<b>LEMBAR PENGESAHAN SKRIPSI .....</b>	<b>ii</b>
<b>PENGESAHAN UJIAN .....</b>	<b>iii</b>
<b>PERNYATAAN .....</b>	<b>iv</b>
<b>ABSTRAK .....</b>	<b>v</b>
<b>KATA PENGANTAR .....</b>	<b>vii</b>
<b>DAFTAR ISI .....</b>	<b>ix</b>
<b>DAFTAR GAMBAR .....</b>	<b>xii</b>
<b>DAFTAR TABEL .....</b>	<b>xiv</b>
<b>BAB 1 PENDAHULUAN .....</b>	<b>1</b>
1.1 Latar Belakang .....	1
1.2 Identifikasi Masalah .....	5
1.3 Rumusan Masalah .....	6
1.4 Batasan Masalah .....	6
1.5 Tujuan Penelitian .....	7
1.6 Manfaat Penelitian .....	8
1.7 Metodologi Penelitian .....	8
1.7.1 Pengumpulan Data .....	8
1.7.2 Tahapan <i>Data Mining</i> .....	9
1.8 Sistematika Penulisan .....	10
<b>BAB 2 TINJAUAN PUSTAKA .....</b>	<b>13</b>
2.1 Menjauhi Sifat Lemah dan Malas .....	13



2.2	<i>Data Mining</i>	15
2.3	Klasifikasi	16
2.4	SEMMA	19
2.5	<i>Split Validation</i>	21
2.6	<i>Confusion Matrix</i>	22
2.7	Algoritma	24
2.7.1	C4.5	24
2.7.2	<i>Naïve Bayes Classification</i>	29
2.8	Python	30
2.9	Library	31
2.9.1	<i>Numpy</i>	34
2.9.2	<i>Pandas</i>	35
2.9.3	<i>Matplotlib</i>	35
2.9.4	<i>Seaborn</i>	36
2.10	Mahasiswa	36
2.11	<i>Drop Out</i>	37
2.12	Penelitian Sejenis	39
2.13	Ranah Sejenis	43
<b>BAB 3 METODELOGI PENELITIAN</b>		<b>46</b>
3.1	Metode Pengumpulan Data	46
3.1.1	Studi Pustaka	46
3.1.2	Wawancara	46
3.1.2	Obeservasi	47

3.2 Tahapan Penelitian .....	48
3.3 Kerangka Penelitian .....	50
<b>BAB 4 HASIL DAN PEMBAHASAN .....</b>	<b>52</b>
4.1 <i>Sample</i> .....	52
4.2 <i>Explore</i> .....	53
4.3 <i>Modify</i> .....	56
4.4 <i>Model</i> .....	57
4.5 <i>Access</i> .....	75
4.6 Evaluasi .....	83
<b>BAB 5 PENUTUP .....</b>	<b>85</b>
5.1 Kesimpulan .....	85
5.2 Saran .....	86
<b>DAFTAR PUSTAKA .....</b>	<b>87</b>
<b>LAMPIRAN .....</b>	<b>xv</b>

Universitas Islam Negeri  
YARIF HIDAYATULLAH JAKARTA

## DAFTAR GAMBAR

<b>Gambar 1.1</b> Grafik Data Kelulusan Mahasiswa .....	3
<b>Gambar 2.1</b> Konsep Klasifikasi .....	17
<b>Gambar 2.2</b> Contoh Klasifikasi menurut (Gorunescu, 2011) .....	19
<b>Gambar 2.3</b> SEMMA Data Mining Process menurut (Wilson, 2021) .....	21
<b>Gambar 2.4</b> Contoh Pohon Keputusan .....	24
<b>Gambar 2.5</b> Logo library Numpy.....	34
<b>Gambar 2.6</b> Logo library Pandas .....	35
<b>Gambar 2.7</b> Logo library Matplotlib.....	36
<b>Gambar 2.8</b> Logo library Seaborn .....	36
<b>Gambar 2.9</b> Ranah Penelitian .....	44
<b>Gambar 3.1</b> Kerangka Penelitian .....	51
<b>Gambar 4.1</b> Template Email Permohonan Permintaan Data .....	52
<b>Gambar 4.2</b> Data Mahasiswa yang Dikirimkan oleh Pustipanda.....	53
<b>Gambar 4.3</b> Faktor-Faktor Mahasiswa Berpotensi DO .....	54
<b>Gambar 4.4</b> Dataset Mengubahnya Menjadi Numerik .....	56
<b>Gambar 4.5</b> Decision Tree C4.5.....	59
<b>Gambar 4.6</b> Diagram Lingkaran Hasil Akhir Algoritma C4.5 .....	60
<b>Gambar 4.7</b> Diagram Lingkaran Hasil Akhir Algoritma NBC .....	61
<b>Gambar 4.8</b> Confusion Matrix pada Algoritma C4.5 .....	76
<b>Gambar 4.9</b> Hasil Pengukuran Kinerja menggunakan Algoritma C4.5 .....	77
<b>Gambar 4.10</b> Confusion Matrix pada Algoritma NBC.....	79

**Gambar 4.11** Hasil Pengukuran menggunakan Algoritma NBC ..... 81

**Gambar 4.12** ROC *Curve* pada Algoritma NBC..... 82



Universitas Islam Negeri  
**YARIF HIDAYATULLAH JAKARTA**

## DAFTAR TABEL

<b>Tabel 1.1</b> Data Status Kelulusan Mahasiswa Sistem Informasi .....	3
<b>Tabel 2.1</b> Bentuk <i>Confusion Matrix</i> .....	22
<b>Tabel 2.2</b> Hasil Hitungan Akurasi.....	31
<b>Tabel 2.3</b> Hasil Akurasi .....	31
<b>Tabel 2.4</b> Penelitian Sejenis.....	40
<b>Tabel 3.1</b> Waktu dan Tempat Penelitian .....	47
<b>Tabel 3.2</b> Perangkat Penelitian .....	47
<b>Tabel 3.3</b> <i>Dataset</i> Mahasiswa.....	48
<b>Tabel 4.1</b> <i>Table Explore</i> Data Mahasiswa.....	55
<b>Tabel 4.2</b> Perbandingan Data <i>Train</i> dan Data <i>Test</i> Metode C4.5 dan NBC.....	57
<b>Tabel 4.3</b> Hasil dari model Algoritma C4.5 .....	62
<b>Tabel 4.4</b> Hasil dari Model Algoritma <i>Naive Bayes Classification</i> .....	74
<b>Tabel 4.5</b> Hasil <i>Accuracy</i> , <i>Precision</i> , dan <i>Recall</i> pada metode C4.5.....	75
<b>Tabel 4.6</b> <i>Confusion Matrix</i> pada Algoritma C4.5.....	76
<b>Tabel 4.7</b> Hasil <i>Accuracy</i> , <i>Precision</i> , dan <i>Recall</i> pada Algoritma NBC.....	79
<b>Tabel 4.8</b> <i>Confusion Matrix</i> pada Algoritma NBC .....	80
<b>Tabel 4.9</b> Hasil Perbandingan Kinerja .....	83

# **BAB 1**

## **PENDAHULUAN**

### **1.1 Latar Belakang**

UIN Syarif Hidayatullah Jakarta adalah universitas yang terdiri atas beberapa fakultas salah satunya adalah Fakultas Sains dan Teknologi. Dalam pedoman akademik Universitas Islam Negeri Syarif Hidayatullah Jakarta bidang pendidikan tahun 2018 pada BAB II disebutkan bahwa Program Sarjana (S1) reguler memiliki beban studi sekurang-kurangnya 144 (seratus empat puluh empat) SKS (Satuan Kredit Semester) dan sebanyak-banyaknya 150 (seratus lima puluh) sks yang dijadwalkan untuk 8 (delapan) semester atau 4 tahun dan dapat ditempuh dalam waktu kurang dari 8 (delapan) semester dan paling lama 12 (dua belas) semester atau 6 tahun (Fadhilah *et al.*, 2018).

Informasi mengenai kelulusan mahasiswa adalah hal terpenting untuk membantu program studi (prodi) untuk dapat bersaing dan lebih berkembang. Informasi ini sangat diperlukan untuk mengoptimalkan kemungkinan mahasiswa dapat menyelesaikan masa studinya, guna menjaga kualitas pendidikan di program studi. Pengukuran ketepatan kelulusan mahasiswa menggambarkan kompetensi lulusan tersebut dan keefektifan kurikulum yang diterapkan sehingga dapat digunakan sebagai acuan dalam pengambilan keputusan setiap tahunnya terhadap kebutuhan masyarakat dunia khususnya tingkat daerah (Munawir & Iqbal, 2019).

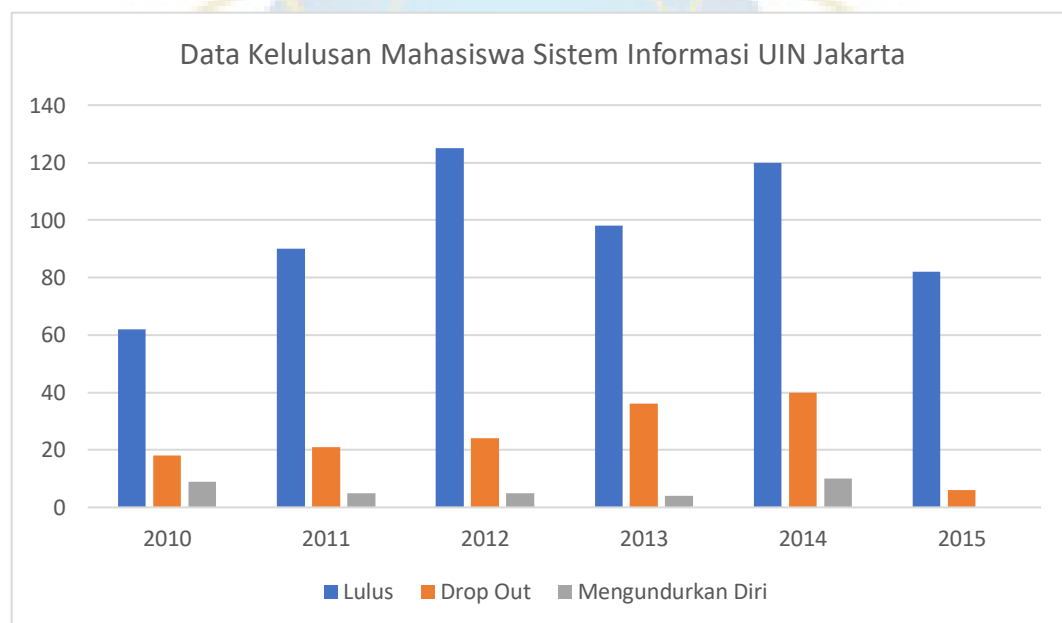
Penelitian menggunakan metode Algoritma Decision C4.5 (Sinaga *et al.*, 2021) dan (Rahman *et al.*, 2020) menunjukkan tingkat akurasi yang cukup tinggi. Namun, penelitian (Rahman *et al.*, 2020) menyarankan peningkatan jumlah *record* data *training* untuk meningkatkan akurasi, karena jumlah data *training* memengaruhi hasil akhir. Lebih lanjut, perbandingan dengan metode lain juga diperlukan untuk memvalidasi klasifikasi. Di sisi lain, penelitian (Yuniarti *et al.*, 2020) dengan 5.934 data mencapai akurasi 99,41%, menunjukkan bahwa penggunaan data yang lebih besar dapat meningkatkan performa metode, seperti yang terbukti dalam penelitian (Iskandar *et al.*, 2021) menggunakan algoritma *Naïve Bayes Classification*.

Pihak Akademik Universitas Islam Negeri (UIN) Syarif Hidayatullah Jakarta telah mengeluarkan Surat Keputusan (SK) Rektor No. 154 Tahun 2019 tentang Pemberhentian Mahasiswa yang sudah tidak aktif dan melewati batas semester yang telah ditentukan dalam peraturan, dalam SK yang dikeluarkan oleh Rektor UIN Syarif Hidayatullah Jakarta ada 828 mahasiswa yang masuk ke daftar *drop out*. Penurunan tingkat kelulusan secara signifikan menjadi permasalahan yang serius, bahkan dapat mempengaruhi akreditasi perguruan tinggi. Oleh karena itu pemantauan dan evaluasi secara berkala terhadap kecenderungan tingkat kelulusan mahasiswa diperlukan (Sudriyanto *et al.*, 2021).

Menurut data Akademik UIN Syarif Hidayatullah Jakarta, jumlah mahasiswa Prodi Sistem Informasi, Fakultas Sains dan Teknologi yang tercatat *drop out* mengalami peningkatan presentase sebesar 7,95% dari tahun masuk mahasiswa 2012-2014. Data ditunjukkan pada Tabel 1.1 dan Gambar 1.1.

**Tabel 1.1** Data Status Kelulusan Mahasiswa Sistem Informasi

TAHUN MASUK	JUMLAH MAHASISWA	LULUS	<i>DROP OUT</i>	MENGUNDURKAN DIRI	<i>DROP OUT (%)</i>
2010	89	62	18	9	20,22%
2011	116	90	21	5	18,10%
2012	154	125	24	5	15,58%
2013	138	98	28	12	20,29%
2014	170	120	40	10	23,53%
2015	88	82	6	0	6,82%

**Gambar 1.1** Grafik Data Kelulusan Mahasiswa Sistem Informasi UIN Jakarta (Akademik Pusat UIN Syarif Hidayatullah Jakarta, 2022)

Dalam proses klasifikasi data akademik mahasiswa, banyak model algoritma yang dapat digunakan, di antaranya adalah model Algoritma C4.5 dan Algoritma *Naïve Bayes Classification*. Algoritma C4.5 menggunakan konsep *information gain* atau *entropy reduction* untuk memilih pembagian yang optimal (Septiani, 2017). Algoritma C4.5 ini menggunakan metode *Decission Tree*. Menurut Septiani (2017)



*Decision Tree* mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Dapat disimpulkan bahwa kelebihan algoritma C4.5 dapat menghasilkan pohon keputusan yang mudah diinterpretasikan, memiliki tingkat akurasi yang dapat diterima, efisien dalam menangani atribut bertipe diskret dan dapat menangani atribut bertipe diskret dan numerik. Sedangkan algoritma *Naive Bayes Classification* adalah model yang sederhana, namun dapat bersaing dengan model algoritma lainnya. Implementasinya pun tidak terlalu rumit, cocok untuk mengevaluasi probabilitas bersyarat.

Lain halnya dengan Algoritma *Naïve Bayes* atau yang bisa disebut dengan *Bayesian classification*. *Bayesian classification* adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu *class*. *Bayesian classification* didasarkan pada teorema *Bayes* yang memiliki kemampuan klasifikasi serupa dengan *decision tree* dan *neural network*. *Bayesian classification* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam *database* dengan data yang besar (Annur, 2018).

Penelitian Sinaga *et al.* (2021) menunjukan tentang penerapan algoritma *Decision Tree* untuk klasifikasi mahasiswa berpotensi *drop out* di Universitas Advent Indonesia menghasilkan *accuracy* 90.00%, *precision* 87,50%, dan *recall* sebesar 100%. Pada penelitian Rahman *et al.* (2020) mengenai prediksi kelulusan mahasiswa menggunakan algoritma C4.5 dengan studi kasus Universitas Peradaban mengklasifikasi dengan nilai *accuracy* sebesar 88,74% *precision* sebesar 91,79%, dan *recall* 95,34%. Penelitian Yuniarti *et al.* (2020) mengenai identifikasi potensi keberhasilan studi menggunakan *naïve bayes Classification* memberikan akurasi

95,8% sampai dengan 99,41%. Dari ketiga riset tersebut membuktikan bahwa penggunaan algoritma C4.5 dan *Naïve Bayes Classification* memiliki *accuracy*, *precision*, dan *recall* yang tinggi untuk mengklasifikasi mahasiswa berpotensi *drop out*. Oleh karena itu Fakultas Sains dan Teknologi, Universitas Islam Negeri Syarif Hidayatullah Jakarta memerlukan sebuah informasi untuk untuk mengetahui mahasiswanya yang berpotensi *drop out* dalam masa studinya, agar nantinya dapat melakukan evaluasi, baik dari mahasiswa itu sendiri ataupun dari program studi mahasiswa yang terkait.

Dari hasil pemaparan maka penulis ingin mengklasifikasi mahasiswa yang berpotensi *drop out* pada Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri (UIN) Syarif Hidayatullah Jakarta. Oleh karena itu penelitian ini berjudul “Implementasi Algoritma C4.5 dan *Naïve Bayes Classification* untuk Mengklasifikasi Mahasiswa Berpotensi *Drop Out* (Studi Kasus: Prodi Sistem Informasi, Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta)”.

## 1.2 Identifikasi Masalah

Dalam latar belakang tersebut, dapat diidentifikasi beberapa permasalahan sebagai berikut:

- a. Kenaikan persentase mahasiswa Prodi Sistem Informasi UIN Syarif Hidayatullah Jakarta yang berstatus *drop out* pada tahun masuk 2012-2015 sebesar 7,95% dan kenaikan jumlah mahasiswa yang berstatus *drop out* pada tahun 2010-2014 sebesar 32 mahasiswa.

- b. Tidak adanya informasi bagi mahasiswa dan pihak prodi mengenai mahasiswa Prodi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri Syarif Hidayatullah Jakarta yang terancam *drop out* dalam masa studinya.
- c. Banyaknya hasil penelitian yang menunjukkan bahwa kinerja algoritma C4.5 dan *Naïve Bayes Classification* memiliki *accuracy*, *precision*, dan *recall* yang tinggi untuk mengklasifikasi mahasiswa berpotensi *drop out*.

### 1.3 Rumusan Masalah

Berdasarkan identifikasi masalah, rumusan permasalahan dalam penelitian ini yaitu:

- a. Apa saja kriteria yang digunakan untuk mengklasifikasi mahasiswa berpotensi *drop out* untuk mahasiswa Prodi Sistem Informasi, Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta?
- b. Bagaimana menerapkan algoritma C4.5 dan *Naïve Bayes Classification* untuk memprediksi mahasiswa berpotensi *drop out*?
- c. Bagaimana hasil prediksi mahasiswa berpotensi *drop out* di Prodi Sistem Informasi, Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta tahun masuk 2018?
- d. Bagaimana kinerja algoritma C4.5 dan *Naïve Bayes Classification* untuk mengklasifikasi mahasiswa berpotensi *drop out*?

#### 1.4 Batasan Masalah

Berdasarkan masalah yang telah dirumuskan, adapun batasan atau ruang lingkup masalah sebagai berikut:

- a. Data mahasiswa yang digunakan sebagai data *training* menggunakan data tahun masuk mahasiswa 2010-2015 dan data *testing* menggunakan data tahun masuk mahasiswa 2018.
- b. Mengklasifikasikan jumlah, persentase, memvisualkan dengan grafik, kriteria mahasiswa Prodi Sistem Informasi, Fakultas Sains dan Teknologi, UIN Jakarta tahun masuk 2018 yang berpotensi *drop out*.
- c. Pada penelitian dalam hal implementasi dan pengujian peneliti menggunakan *website Google Colab*.

#### 1.5 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

- a. Mengetahui kriteria apa saja yang digunakan oleh Prodi Sistem Informasi UIN Syarif Hidayatullah Jakarta untuk mengklasifikasi mahasiswa berpotensi *drop out*.
- b. Memahami implementasi algoritma C4.5 dan *Naïve Bayes Classification* dalam menyajikan informasi dalam klasifikasi mahasiswa Prodi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri Syarif Hidayatullah Jakarta tahun masuk 2018 yang berpotensi *drop out* dalam masa studinya.

- c. Mendapatkan hasil prediksi dari algoritma C4.5 dan *Naïve Bayes Classification* pada mahasiswa Prodi Sistem Informasi, UIN Syarif Hidayatullah Jakarta tahun masuk 2018.
- d. Mendapatkan kinerja algoritma C4.5 dan *Naïve Bayes Classification* untuk mengklasifikasi mahasiswa berpotensi *drop out* pada Prodi Sistem Informasi UIN Syarif Hidayatullah Jakarta.

## 1.6 Manfaat Penelitian

### a. Manfaat bagi peneliti:

Bagi peneliti, penelitian ini merupakan salah satu cara untuk menambah pengetahuan serta peneliti dapat menerapkan ilmu yang telah didapatkan saat dimasa perkuliahan sebagai salah satu syarat untuk meraih gelar Sarjana.

### b. Manfaat bagi universitas:

- Sebagai bahan referensi untuk penelitian sejenis selanjutnya.
- Sebagai bahan evaluasi untuk mengembangkan keilmuan yang berhubungan dengan penelitian sejenis.
- Hasil dari penelitian ini dapat digunakan sebagai informasi dalam mengukur mahasiswa yang berpotensi *drop out* dalam masa studinya.

## 1.7 Metodologi Penelitian

Metodologi yang digunakan dalam penelitian ini adalah:

### 1.7.1 Metode Pengumpulan Data

#### a. Wawancara

Metode wawancara bertujuan untuk mengetahui masalah apa yang sedang dihadapi dan dibutuhkan apa saja dalam pengukuran ini dengan Sekretaris Program Studi Sistem Informasi UIN Syarif Hidayatullah Jakarta.

b. Observasi

Metode ini dilakukan dengan mengumpulkan data serta informasi yang diperlukan untuk penelitian dengan mengamati langsung sistem yang berjalan.

c. Studi Pustaka

Metode ini dilakukan dengan mencari dan mempelajari teori dan hasil penelitian yang berhubungan dengan C4.5 dan *Naïve Bayes Classification* yang bersumber dari buku, jurnal, *website*, dan prosiding konferensi.

### 1.7.2 Tahapan *Data Mining*

Metodologi penelitian yang digunakan dalam penelitian kali ini adalah SEMMA (*Sample, Explore, Modify, Model, Assess*) (Alizah *et al.*, 2020):

a. *Sample*

Pada proses ini, dimana *data mining* yang digunakan untuk mengumpulkan sampel untuk membentuk informasi yang penting dan signifikan. Data pada penelitian kali ini didapatkan dari bagian Pusat Akademik UIN Syarif Hidayatullah Jakarta.

b. *Explore*

Pada proses ini, *data mining* yang dapat digunakan untuk mencari kumpulan data serta menjadikannya sebuah informasi yang terkait, untuk

mendapatkan pengertian dan ide. Data yang digunakan pada penelitian kali ini adalah data nilai dan status mahasiswa Prodi Sistem Informasi, Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta tahun masuk 2010, 2011, 2012, 2013, 2014, 2015, dan 2018 dan mengobeservasi kriteria-kriteria apa saja yang menjadi penentu mahasiswa berpotensi *drop out*.

c. *Modify*

Pada proses ini, *data mining* yang dapat digunakan untuk memodifikasikan dan memfokuskan proses pemilihan model. Beberapa proses yang dilakukan yaitu: *case folding*, *cleaning*, *tokenize*, *normalize*, *stopword*, dan *stemming*.

d. *Model*

Pada proses ini, *data mining* yang digunakan untuk untuk mencari kombinasi data yang mengklasifikasi hasil terpercaya yang diinginkan secara otomatis. Dilakukan klasifikasi sesuai kategori untuk melihat apakah mahasiswa yang bersangkutan termasuk mahasiswa berpotensi *drop out* atau tidak menggunakan algoritma C4.5 dan *Naïve Bayes Classification*.

e. *Assess*

Proses ini merupakan mengevaluasi hasil kinerja penerapan kedua algoritma yang meliputi tingkat yaitu akurasi, presisi, dan *recall* (*confusion matix*).

## 1.8 Sistematika Penulisan

### BAB 1 PENDAHULUAN

Dalam bab ini dijelaskan secara singkat latar belakang masalah, identifikasi masalah, rumusan masalah, batasan masalah, tujuan dan juga manfaat penelitian, metode penelitian, serta sistematika penulisan.

## **BAB 2 KAJIAN PUSTAKA**

Penulis memaparkan teori apa saja yang dipakai dan mendukung implementasi algoritma C4.5 dan *Naïve Bayes Classification* untuk mengklasifikasi mahasiswa berpotensi *drop out*.

## **BAB 3 METODOLOGI PENELITIAN**

Dalam bab ini penulis memaparkan tentang metode yang digunakan dalam penelitian ini, yaitu: metode pengumpulan data dan tahapan *data mining*.

## **BAB 4 HASIL DAN PEMBAHASAN**

Pada bab ini membahas tahapan *data mining*, hasil prediksi, dan evaluasi kinerja yang diperoleh dari implementasi algoritma C4.5 dan *Naïve Bayes Classification* untuk mahasiswa yang berpotensi *drop out*.

## **BAB 5 PENUTUP**

Bab ini membahas kesimpulan dari uraian yang sudah dijelaskan pada bab-bab sebelumnya dan saran dalam pengembangan lebih lanjut.



## **BAB 2**

### **TINJAUAN PUSTAKA**

#### **2.1 Menjauhi Sifat Lemah dan Malas**

Sebagai umat yang diperintahkan untuk selalu menuntut ilmu, kita diperintahkan untuk menjauhi 2 sifat yaitu sifat lemah dan malas. Seorang muslim yang baik dapat dilihat dari sikapnya yang rajin beribadah dan bekerja. Rasulullah Shallahu Alaihi Wa Salam dalam hidupnya selalu memohon kepada Allah untuk dijauhkan dari sifat lemah dan malas, seperti yang tertera pada Hadist Riwayat Al-Bukhari sebagai berikut:

اللَّهُمَّ إِنِّي أَعُوذُ بِكَ مِنَ الْعَجْزِ وَالْكَسَلِ وَالْجُبْنِ وَالْهَرَمِ وَالْبُخْلِ

“Ya Allah sesungguhnya aku memohon perlindungan kepadaMu dari kelemahan, kemalasan, pengecut, penyakit tua, dan kekikiran”. (HR. Al-Bukhari)

Sifat lemah dan malas adalah dua sifat yang berbeda, sifat lemah dapat diartikan dengan tidak memilikinya kemampuan, lain halnya dengan sifat malas, sifat malas adalah ketidak inginan untuk melakukan sesuatu walaupun mampu melakukannya, namun karena ketidak inginan itu membuatnya tidak melakukan sesuatu.

Sifat lemah dan malas adalah sebuah penyakit yang dapat membuat seseorang tidak bergerak serta meninggalkan kewajiban yang dimilikinya, yang berakibat terbukanya pintu-pintu keburukan, yang bisa kita gambarkan seperti mahasiswa yang sedang menyelesaikan masa kuliah, namun kedua sifat tersebut (lemah dan malas) membuat mahasiswa dikenakan sanksi yaitu dikeluarkan dari perguruan tinggi

(*drop out*) tempat menuntut ilmu, maka sebaiknya sifat yang harus selalu kita terapkan adalah sifat bersungguh-sungguh, seperti yang tertera pada Hadist Riwayat Ahmad sebagai berikut:

أَحْرِصْ عَلَى مَا يَنْفَعُكَ وَاسْتَعِزْ بِاللَّهِ وَلَا تَعْجُزْ

“Bersungguh-sungguhlah terhadap apa-apa yang bermanfaat bagimu, mohonlah pertolongan Allah, dan janganlah lemah”. (HR. Ahmad 9026, Muslim 6945, dan yang lainnya).

Dalam Al-Qur'an surat an-Nisa ayat 71-72, Allah berfirman:

يَا أَيُّهَا الَّذِينَ آمَنُوا خُذُوا حِذْرَكُمْ فَانْفِرُوا ثُبَاتٍ أَوْ انفِرُوا جَمِيعًا ۚ  
وَإِنَّ مِنْكُمْ لَمَنْ لَيُبَطِّئَنَّ فَإِنْ أَصَابَكُمْ مُصِيبَةٌ قَالَ قَدْ أَنْعَمَ اللَّهُ عَلَيَّ إِذْ لَمْ أَكُنْ مَعَهُمْ شَهِيدًا ۚ

“Wahai orang-orang yang beriman! Bersiap-siagalah kamu, dan majulah (ke medan pertempuran) secara berkelompok, atau majulah bersama-sama (serentak). Dan sesungguhnya di antara kamu pasti ada orang yang sangat enggan (ke medan pertempuran). Lalu jika kamu ditimpa musibah dia berkata: Sungguh, Allah telah memberikan nikmat kepadaku karena aku tidak ikut berperang bersama mereka.”

Dari penjelasan ayat diatas dapat diketahui bahwa Allah sangat tidak menyukai sifat malas dan lemah. Sebagai seorang mahasiswa sudah seharusnya kita menjalankan tugas dan kewajiban kita sebagai seorang mahasiswa dengan belajar dan menyelesaikan studi dengan tepat waktu. Sebagian besar mahasiswa di *drop out* dari kampus disebabkan memiliki sifat malas, lemah dan tidak mau berjuang.

Sifat lemah dan malas akan membuat segala hal yang sedang dan akan dilakukan menjadi terhambat. Ayat dan hadits tersebut sangat sesuai dengan penelitian yang diambil penulis terkait *drop out*, sifat lemah dan malas menjadi

salah satu dari sekian banyak faktor yang mempengaruhi seorang mahasiswa di *drop out*.

## 2.2 *Data Mining*

Menurut Rony (2021), *data mining* merupakan proses pengumpulan serta pengolahan informasi yang bertujuan untuk mengekstrak informasi bernilai pada data. Proses pengumpulan serta ekstraksi data tersebut dapat dicoba memanfaatkan *software* dengan dukungan perhitungan statistika, matematika, maupun teknologi *Artificial Intelligence* (AI). Selain itu, *data mining* kerap disebut juga *Knowledge Discovery in Database* (KDD).

Secara umum, ada sebagian prosedur yang digunakan untuk melakukan *data mining*. Berikut ini merupakan metodenya:

a. *Association*

Teknik yang pertama adalah *association*. *Association* merupakan suatu metode atau cara berbasis aturan yang kerap digunakan untuk mendapatkan asosiasi dan hubungan *variabel* dalam satu *set* data. Biasanya analisis ini terdiri atas pernyataan “*if* atau *then*” sederhana.

b. *Classification*

Selanjutnya *classification* adalah metode yang paling sering digunakan dalam proses *data mining*. *Classification* adalah proses untuk mengklasifikasi suatu kelas dalam suatu objek penelitian.

c. *Regression*

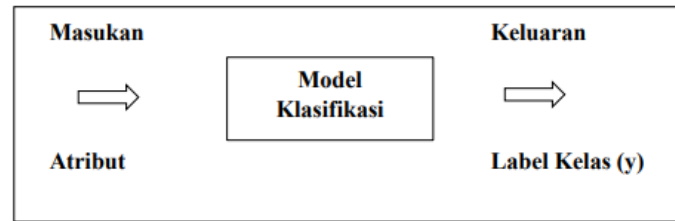
*Regression* adalah suatu cara atau teknik yang menjelaskan variabel *dependent* melalui proses analisis variabel *independen*. Sebagai contoh, mengklasifikasi penjualan suatu produk berdasarkan hubungan antara harga produk dengan tingkat pendapatan rata-rata pelanggan.

d. *Clustering*

*Clustering* digunakan dalam membagi kumpulan data menjadi beberapa kelompok berdasarkan kemiripan atribut yang dimiliki. Contoh kasusnya adalah *Customer Segmentation*, di mana pelanggan dibagi ke dalam beberapa grup berdasarkan tingkat kemiripannya.

### 2.3 Klasifikasi

Salah satu metode algoritma yang umum digunakan pada konsep *data mining* untuk mencari solusi dari permasalahan yang sering muncul adalah metode klasifikasi. Klasifikasi merupakan suatu proses untuk menemukan model atau fungsi yang dapat data untuk nantinya dimasukkan ke dalam kelas tertentu dari sejumlah kelas yang tersedia (Budiman *et al.*, 2015). Tan *et al.* (2006) klasifikasi ialah sebuah proses untuk menemukan model yang menjelaskan atau membedakan konsep atau kelas data. Klasifikasi bertujuan untuk memperkirakan kelas dari suatu objek yang kelasnya tidak diketahui (Larose & Larose, 2014). Selain itu, klasifikasi adalah proses pengkategorian yang dilakukan terhadap sekumpulan dokumen, dimana klasifikasi ini sangat penting untuk kemudahan pengguna dalam mencari dokumen (Indriani, 2014). Secara umum, konsep klasifikasi dapat dilihat pada Gambar 2.1.



**Gambar 2.1** Konsep Klasifikasi

Di dalam klasifikasi akan diberikan sejumlah *record* yang dikenal dengan *training set*. *Training set* terdiri atas beberapa atribut, atribut ini dapat berupa kontinyu ataupun kategoris, atribut inilah yang nantinya menunjukkan kelas untuk *record*. Ada dua jenis model klasifikasi, yakni pemodelan deskriptif (*descriptive modelling*) dan pemodelan prediktif (*predictive modelling*). *Descriptive modelling* merupakan model klasifikasi yang dapat berfungsi sebagai alat penjelasan untuk membedakan objek-objek yang ada dalam kelas-kelas yang berbeda. Sedangkan, *predictive modelling* merupakan model klasifikasi yang dapat digunakan untuk mengklasifikasi label kelas *record* yang tidak diketahui (Larose & Larose, 2014).

Menurut Gorunescu, proses klasifikasi mempunyai empat komponen dasar, yakni (Gorunescu, 2011):

1. Kelas

*Class*/kelas ialah variabel dependen dari model yang merupakan variabel kategorikal “label” yang diletakkan pada objek setelah klasifikasinya. Sebagai contoh adanya kelas bintang, kelas gempa bumi, loyalitas pelanggan, dan sebagainya.

2. Prediktor

*Predictors*/prediktor ialah variabel dependen dari model yang diwakili oleh karakteristik (atribut) dari data yang akan diklasifikasikan dan berdasarkan pada klasifikasi yang dibuat. Contohnya adalah konsumsi alkohol, catatan geologi spesifik, musim, arah angin dan kecepatan, merokok, lokasi terjadi fenomena dan lainnya.

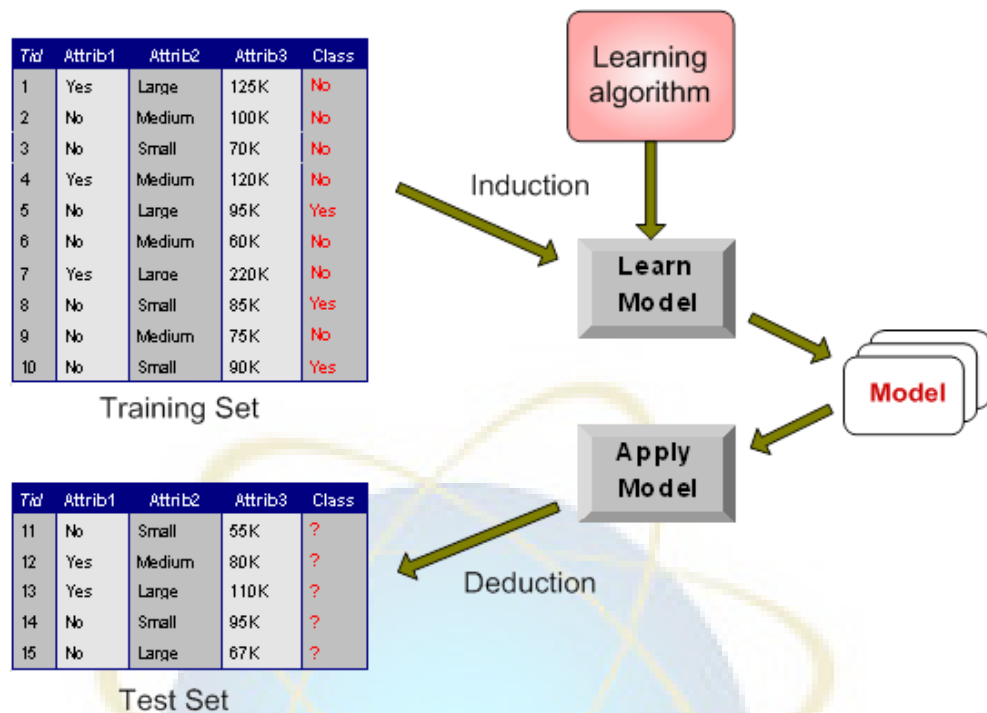
### 3. Pelatihan Dataset

*Training* dataset/pelatihan dataset adalah kumpulan data yang berisi nilai untuk dua komponen sebelumnya, yang digunakan untuk “melatih” model sehingga dapat mengenali kelas yang sesuai berdasarkan prediktor yang tersedia. Misalnya, set tersebut adalah kelompok pelanggan supermarket, *database* tentang badai (<https://opendata.jabarprov.go.id/id/dataset/jumlah-kejadian-bencana-angin-topanangin-puting-beliung-berdasarkan-kabupatenkota-di-jawa-barat>), *database* kecelakaan lalu lintas menurut profesi (<https://data.jakarta.go.id/dataset/data-jumlah-korban-kecelakaan-lalu-lintas-menurut-profesi/resource/9968fb00-5119-441d-97e7-4007663b508c>), dan lain sebagainya.

### 4. Pengujian Dataset

*Testing* dataset/pengujian dataset adalah data baru yang akan diklasifikasikan oleh model (pengklasifikasi) yang dibangun, dan akurasi klasifikasi (kinerja model) dapat dievaluasi.

Dari penjelasan pada point sebelumnya, Gambar 2.2 adalah contoh klasifikasi menurut Gorunescu (2011).



**Gambar 2.2** Contoh Klasifikasi menurut (Gorunescu, 2011)

## 2.4 SEMMA

*Sample, Explore, Modify, Model, Assess* atau yang biasa disingkat SEMMA adalah suatu proses dalam melakukan sebuah proses *data mining*. Pada umumnya, penerapannya SAS Institute membagi siklus SEMMA menjadi 5 (lima) tahapan untuk proses *data mining* (Wilson, 2021). Mengutip dari (Azevedo & Santos, 2008) SAS Institute mendefinisikan lima proses tahapan pada SEMMA yaitu: *sample, explore, modify, model, assess*.

### a. Sample

Pada proses ini, dimana *data mining* yang digunakan untuk mengumpulkan sampel untuk membentuk informasi yang penting dan signifikan. Data pada

penelitian kali ini didapatkan dari bagian Pusat Akademik UIN Syarif Hidayatullah Jakarta.

*b. Explore*

Pada proses ini, *data mining* yang dapat digunakan untuk mencari kumpulan data serta menjadikannya sebuah informasi yang terkait, untuk mendapatkan pengertian dan ide. Data yang digunakan pada penelitian kali ini adalah data nilai dan status mahasiswa Prodi Sistem Informasi, Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta tahun masuk 2010, 2011, 2012, 2013, 2014, 2015, dan 2018.

*c. Modify*

Pada proses ini, *data mining* yang dapat digunakan untuk memodifikasikan untuk memfokuskan proses pemilihan model. Beberapa proses yang dilakukan yaitu: *case folding*, *cleaning*, *tokenize*, *normalize*, *stopword*, dan *stemming*.

*d. Model*

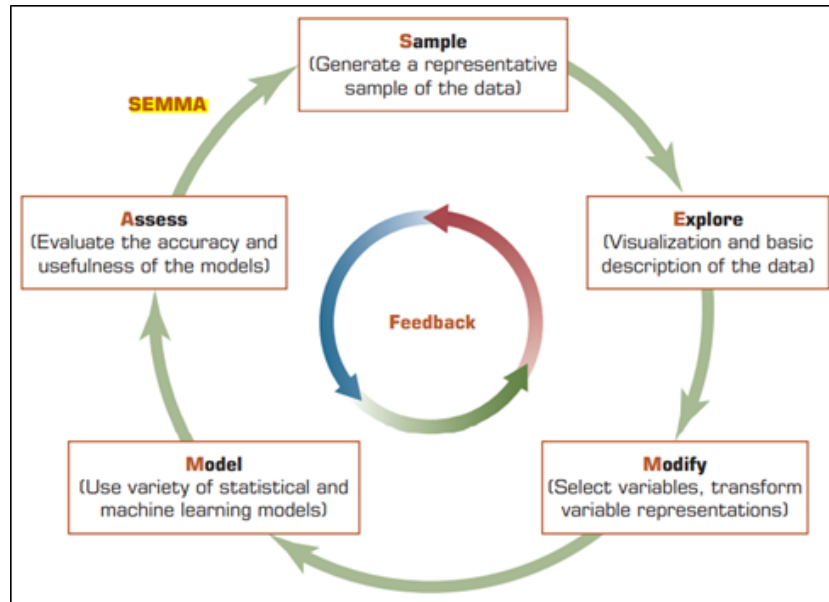
Pada proses ini *data mining* yang digunakan untuk mencari kombinasi data yang mengklasifikasi hasil terpercaya yang diinginkan secara otomatis. Dilakukan klasifikasi sesuai kategori untuk melihat apakah mahasiswa yang bersangkutan termasuk mahasiswa berpotensi *drop out* atau tidak menggunakan algoritma C4.5 dan *Naïve Bayes Classification*.

*e. Assess*

Pada proses ini *data mining* yang digunakan untuk mengevaluasi kegunaan dan keandalan penemuan dari data proses *data mining* dan memperkirakan



seberapa baik kinerja tersebut. Nilai yang dihasilkan adalah *confusion matrix* yaitu akurasi, presisi, dan *recall*.



**Gambar 2.3** SEMMA Data Mining Process menurut (Wilson, 2021)

## 2.5 Split Validation

*Split Validation* adalah teknik validasi yang melibatkan pembagian dataset menjadi dua bagian secara acak, di mana satu bagian digunakan untuk data *training* dan bagian lainnya untuk data *testing*. Dalam *split validation*, eksperimen pelatihan dilakukan berdasarkan perbandingan pembagian yang telah ditentukan sebelumnya, dan bagian sisanya dari perbandingan pembagian dianggap sebagai data pengujian. Data *training* digunakan untuk proses pembelajaran, sementara data *testing* belum pernah digunakan untuk pembelajaran dan berfungsi untuk mengevaluasi akurasi atau kebenaran hasil pembelajaran (Suherma & Muzaky, 2019).

Dengan menggunakan *split validation* klasifikasi diperoleh dengan menggunakan teknik *systematic random sampling*, yaitu dengan membagi ukuran

populasi dengan ukuran sampel yang dikehendaki. Penentuan unsur selanjutnya ditempuh dengan cara menggunakan *interval* sampel (Manullang *et al.*, 2021).

## 2.6. Confusion Matrix

*Confusion matrix* adalah tabel yang menyatakan klasifikasi jumlah data uji yang benar dan jumlah data uji yang salah (Dwi & Surya, 2021). *Confusion Matrix* adalah pengukuran performa untuk masalah klasifikasi machine learning dimana keluaran dapat berupa dua kelas atau lebih. *Confusion Matrix* adalah tabel dengan 4 kombinasi berbeda dari nilai prediksi dan nilai aktual. Ada empat istilah yang merupakan representasi hasil proses klasifikasi pada confusion matrix yaitu *TRUE POSITIVE*, *TRUE NEGATIVE*, *FALSE POSITIVE*, DAN *FALSE NEGATIVE* (Anggreany, 2020).

**Tabel 2.1** Bentuk *Confusion Matrix*

<i>Confusion Matrix</i>		Nilai Sebenarnya	
		<i>TRUE</i>	<i>FALSE</i>
Nilai Prediksi	<i>TRUE</i>	TP ( <i>True Positive Correct result</i> )	FP ( <i>False Positive Unexpected result</i> )
	<i>FALSE</i>	FN ( <i>False Negative Missing result</i> )	TN ( <i>True Negative Correct absence of result</i> )

Pengujian keakuratan hasil pencarian akan dievaluasi dengan dengan nilai *recall*, *precision*, *accuracy*, dan lainnya (Melita *et al.*, 2018). Tabel 2.1 merupakan table dari bentuk *Confusion Matrix*. Nilai TP (*true positive*) dan TN (*true negative*) menunjukkan tingkat ketepatan dari klasifikasi. Semakin tinggi nilai TP dan TN maka semakin baik pula tingkat klasifikasi dari akurasi, presisi dan *recall*. Jika label prediksi keluaran bernilai benar (*true*) dan nilai sebenarnya bernilai salah (*false*)

disebut sebagai *false positive* (FP). Sedangkan, jika nilai prediksi label keluaran bernilai salah (*false*) dan nilai sebenarnya bernilai benar (*true*) maka disebut sebagai *false negative* (FN) (Arifin & Sasongko, 2018). Secara matematis untuk menghitung nilai akurasi, presisi dan *recall* dapat dilihat pada persamaan (2.1)-(2.3) berikut:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (2.1)$$

$$Presisi = \frac{TP}{TP + FP} \times 100\% \quad (2.2)$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (2.3)$$

Keterangan:

TP : *True positive* adalah jumlah data positif yang terklasifikasi dengan benar oleh sistem

TN : *True negative* adalah jumlah data negative yang terklasifikasikan dengan benar oleh sistem

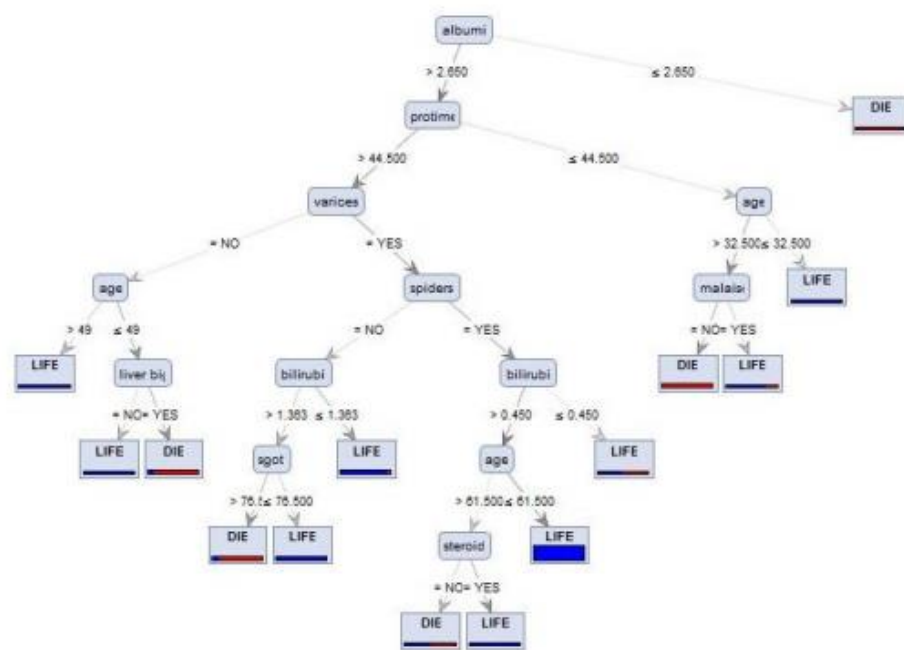
FN : *False negative* adalah jumlah data negative yang terklasifikasikan salah oleh sistem

FP : *False Positive* adalah jumlah data positif yang terklasifikasi salah oleh sistem.

## 2.7 Algoritma

### 2.7.1 C4.5

*Decision tree* merupakan model prediksi yang secara rekursif membagi ruang kovariat ke dalam ruang bagian sedemikian sehingga setiap ruang bagian membentuk dasar untuk fungsi prediksi yang berbeda. Selain itu, *decision tree* dapat digunakan pada berbagai tugas belajar seperti klasifikasi, regresi, bahkan analisis survival. *Decision tree* telah menjadi salah satu pendekatan yang populer di dalam ilmu data. Selain itu, *decision tree* menyerupai struktur *flowchart* dimana masing-masing internal node-nya dinyatakan sebagai atribut pengujian.



**Gambar 2.4** Contoh Pohon Keputusan

C4.5 *Decision Tree*, yang dihasilkan dari algoritma ID3, C4.5 adalah algoritma yang paling sering digunakan serta berpengaruh saat ini. Dibandingkan dengan ID3, ada beberapa peningkatan yang terdapat pada Algoritma C4.5.

Pertama, model dengan algoritma C4.5 mengambil rasio perolehan informasi sebagai kriteria pemilihan atribut sedangkan model dengan algoritma ID3 menggunakan *information gain subtree*. Kedua, ketika membangun pohon keputusan, untuk menghindari *overfitting*, pemangkasan dapat dilakukan. Ketiga, data yang tidak lengkap dan data diskrit dapat ditangani oleh Pohon Keputusan C4.5 (Wang *et al.*, 2019). Gambar 2.4 merupakan contoh pohon keputusan hasil algoritma C4.5.

Pada dasarnya, algoritma pohon keputusan C4.5 menerapkan konsep entropi informasi untuk membangun pohon keputusan melalui data pembelajaran yang ada. Dengan cara memilih atribut dengan tingkat perolehan informasi tertinggi di set sampel saat ini sebagai atribut pengujian untuk membagi set sampel. Sebagai untuk mekanisme optimasi, *bootstrap* agregasi (*bagging*) prosedur bertujuan untuk meningkatkan akurasi prediksi yang dilaporkan oleh satu pohon klasifikasi, diketahui bahwa pengklasifikasi agregat dapat secara efektif meningkatkan akurasi prediksi. Sementara itu, faktor kunci untuk peningkatan akurasi adalah kemungkinan ketidakstabilan prediksi metode, yaitu apakah perubahan kecil dalam perangkat pembelajaran menghasilkan perubahan pada prediktor. Juga, prosedur yang tidak stabil cenderung mendapat manfaat dari agregasi, dan pohon klasifikasi adalah pengklasifikasi yang tidak stabil (Lee *et al.*, 2018).

Hal pertama yang dilakukan untuk membentuk pohon keputusan adalah menentukan variabel mana yang menjadi akar dari pohon keputusan tersebut. Cara menentukan variabel yang menjadi akar adalah dengan menggunakan *entropy*, *gain*, *split info*, dan *gain ratio* (Mubarok, 2018).

i. *Entropy*

*Entropy* adalah suatu parameter untuk mengukur tingkat keberagaman (heterogenitas) dari kumpulan data. Jika nilai dari *entropy* semakin besar, maka tingkat keberagaman suatu kumpulan data semakin besar. Rumus untuk menghitung *entropy* sebagai berikut:

$$\text{Entropy}(S) = \sum_{i=1}^m p_i \log_2(p_i) \quad (2.4)$$

dimana:

$m$  = jumlah kelas klasifikasi

$p_i$  = jumlah proporsi sampel (peluang) untuk kelas  $i$

Sedangkan rumus untuk *entropy* pada masing-masing variabel adalah:

$$\text{Entropy}_A(S) = \sum_v \frac{|S_v|}{|S|} \text{Entropy}(S_v) \quad (2.5)$$

dimana:

$A$  = variabel

$V$  = nilai yang mungkin untuk variabel  $A$

$|S_v|$  = jumlah sample untuk nilai  $v$

$|S|$  = jumlah untuk seluruh sample data

$\text{Entropy}(S_v)$  = Entropy untuk sampel yang memiliki nilai  $v$

ii. *Gain*

*Gain* adalah ukuran efektifitas suatu variabel dalam mengklasifikasikan data.

*Gain* dari suatu variabel merupakan selisih antara nilai *entropy* total dengan *entropy* dari variabel tersebut. Secara matematis *Gain* dapat ditulis sebagai berikut:

$$Gain(A) = Entropy(S) - Entropy_A(S) \quad (2.6)$$

Pada algoritma C4.5, nilai gain digunakan untuk menentukan variabel mana yang menjadi *node* dari suatu pohon keputusan. Suatu variabel yang memiliki gain tertinggi akan dijadikan *node* di pohon keputusan.

iii. *Split Info*

*Split info* digunakan sebagai pembagi dari  $Gain(A)$  yang akan menghasilkan *Gain Ratio*.

$$SplitInfo_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \cdot \log_2 \left( \frac{|D_j|}{|D|} \right) \quad (2.7)$$

iv. *Gain Ratio*

*Gain Ratio* merupakan salah satu ukuran lain yang digunakan untuk mengatasi masalah pada atribut yang memiliki nilai sangat bervariasi. *Gain Ratio* tertinggi dipilih sebagai atribut *test* untuk simpul.

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)} \quad (2.8)$$

Algoritma C4.5 merupakan algoritma yang digunakan dalam pembangunan pohon keputusan untuk klasifikasi. Dalam konteks mengklasifikasi mahasiswa berpotensi *drop out*, langkah-langkah algoritma C4.5 dapat dijelaskan sebagai berikut:

1. **Pembentukan Pohon Keputusan:** Algoritma C4.5 memulai dengan pembentukan pohon keputusan dari data *training*. Pohon keputusan ini akan menjadi model untuk mengklasifikasikan mahasiswa berpotensi *dropout*.

2. **Pemilihan Atribut:** Algoritma C4.5 menggunakan metode *information gain* untuk memilih atribut terbaik yang akan digunakan untuk membagi data. Atribut yang dipilih adalah atribut yang memberikan informasi terbesar dalam mengklasifikasikan data.
3. **Pembagian Data:** Setelah atribut terbaik dipilih, data latih dibagi berdasarkan nilai atribut tersebut. Proses ini dilakukan secara rekursif untuk setiap cabang dalam pohon keputusan.
4. **Penanganan Atribut yang Hilang:** Algoritma C4.5 dapat menangani nilai atribut yang hilang dengan cara mengabaikan data yang memiliki nilai atribut yang hilang saat melakukan perhitungan *information gain*.
5. **Menghentikan Pohon:** Pohon keputusan berhenti tumbuh ketika salah satu kondisi berikut terpenuhi: semua data pada cabang memiliki kelas yang sama, tidak ada atribut lagi yang dapat dipilih, atau telah mencapai batasan kedalaman pohon yang ditentukan.
6. **Pruning (Pemangkasan):** Setelah pohon keputusan terbentuk, algoritma C4.5 melakukan pruning untuk mengurangi kompleksitas pohon dan mencegah *overfitting*. Ini dilakukan dengan menghapus cabang-cabang yang tidak signifikan.
7. **Klasifikasi:** Setelah pohon keputusan selesai dibangun, algoritma C4.5 dapat digunakan untuk mengklasifikasikan mahasiswa baru berdasarkan atribut-atribut yang ada dalam data mereka.



Dengan langkah-langkah ini, algoritma C4.5 dapat membantu dalam mengidentifikasi mahasiswa yang berpotensi drop out berdasarkan pola atribut yang ada dalam data mahasiswa.

### 2.7.2 *Naïve Bayes Classification*

*Naïve Bayes Classification* atau yang bisa disebut dengan NBC adalah metode pengklasifikasian probabilitas sederhana berdasarkan pada teorema *bayes*. Teorema *bayes* dikombinasikan dengan “*Naïve*” yang berarti setiap atribut atau *variable* bersifat bebas (*independent*) (Arifin & Sasongko, 2018). *Naïve Bayes* merupakan turunan dari konsep *teorema Bayes*. Adapun kelebihan dari *Naïve Bayes* adalah sederhana, cepat, dan berakurasi tinggi, Algoritma *Naïve Bayes* merupakan suatu algoritma klasifikasi pada *data mining* yang memanfaatkan kemungkinan dan stasistika sederhana yang ditemukan oleh ilmuwan Inggris yaitu Thomas Bayes (Muhathir & Santoso, 2020).

*Naïve Bayes* adalah salah satu teknik klasifikasi dengan memprediksi probabilitas berdasarkan data masa sebelumnya yang sudah ada (Febrianti, 2020). Selain itu, metode ini dapat digunakan untuk klasifikasi data yang bersifat kuantitatif dan kualitatif, tidak memerlukan jumlah data yang banyak, bisa digunakan untuk klasifikasi untuk dua kelas atau lebih (*multiclass*), nilai yang hilang bisa diabaikan dalam perhitungan (Widianto, 2019).

Kekurangan dari *Naïve Bayes* yaitu atribut dari suatu data adalah independen dan tidak memiliki keterkaitan satu sama lain (Fairuz, Ramadhani & Tanjung, 2021). Metode ini juga memiliki kekurangan dimana variabel independen sehingga

tidak memperhitungkan korelasi antar variabel dimana pada realitanya terkadang ada korelasi antar variabel. Selain itu, *data train* sangat berpengaruh karena dalam menghasilkan klasifikasi, metode ini bekerja sesuai dengan pengetahuan awal. Metode *Naïve Bayes* tidak bisa digunakan untuk mendeteksi gambar, hanya klasifikasi data berupa teks dan numerik (Widianto, 2019). Secara matematis rumus *Naïve Bayes* adalah sebagai berikut:

$$P(C|X) = \frac{P(X_1|C)P(X_2|C) \dots P(X_n|C)P(C)}{P(X_1)P(X_2) \dots P(X_n)} \quad (2.9)$$

Rumus perhitungan *Naïve Bayes* dimana  $X$  adalah data dengan kelas yang belum diketahui dan  $C$  merupakan suatu kelas pada dataset. *Posterior* yaitu probabilitas data  $X$  pada kelas  $C$  atau  $P(C/X)$  adalah hasil perkalian dari *likelihood* dan *prior* dibagi *evidence*. *Likelihood* yaitu probabilitas atribut data  $X$  pada kelas  $C$  atau  $P(x|C)$ , *prior* yaitu probabilitas kelas  $C$  dari total data set atau  $P(C)$ , dan *evidence* yaitu probabilitas atribut data  $X$  dari seluruh total data set atau  $P(x)$ .

Beberapa penelitian terkait penggunaan *data mining* telah dilakukan untuk menggali serta memperoleh model dari data yang telah diproses dan diukur tingkat keakuratannya. Salah satunya adalah penelitian yang menerapkan dua metode yakni *Naïve Bayes* dan optimasi *Particle Swarm Optimization* (PSO) sebagai penentu atribut mana saja yang akan dihilangkan (Sudriyanto *et al.*, 2021), dalam mengklasifikasi tingkat kelulusan mahasiswa tepat waktu. Hasil penelitian menunjukkan akurasi dalam mengklasifikasi kelulusan mahasiswa sebesar 89,46%

dimana *class precision* untuk *precision* terlambat adalah sebesar 89,68%; *precision* mahasiswa yang lulus tepat waktu adalah 89,29% dengan prediksi tepat dan *true* tepat sebanyak 200 data. Data hasil hitung akurasi dapat dilihat pada Tabel 2.1.

**Tabel 2.2** Hasil Hitungan Akurasi

	<i>true</i> .TERLAMBAT	<i>true</i> .TEPAT	<i>class precision</i>
<i>pred</i> .TERLAMBAT	139	16	89.68%
<i>pred</i> .TEPAT	24	200	89.29%
<i>class recall</i>	85.28%	92.59%	

Accuracy: 89.46% +/- 3.09% (micro average:89.45%)

Wenty *et al.* (2020) melakukan penelitian dengan menggunakan metode *Naïve Bayes*, dimana hasil penelitian menunjukkan bahwa klasifikasi dengan menggunakan metode *Naïve Bayes* untuk variabel input dengan target IPs 2/4 memperoleh akurasi sebesar 99,41% (Tabel 2.2) (Yuniarti *et al.*, 2020). Hal ini menunjukkan bahwa metode *Naïve Bayes* dapat menunjukkan akurasinya dalam menentukan potensi keberhasilan studi dengan variabel ada.

**Tabel 2.3** Hasil Akurasi

Proporsi Data	IPs 2/4	DO/AKTIF	IP-Matkul Prodi	IP-Matkul Non Prodi
80% : 20%	99.410	96.671	95.871	97.89

Penelitian lainnya, adalah penelitian yang dilakukan oleh Haditsah pada tahun 2018, dimana dilakukan penelitian mengenai klasifikasi masyarakat miskin menggunakan metode *Naïve Bayes* di wilayah pemerintahan Kecamatan Tibawa Kab. Gorontalo. Atribut yang digunakan dalam penelitian tersebut adalah umur, pekerjaan, pendidikan, penghasilan, tanggungan serta status. Hasil penelitian menunjukkan bahwa, penggunaan metode *Naïve Bayes* terhadap dataset yang telah

diambil pada objek penelitian diperoleh tingkat akurasi sebesar 73% termasuk dalam kategori *good*. Sementara nilai *Precision* sebesar 92% dan *Recall* sebesar 86% (Annur, 2018).

Berangkat dari beberapa penelitian yang telah dilakukan diatas dengan tingkat akurasi yang cukup baik. Metode *Naïve Bayes* cocok digunakan untuk mengklasifikasi mahasiswa yang berpotensi *drop out*. Selain itu, metode lain yang digunakan dalam penelitian ini adalah metode algoritma C4.5.

Algoritma *Naive Bayes Classification* adalah metode klasifikasi yang berbasis pada teorema Bayes dengan asumsi bahwa setiap atribut adalah independen satu sama lain. Dalam konteks mengklasifikasi mahasiswa berpotensi drop out, langkah-langkah algoritma Naive Bayes Classification dapat dijelaskan sebagai berikut:

1. Pemahaman Probabilitas: Algoritma *Naive Bayes Classification* menggunakan pemahaman probabilitas untuk menentukan kelas mana yang paling mungkin untuk setiap instance data. Ini didasarkan pada teorema Bayes, yang menggambarkan hubungan antara probabilitas *posterior*, probabilitas *prior*, dan *likelihood*.
2. Perhitungan Probabilitas Kelas *Prior*: Langkah pertama adalah menghitung probabilitas *prior* untuk setiap kelas, yaitu probabilitas bahwa seorang mahasiswa akan *drop out* atau tidak, berdasarkan data *training*.
3. Perhitungan Probabilitas *Likelihood*: Selanjutnya, algoritma menghitung probabilitas *likelihood* dari setiap atribut untuk setiap kelas. Ini melibatkan menghitung seberapa sering nilai atribut muncul dalam setiap kelas.

4. Perhitungan Probabilitas *Posterior*: Probabilitas *posterior* adalah probabilitas bahwa seorang mahasiswa termasuk dalam suatu kelas tertentu setelah melihat nilai atributnya. Ini dihitung dengan menggabungkan probabilitas *prior* dan *likelihood* menggunakan teorema Bayes.
5. Prediksi Kelas: Setelah probabilitas *posterior* dihitung untuk setiap kelas, algoritma memprediksi kelas yang paling mungkin untuk setiap *instance* data. Mahasiswa akan diklasifikasikan sebagai berpotensi *drop out* atau tidak berdasarkan probabilitas tertinggi.
6. Evaluasi Model: Langkah terakhir melibatkan evaluasi kinerja model *Naive Bayes Classification* menggunakan data uji atau validasi. Ini dilakukan dengan membandingkan prediksi model dengan kelas sebenarnya dari data uji untuk mengukur akurasi dan kinerja keseluruhan model.

Dengan langkah-langkah ini, algoritma Naive Bayes Classification dapat membantu dalam mengklasifikasikan mahasiswa berpotensi drop out dengan memanfaatkan informasi probabilitas dari atribut-atribut yang relevan dalam data mahasiswa.

## 2.8 Python

*Python* merupakan salah satu bahasa pemrograman yang banyak digunakan oleh perusahaan besar maupun para *developer* untuk mengembangkan berbagai macam aplikasi berbasis *desktop*, *web*, dan *mobile* (Romzi & Budi, 2020). Selain itu, *Python* adalah salah satu dari bahasa pemrograman tingkat tinggi. *Python* terkenal pada kalangan *programmer* karena penggunaannya yang lebih sederhana

dari bahasa pemrograman lainnya. Selain itu, *Python* memiliki struktur sintak yang rapi dan mudah dipahami oleh *programmer* (Wiratmaja *et al.*, 2021).

Berikut adalah contoh penulisan kode menggunakan *python* untuk menampilkan kata “Halo, dunia!”.

```
print("Halo, dunia!")
```

## 2.9 Library

### 2.9.1 NumPy

*Numpy* merupakan *library* yang akan digunakan untuk kebutuhan *scientific* dan matematis (Endang & Rully, 2020). Penggunaan *library* lain pada pengolahan citra digital ini adalah *Numerical Python (Numpy)*. *Numpy* ialah *library* pada bahasa pemrograman *python* yang fokus pada *scientific computing*. Kelebihan dari *library* ini adalah konsumsi memori yang lebih kecil serta *runtime* yang lebih cepat dibandingkan dengan fungsi list pada pemrograman *python*. Penggunaannya ialah mengolah seluruh *pixel* untuk diambil nilainya dengan resolusi kamera endoskop yakni 640x480 *pixel*. Apabila tidak menggunakan *library Numpy* ini mengolah seluruh *pixel* perlu satu persatu, akan sangat memakan waktu dan memberi beban berat pada kinerja kontroler (Nu'man *et al.*, 2020).



**Gambar 2.5** Logo library Numpy

### 2.9.2 *Pandas*

*Pandas Library* merupakan *library* yang digunakan untuk membaca berbagai format data seperti *file*, *.txt*, *.csv*, atau lainnya (Miftah, 2021). *Pandas* merupakan *library python* yang memiliki kemampuan pengolahan data dalam bentuk tabel (baris-kolom) dan kalkulasi statistik. *Pandas* dibuat untuk memenuhi kebutuhan akademis dan industri data *science* dalam *preprocessing* data menggunakan *python*.



**Gambar 2.6** Logo library *Pandas*

### 2.9.3 *Matplotlib*

*Matplotlib* adalah sebuah pustaka Python yang dapat dijalankan secara lintas-*platform* dan digunakan untuk menghasilkan grafik 2 dimensi berkualitas tinggi.

Pustaka ini dapat diintegrasikan dalam skrip Python, *Interpreter* Python, *iPython*, *server*, dan enam toolkit GUI. Dengan *Matplotlib*, pembuatan plot, histogram,

spektra, diagram batang, *errorchart*, serta scatterplot menjadi lebih mudah.

Kemudahan yang ditawarkan oleh *Matplotlib* memungkinkan pengembangan aplikasi laporan profesional, analisis interaktif, dan *dashboard* yang komprehensif,

baik sebagai bagian dari web atau aplikasi GUI (Antarmuka Pengguna Grafis)

(Defrianto & Firmansyah, 2019).



**Gambar 2.7** Logo Library Matplotlib

#### 2.9.4 Seaborn

Seaborn merupakan sebuah *library* yang dibuat berdasarkan *Matplotlib*, dirancang secara khusus untuk visualisasi data statistik. Seaborn menawarkan beragam fungsi dan estetika yang lebih canggih untuk menciptakan visualisasi data yang menarik serta informative (Kelly Hermanto *et al.*, 2023).



**Gambar 2.8** Logo Library Seaborn

#### 2.10 Mahasiswa

Mahasiswa adalah parameter penting dalam penyelenggaraan program studi yang berkaitan dengan prestasi, kompetensi, dan presensi mahasiswa yang seharusnya mendapatkan perhatian lebih serius dalam evaluasi kinerja mahasiswa (Fatma Ayu Rahman *et al.*, 2020). Mahasiswa dapat didefinisikan sebagai individu yang sedang menuntut ilmu ditingkat perguruan tinggi, baik negeri maupun swasta



atau Lembaga lain yang setingkat dengan perguruan tinggi (Hulukati & Djibran, 2018).

Mahasiswa memiliki batas waktu untuk menyelesaikan studi yang mereka lakukan, jika melewati masa waktu studi yang sudah ditentukan, maka pihak perguruan tinggi dapat melakukan *drop out*. Dilihat dari kondisi saat ini, potensi *drop out* semakin meningkat mengingat batasan masa kuliah yang tidak boleh melebihi 14 semester atau selama 7 tahun. Hal inilah yang membuat banyak mahasiswa yang terancam *drop out*. Data status kelulusan mahasiswa sistem informasi pada Tabel 1.1 menunjukkan presentase *drop out* yang cukup tinggi.

Pada penelitian ini, akan dibahas implementasi dari algoritma C4.5 dan *Naïve Bayes Classification* untuk mengklasifikasi mahasiswa berpotensi *drop out*. Mahasiswa yang dimaksud dalam penelitian ini adalah mahasiswa Prodi Sistem Informasi, UIN Syarif Hidayatullah Jakarta dengan tahun masuk 2010, 2011, 2012, 2013, 2014, 2015, dan 2018.

### **2.11 Drop Out**

*Drop out* merupakan salah satu bentuk dari kegagalan mahasiswa dalam mengikuti proses pendidikan pada perguruan tinggi (Fakhriza & Heri, 2018). *Drop out* merupakan salah satu usaha untuk mencegah terjadinya *overfitting* dan juga mempercepat proses *learning* (Santoso & Ariyanto, 2007). Universitas Peradaban Bumiayu melakukan penelitian dan menemukan masalah mahasiswa tidak lulus tepat waktu dan bahkan di *drop out* dikarenakan banyak mahasiswa mengikuti kegiatan kampus, sibuk bekerja, bermasalah dengan kampus, menjadi aktivis, bahkan anggapan bahwa kuliah hanya sekedar untuk mendapatkan ijazah saja,

membuat banyak mahasiswa yang akhirnya tidak lulus tepat waktu bahkan di *drop out* dari kampus (Rahman *et al.*, 2020).

Dalam penelitian yang dilakukan oleh (Samuel *et al.*, 2021) diperoleh 17 aturan yang dapat digunakan untuk menentukan mahasiswa yang berpotensi *drop out*. Selain itu, Alizah *et al.* (2020) juga melakukan penelitian dengan menggunakan metode algoritma C4.5 untuk mengklasifikasikan mahasiswa-mahasiswa yang berpotensi *drop out*. Penelitian tersebut menghasilkan 9 aturan untuk klasifikasi potensi mahasiswa *drop out*.

Sinaga *et al.* (2021) melakukan penelitian dengan hasil penelitian menunjukkan metode klasifikasi dengan konsep algoritma *Decision Tree* C4.5 menghasilkan akurasi sebesar 90,00%. Hasil tersebut diharapkan dapat meningkatkan keinginan lembaga Universitas atau Perguruan Tinggi untuk memberikan respond, pikiran yang baik, pandangan, dan kebijakam baru bagi mahasiswa yang memiliki permasalahan dalam perkuliahan, dengan kata lain membantu memaksimalkan mahasiswa dalam upaya peningkatan presentase minat kuliah (Sinaga *et al.*, 2021).

Selain itu, sebagian besar masalah akademik yang terjadi di lingkungan perguruan tinggi mengakibatkan mahasiswa mengundurkan diri dan *drop out*.

Penentuan *drop out* dari sebuah lembaga perguruan tinggi bukanlah hal yang mudah dilakukan, hal ini dikarenakan harus melihat dari berbagai variabel dan kriteria akademis yang telah ditetapkan oleh perguruan tinggi secara jelas dan diketahui oleh mahasiswa sejak awal menjadi mahasiswa. Dari Tabel 1.1 Data status mahasiswa jurusan sistem informasi menunjukkan presentase *drop out* yang cukup

tinggi. Oleh karena itu, diperlukan analisis dan penelitian untuk mengetahui klasifikasi mahasiswa jurusan sistem informasi yang berpotensi *drop out*.

## 2.12 Penelitian Sejenis

Penelitian sejenis merupakan kumpulan tinjauan pustaka yang berisi penelitian yang sudah dilakukan sebelumnya yang berhubungan dengan tema penelitian yang penulis teliti guna menjadi referensi dalam penelitian yang disajikan pada Tabel 2.3.



Tabel 2.4 Penelitian Sejenis

No	Penulis	Metode dan Tools	Data Set	Akurasi	Kekurangan dan Kelebihan
1	(Sinaga et al., 2021)	<ul style="list-style-type: none"> <li>- <i>Decision Tree</i> C4.5</li> <li>- <i>Rapid Miner Studio</i></li> </ul>	<ul style="list-style-type: none"> <li>- Mahasiswa Fakultas Teknologi Informasi, Informatika dan Sistem Informasi</li> <li>- Angkatan 2017-2018 yang telah melewati masa studi empat semester</li> <li>- Sebanyak 98 data mahasiswa.</li> </ul>	Menghasilkan <i>accuracy</i> sebesar 90.00%, hasil dari <i>precision</i> adalah 87.50, dan hasil dari <i>recall</i> sebesar 100%.	<ul style="list-style-type: none"> <li>- Menghasilkan akurasi 90.00% yang tergolong sangat tinggi.</li> <li>- Perlu ditambahkan jumlah data serta atribut-atribut yang berkaitan dengan kegiatan mahasiswa sehingga dapat memberikan hasil dan analisis yang lebih baik lagi.</li> </ul>
2	(Rahman et al., 2020)	<ul style="list-style-type: none"> <li>- <i>Decision Tree</i> C4.5</li> </ul>	<ul style="list-style-type: none"> <li>- Data mahasiswa yang sudah lulus tahun 2016 di Universitas Peradaban</li> <li>- Sebanyak 151 data mahasiswa.</li> </ul>	Menghasilkan <i>accuracy</i> sebesar 88,74%, hasil dari <i>precision</i> adalah 91,79%, dan hasil dari <i>recall</i> sebesar 95,34%.	<ul style="list-style-type: none"> <li>- Jumlah <i>record</i> data yang digunakan untuk proses <i>training</i> sebaiknya ditingkatkan lagi karena jumlah data training mempengaruhi nilai akurasi</li> <li>- Penelitian ini perlu dilakukan penelitian lebih lanjut untuk membandingkan hasil prediksi,</li> </ul>
3.	(Iskandar et al., 2021)	<i>Naïve Bayes Classification</i>	<ul style="list-style-type: none"> <li>- Data mahasiswa yang sudah lulus tahun 2019 di Universitas Negeri</li> </ul>	Didapat pengujian pada perbandingan data <i>training</i> dan data <i>testing</i> sebesar	<ul style="list-style-type: none"> <li>- Menghasilkan akurasi yang baik, dengan nilai kelayakannya 5,04919.</li> </ul>

			Medan, jurusan Matematika - Sebanyak 318 data mahasiswa.	80:20 menghasilkan akurasi tertinggi dengan 79%.	- Pengujian pada perbandingan data training dan data testing sebesar 80:20 menghasilkan akurasi tertinggi dengan 79%,
4.	(Yuniarti et al., 2020)	- <i>Naïve Bayes Classification</i> - <i>Machine Learning</i> dengan tools WEKA	- Sebanyak 5.934 data bersih	Hasil menunjukkan klasifikasi dengan <i>naïve bayes Classification</i> memperoleh akurasi sebesar 99.41%.	- Menghasilkan akurasi sebesar 99,41%.
5.	(Yuwono et al., 2021)	- <i>Naïve Bayes Classification</i>	- Data pedagang kaki lima dan Usaha <i>Online</i> - 30 data yang didapatkan melalui menyebarkan <i>link</i> kuisioner kepada pelaku usaha.	Hasil pengujian <i>confusion matrix</i> dengan teknik split validasi, penggunaan metode klasifikasi <i>Naïve Bayes</i> terhadap dataset yang telah diambil pada objek penelitian diperoleh tingkat akurasi 75%,	- Mengingat nilai akurasi masih berada pada angka 75%, maka masih sangat dimungkinkan untuk dapat dilakukan penelitian selanjutnya, - Diharapkan dapat digunakan Dataset dalam jumlah yang lebih besar atau dengan sejumlah variabel lainnya guna meningkatkan performa dari metode yang digunakan.

Dari Tabel 2.1 Terdapat beberapa penelitian yang sejenis. Adapun perbandingan dan perbedaannya dengan penelitian ini adalah sebagai berikut:

- a. Penelitian dengan metode Algoritma *Decision C4.5* yang dilakukan oleh Sinaga *et al.* (2021) dan Rahman *et al.* (2020) menghasilkan nilai akurasi yang cukup tinggi. Namun, pada penelitian yang dilakukan oleh Rahman *et al.* (2020), jumlah *record* data yang digunakan untuk proses *training* sebaiknya ditingkatkan lagi. Hal ini, disebabkan nilai data *training* mempengaruhi nilai akurasi yang diperoleh. Selain itu, perlu dilakukan penelitian lebih lanjut dengan metode lain untuk membandingkan hasil prediksi.
- b. Pada penelitian Yuniarti *et al.* (2020) dengan dataset sebanyak 5.934 data bersih yang digunakan menghasilkan akurasi data sebesar 99,41% yang tergolong sangat tinggi. Hal ini membuktikan bahwa semakin banyak data yang digunakan dapat meningkatkan performa dari metode yang digunakan. Data yang banyak akan menghasilkan nilai akurasi yang cukup baik dan meningkatkan performa metode algoritma *Naïve Bayes Classification* seperti yang dilakukan oleh Iskandar *et al.* (2021).
- c. Berbeda dengan penelitian yang dilakukan oleh Yuniarti *et al.* (2020), penelitian yang dilakukan oleh Yuwono *et al.* (2021) data yang digunakan adalah sebanyak 30 data. Jumlah data yang sedikit ini, mempengaruhi nilai akurasi yang dihasilkan sehingga penelitian lebih lanjut perlu dilakukan. Jumlah data ini juga mempengaruhi performa dari metode yang digunakan.
- d. Penelitian dengan metode algoritma *Naïve Bayes Classification* juga dilakukan oleh Iskandar *et al.* (2021), Yuniarti *et al.* (2020), dan Yuwono *et al.* (2021).

Sedangkan penelitian dengan metode algoritma *Decision Tree* C4.5 dilakukan oleh Sinaga *et al.* (2021) dan Rahman *et al.* (2020).

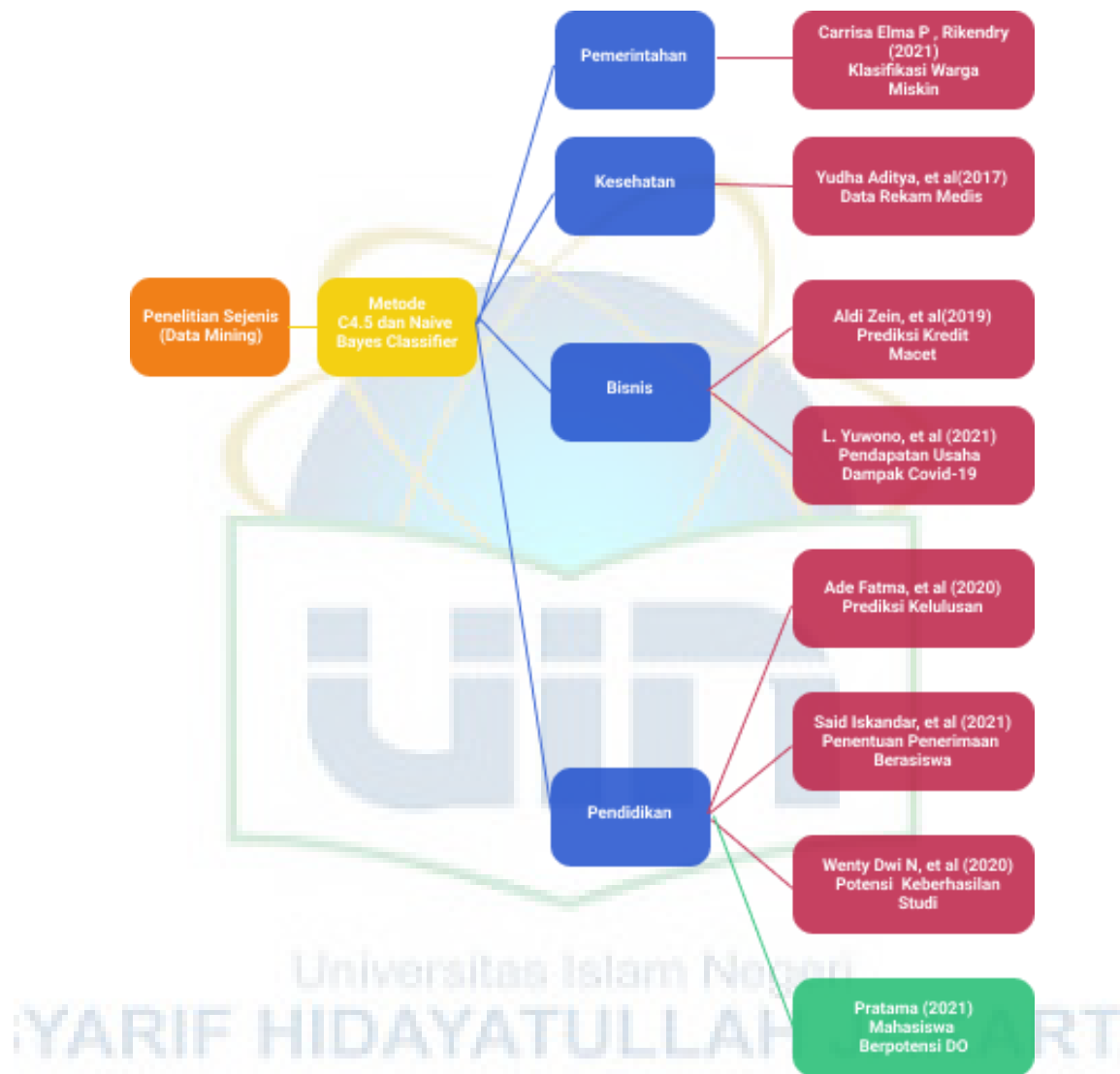
- e. Penelitian dengan menggunakan metode dan *tools* selain dilakukan oleh Sinaga *et al.* (2021) juga dilakukan oleh Yuniarti *et al.* (2020).
- f. Jumlah dataset yang digunakan dalam beberapa penelitian sejenis prediksi menggunakan algoritma *Decision Tree* C4.5 dan *Naïve Bayes Classification* berbeda-beda.

Berdasarkan perbedaan dan perbandingan literatur sejenis, maka penulis memutuskan untuk melakukan penelitian lebih lanjut dengan menggunakan metode algoritma *Decision Tree* C4.5 dan *Naïve Bayes Classification* sebagai perbandingan untuk mengklasifikasi mahasiswa berpotensi *drop out*. Penulis menemukan tule untuk membantu prediksi potensi *drop out* mahasiswa berdasarkan dataset mahasiswa jurusan Sistem Informasi UIN Syarif Hidayatullah Jakarta angkatan tahun 2010,2011, 2012, 2013, 2014, 2015, dan 2018. Penulis juga menggunakan pengujian *confusing matrix* untuk menemukan *accuracy*, *precision*, *recall*, validasi algoritma C4.5 dan validasi *Naïve Bayes Classification*.

### 2.13 Ranah Penelitian

Pada tahap ini, menggambarkan ranah penelitian sejenis yang dilakukan penulis berdasarkan literatur yang dibandingkan dimana menggunakan metode C4.5 dan *Naïve Bayes Classification* pada bidang tertentu. Berdasarkan penelitian sejenis yang identik dengan ranah penelitian sebelumnya, maka ranah penelitian

penulis berkaitan dengan bidang Pendidikan. Gambar 2.9 adalah ilustrasi dari ranah penelitian.



**Gambar 2.9** Ranah Penelitian

Ranah pada penelitian ini adalah berfokus pada mahasiswa jurusan sistem informasi yang berpotensi untuk di *drop out*. Mahasiswa jurusan sistem informasi yang dimaksud dalam penelitian ini adalah mahasiswa jurusan sistem informasi angkatan tahun 2010, 2011, 2012, 2013, 2014, 2015, dan 2018. Penulis sendiri



merupakan mahasiswa angkatan tahun 2017. Hal yang membedakan penelitian ini dengan penelitian-penelitian sebelumnya adalah dalam penelitian ini digunakan 2 metode sekaligus yakni, metode *C4.5* dan *Naïve Bayes Classification*. Sedangkan dataset yang digunakan adalah dataset dari 7 angkatan.



## **BAB 3**

### **METODOLOGI PENELITIAN**

#### **3.1 Metode Pengumpulan Data**

Dalam melakukan pengumpulan data untuk penelitian ini, penulis menggunakan 3 macam metode pengumpulan data sebagai berikut:

##### **3.1.1 Studi Pustaka**

Studi pustaka digunakan mencari dan mempelajari teori-teori dari artikel, jurnal dan juga situs penyedia layanan yang berhubungan dengan objek dari tugas akhir sebagai data dalam perancangan. Dalam mencari studi pustaka pada penelitian ini, penulis mendapatkannya di Perpustakaan UIN Syarif Hidayatullah Jakarta dan melalui internet secara *online*. Di akhir penelitian ini penulis telah mengumpulkan 3 *e-book*, 5 artikel, dan 28 jurnal.

##### **3.1.2 Wawancara**

Wawancara digunakan untuk mencari dan mempelajari lebih dalam lagi terkait permasalahan dalam penelitian ini. Penulis melakukan wawancara kepada pihak prodi yaitu Ibu Meinarini Catur Utami, M.T., beliau adalah Sekretaris Prodi Sistem Informasi, Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta periode 2014-2018. Wawancara dilakukan seputar *variable* apa saja yang dapat membuat mahasiswa Prodi Sistem Informasi, Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta mendapatkan *drop out*.

Menurut Bu Meinarini ada beberapa faktor yang membuat mahasiswa berpotensi *drop out* di antaranya adalah matakuliah yang belum diselesaikan hingga

semester 8, laporan PKL yang belum diselesaikan, dan belum membuat proposal skripsi.

### 3.1.3 Observasi

#### a. Waktu dan Tempat Penelitian

Penulis menentukan waktu dan tempat penelitian untuk melakukan beberapa tahapan yang harus dikerjakan, yang ditunjukkan pada Tabel 3.1.

**Tabel 3.1** Waktu dan Tempat Penelitian

<b>Tahapan</b>	<b>Waktu</b>	<b>Tempat / Tools</b>
Wawancara	12 November 2022	<i>Google Meet (Gmeet)</i>
Pengumpulan Data	13 September – 18 Oktober 2022	Rumah Penulis dan <i>Google Mail (Gmail)</i>
<i>Data Processing</i>	26 Juni – 27 Juni 2023	<i>Google Colab</i>
<i>Training</i>	28 Juni 2023	<i>Google Colab</i>

#### b. Perangkat Penelitian

Dalam penelitian ini, penulis menggunakan perangkat keras (*hardware*) dan perangkat lunak (*software*) yang ditunjukkan pada Tabel 3.2.

**Tabel 3.2** Perangkat Penelitian

<b>Hardware</b>	Laptop Lenovo Ideapad Slim 3	AMD Ryzen 5 5500U
		8 GB DDR4
		512 GB SSD
		Monitor 14.0 inch
<b>Software</b>	Sistem Operasi	Windows 11 Home Single Language 64 bit
	<i>Tools</i>	<i>Google Colab</i>
	Bahasa Pemrograman	Python 3.8.6

### c. Pengumpulan Dataset

Pengumpulan dataset tentang nilai-nilai dan status akhir mahasiswa prodi Sistem Informasi, Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta tahun masuk 2010, 2011, 2012, 2013, 2014, 2015 dan 2018. Dataset tersebut penulis dapatkan dari bagian Akademik Pusat UIN Syarif Hidayatullah Jakarta. Dengan adanya 6 angkatan untuk menjadi data *training* dan 1 angkatan untuk menjadi data *testing*, jumlah menjadi 7 angkatan yang menjadi dataset penulis pada penelitian kali ini. Rincian pembagian terlihat pada Tabel 3.3.

**Tabel 3.3** Dataset Mahasiswa

No	Dataset	Split	Jumlah
1	Data <i>Training</i>	80%	571
2	Data <i>Testing</i>	20%	150
Total		100%	721

### 3.2 Tahapan Penelitian

Penelitian ini menggunakan metode SEMMA yang terdiri atas *sample*, *explore*, *modify*, *model*, dan *assess*. Berikut tahapan penelitian menggunakan SEMMA sebagai metode penelitian.

#### a. *Sample* (Sampel)

Pada tahap ini peneliti mengumpulkan data mahasiswa prodi Sistem Informasi, Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta tahun masuk 2010, 2011, 2012, 2013, 2014, 2015, dan 2018.

#### b. *Explore* (Eksplorasi)

Setelah mendapatkan data, langkah selanjutnya adalah menjelajahi data tersebut. Pada tahap ini peneliti mengumpulkan data mahasiswa yang relevan. Data yang diperlukan adalah data status mahasiswa (*drop out*/lulus), matakuliah yang mengulang, laporan PKL, dan laporan skripsi mahasiswa pada mahasiswa tahun masuk 2010-2015. Serta data matakuliah yang mengulang, laporan PKL, dan laporan skripsi mahasiswa pada mahasiswa tahun masuk 2018.

c. *Modify* (Modifikasi)

Setelah melakukan penyaringan data dimana hanya data status mahasiswa (*drop out*/lulus), matakuliah yang mengulang, laporan PKL, dan laporan skripsi mahasiswa yang digunakan, selanjutnya adalah *modify* yaitu memodifikasi data dengan melakukan penghapusan data mahasiswa yang berstatus **Mengundurkan Diri**.

d. *Model* (Model)

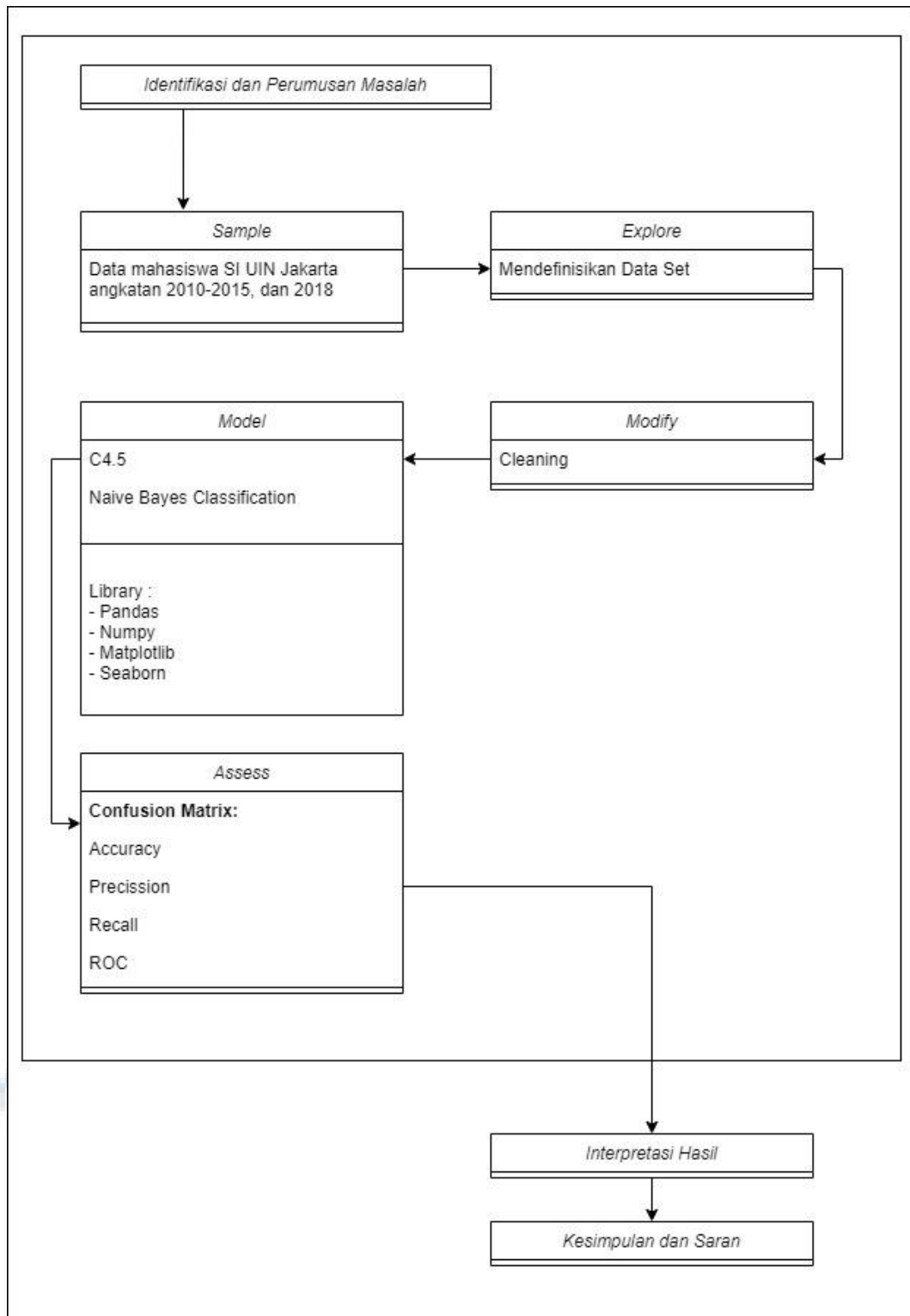
Tahap ini melibatkan pembangunan model prediktif untuk mengklasifikasi mahasiswa yang berpotensi *drop out*. Ditahap ini data set dibagi menjadi data *training* dan data *testing* untuk kemudian diolah. Metode yang digunakan dalam modeling ini adalah algoritma C4.5 dan *Naïve Bayes Classification*.

e. *Assess* (Penilaian)

Setelah membangun model prediktif, langkah terakhir adalah mengevaluasi performa model menggunakan metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan ROC untuk mengukur seberapa baik model dapat mengklasifikasi mahasiswa berpotensi *drop out*.

### 3.3 Kerangka Penelitian

Proses penelitian ini dimulai dengan mengidentifikasi dan merumuskan masalah yang relevan, diikuti dengan pengolahan data menggunakan metode SEMMA yang terdiri atas lima tahap, yaitu *sample*, *explore*, *modify*, *model*, dan *assess*. Tahap *sample* melibatkan tinjauan pustaka serta pengumpulan data. Pada tahap *explore*, *variabel* data didefinisikan untuk data *set* yang digunakan. Selanjutnya, tahap *modify* dilakukan untuk membersihkan data agar lebih terstruktur. Tahap keempat adalah *model*, di mana data set diberi label kelas menggunakan C4.5 dan *Naïve Bayes Classification*. Terakhir, tahap *assess* atau evaluasi kinerja model dilakukan dengan mengukur *accuracy*, *precision*, *recall*, dan ROC. Penelitian ini diakhiri dengan menyajikan kesimpulan dan saran berdasarkan hasil penelitian. Tahapan penelitian ini dapat dilihat pada Gambar 3.1 tentang kerangka penelitian untuk mahasiswa berpotensi *drop out*.



**Gambar 3.1** Kerangka Penelitian

## BAB 4

### HASIL DAN PEMBAHASAN

#### 4.1 *Sample*

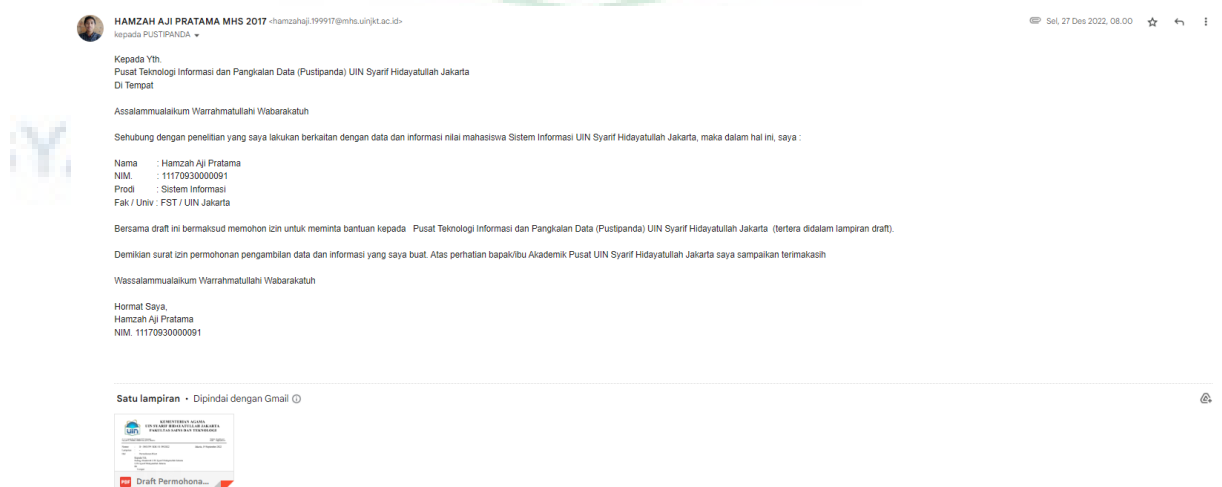
Pada tahap pertama metode SEMMA adalah *Sample* (Sampel), dalam penelitian kali ini, *sample* yang digunakan adalah tinjauan pustaka dan pengumpulan data.

##### 4.1.1 Tinjauan Pustaka

Peneliti melakukan tinjauan pustaka terkait klasifikasi data, metode Algoritma C4.5 dan *Naïve Bayes Classification* dari jurnal-jurnal yang dipaparkan pada Tabel 2.1.

##### 4.1.2 Pengumpulan Data

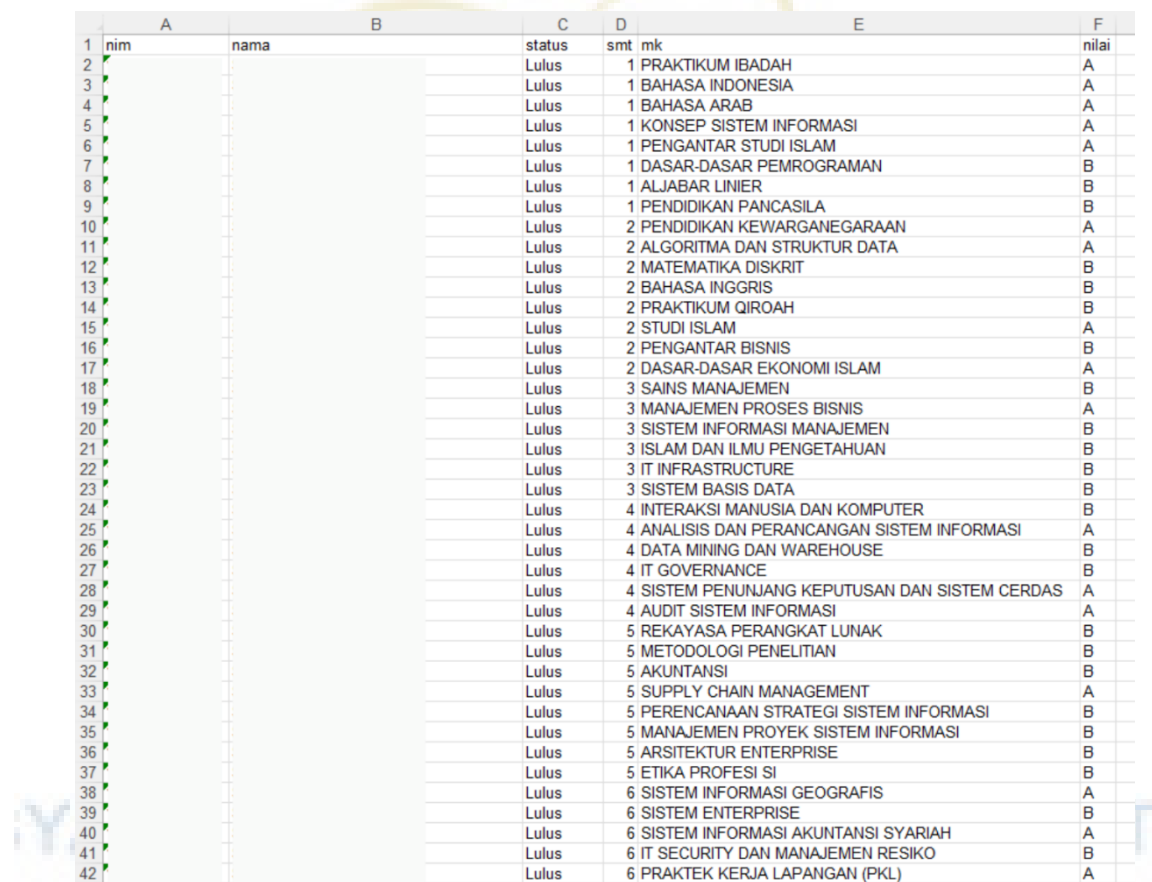
Sumber dataset dalam penelitian ini berasal dari Pusat Teknologi Informasi dan Pangkalan Data (Pustipanda) UIN Syarif Hidayatullah Jakarta yang dikirimkan melalui *email* ke pihak yang bersangkutan.



**Gambar 4.1** *Template* Email Permohonan Permintaan Data



Data yang dikirimkan pihak Pustipanda berformat *Microsoft Excel Open XML Spreadsheet (XLSX)* yang berisi Nama Mahasiswa, NIM Mahasiswa, Status Mahasiswa, Semester, Matakuliah, dan Nilai Matakuliah untuk mahasiswa tahun masuk 2010, 2011, 2012, 2013, 2014, 2015. Serta Nama Mahasiswa, NIM Mahasiswa, Semester, Matakuliah, dan Nilai Mahasiswa untuk mahasiswa tahun masuk 2018, ditunjukkan Gambar 4.2.



	A	B	C	D	E	F
	nim	nama	status	smt	mk	nilai
2			Lulus	1	PRAKTIKUM IBADAH	A
3			Lulus	1	BAHASA INDONESIA	A
4			Lulus	1	BAHASA ARAB	A
5			Lulus	1	KONSEP SISTEM INFORMASI	A
6			Lulus	1	PENGANTAR STUDI ISLAM	A
7			Lulus	1	DASAR-DASAR PEMROGRAMAN	B
8			Lulus	1	ALJABAR LINIER	B
9			Lulus	1	PENDIDIKAN PANCASILA	B
10			Lulus	2	PENDIDIKAN KEWARGANEGARAAN	A
11			Lulus	2	ALGORITMA DAN STRUKTUR DATA	A
12			Lulus	2	MATEMATIKA DISKRIT	B
13			Lulus	2	BAHASA INGGRIS	B
14			Lulus	2	PRAKTIKUM QIROAH	B
15			Lulus	2	STUDI ISLAM	A
16			Lulus	2	PENGANTAR BISNIS	B
17			Lulus	2	DASAR-DASAR EKONOMI ISLAM	A
18			Lulus	3	SAINS MANAJEMEN	B
19			Lulus	3	MANAJEMEN PROSES BISNIS	A
20			Lulus	3	SISTEM INFORMASI MANAJEMEN	B
21			Lulus	3	ISLAM DAN ILMU PENGETAHUAN	B
22			Lulus	3	IT INFRASTRUCTURE	B
23			Lulus	3	SISTEM BASIS DATA	B
24			Lulus	4	INTERAKSI MANUSIA DAN KOMPUTER	B
25			Lulus	4	ANALISIS DAN PERANCANGAN SISTEM INFORMASI	A
26			Lulus	4	DATA MINING DAN WAREHOUSE	B
27			Lulus	4	IT GOVERNANCE	B
28			Lulus	4	SISTEM PENUNJANG KEPUTUSAN DAN SISTEM CERDAS	A
29			Lulus	4	AUDIT SISTEM INFORMASI	A
30			Lulus	5	REKAYASA PERANGKAT LUNAK	B
31			Lulus	5	METODOLOGI PENELITIAN	B
32			Lulus	5	AKUNTANSI	B
33			Lulus	5	SUPPLY CHAIN MANAGEMENT	A
34			Lulus	5	PERENCANAAN STRATEGI SISTEM INFORMASI	B
35			Lulus	5	MANAJEMEN PROYEK SISTEM INFORMASI	B
36			Lulus	5	ARSITEKTUR ENTERPRISE	B
37			Lulus	5	ETIKA PROFESI SI	B
38			Lulus	6	SISTEM INFORMASI GEOGRAFIS	A
39			Lulus	6	SISTEM ENTERPRISE	B
40			Lulus	6	SISTEM INFORMASI AKUNTANSI SYARIAH	A
41			Lulus	6	IT SECURITY DAN MANAJEMEN RESIKO	B
42			Lulus	6	PRAKTEK KERJA LAPANGAN (PKL)	A

**Gambar 4.2** Data Mahasiswa yang Dikirimkan oleh Pustipanda

## 4.2 Explore

Pada pengumpulan data sebelumnya, terdapat 6 kolom yang selanjutnya akan disesuaikan dengan menggunakan 3 faktor menyebabkan mahasiswa

berpotensi *dropout*. Faktor-faktor tersebut adalah matakuliah yang belum diselesaikan hingga semester 8, laporan PKL yang belum diselesaikan, dan belum membuat proposal skripsi. Gambar 4.3 berikut adalah hasil penyesuaian dengan menggunakan 3 faktor penyebab mahasiswa berpotensi *dropout*.

NO	NIM	NAMA	STATUS	Semua Matakul Selesai pada Semester 8 ?	Laporan PKL Selesai ?	Sudah Ambil Skripsi ?
1			Lulus	IYA	IYA	IYA
2			Lulus	IYA	IYA	IYA
3			Lulus	IYA	IYA	IYA
4			Lulus	IYA	IYA	IYA
5			Lulus	IYA	IYA	IYA
6			Lulus	IYA	IYA	IYA
7			Lulus	IYA	IYA	IYA
8			Lulus	IYA	IYA	IYA
9			Drop Out	TIDAK	TIDAK	TIDAK
10			Drop Out	TIDAK	TIDAK	TIDAK
11			Drop Out	TIDAK	TIDAK	TIDAK
12			Drop Out	TIDAK	TIDAK	TIDAK
13			Drop Out	TIDAK	TIDAK	TIDAK
14			Drop Out	TIDAK	TIDAK	TIDAK
15			Lulus	IYA	IYA	IYA
16			Lulus	IYA	IYA	IYA
17			Lulus	IYA	IYA	IYA
18			Drop Out	TIDAK	TIDAK	TIDAK
19			Lulus	IYA	IYA	IYA
20			Lulus	IYA	IYA	IYA
21			Lulus	IYA	IYA	IYA
22			Lulus	IYA	IYA	IYA
23			Lulus	IYA	IYA	IYA
24			Lulus	IYA	IYA	IYA

**Gambar 4.3** Format Penyesuaian menggunakan Faktor-Faktor Mahasiswa Berpotensi DO

Dari data yang didapatkan, peneliti melakukan *explore* secara lebih mendalam, tidak hanya mengetahui jumlah mahasiswa yang berstatus LULUS atau *Drop Out*, namun jumlah mahasiswa disetiap kriteria yang ada dalam penelitian. Tabel adalah hasil *explore* data mahasiswa tahun masuk 2010 sampai dengan 2015.

**Tabel 4.1** Table *Explore* Data Mahasiswa

		2010		2011		2012		2013		2014		2015	
		JML	%	JML	%	JM	%	JML	%	JML	%	JML	%
Status	Jumlah Mahasiswa (Lulus dan Drop Out)	80	100.00	111	100.00	149	100.00	126	100.00	160	100.00	88	100.00
	LULUS	62	77.50	90	81.08	125	83.89	98	77.78	120	75.00	82	93.18
	DROP OUT	18	22.50	21	18.92	24	16.11	28	22.22	40	25.00	6	6.82
Semua Matkul Selesai pada Semester 8 ?	IYA	67	83.75	94	84.68	127	85.23	87	69.05	101	63.13	61	69.32
	TIDAK	13	16.25	17	15.32	22	14.77	39	30.95	59	36.88	27	30.68
Laporan PKL Selesai ?	IYA	66	82.50	91	81.98	124	83.22	79	62.70	104	65.00	42	47.73
	TIDAK	14	17.50	20	18.02	25	16.78	47	37.30	56	35.00	46	52.27
Sudah Ambil Skripsi ?	IYA	68	85.00	97	87.39	128	85.91	92	73.02	129	80.63	78	88.64
	TIDAK	12	15.00	14	12.61	21	14.09	34	26.98	31	19.38	10	11.36

Dari mahasiswa tahun masuk 2010, 2011, 2012, 2013, 2014, 2015 terdapat 136 mahasiswa yang berstatus *Drop Out* dan 578 mahasiswa yang berstatus LULUS.

### 4.3 Modify

Pada tahap *modify*, peneliti melakukan *cleansing* atau penghapusan data mahasiswa tahun masuk 2010, 2011, 2012, 2013, 2014, dan 2015 yang berstatus **Mengundurkan Diri**. Peneliti berfokus pada mahasiswa yang berstatus **Drop Out** dan **Lulus**, sehingga mahasiswa yang berstatus mengundurkan diri perlu dihapus pada data yang tersedia.

Setelah selesai proses *cleansing* pada data yang ada, maka langkah selanjutnya adalah data diubah ke dalam bentuk *numerik*. Perubahan mengonversi atribut yang bersifat kategorikal menjadi bentuk yang dapat diproses oleh algoritma klasifikasi. Gambar 4.4 berikut adalah hasil perubahan data menjadi numerik.

```
[ ] # Membuat objek LabelEncoder
label_encoder = LabelEncoder()

# Melakukan encoding pada semua fitur kecuali kolom "STATUS"
for column in dataset.columns[1:]:
    dataset[column] = label_encoder.fit_transform(dataset[column])

[ ] # Cek dataset yang sudah di encode
dataset
```

	STATUS	Semua Matkul Selesai pada Semester 8 ?	Laporan PKL Selesai ?	Sudah Ambil Skripsi ?
0	Lulus	0	0	0
1	Lulus	0	0	0
2	Lulus	0	0	0
3	Lulus	0	0	0
4	Lulus	0	0	0
...	...	...	...	...
709	Lulus	0	1	0
710	Lulus	0	0	0
711	Lulus	0	0	0
712	Lulus	0	0	0
713	Lulus	0	1	0

714 rows x 4 columns

**Gambar 4.4** Dataset Dilakukan Penyesuaian dengan Mengubahnya Menjadi Numerik

#### 4.4 Model

Tahap pemodelan pada penelitian ini menggunakan dua model yaitu metode C4.5 dan *Naïve Bayes Classification* (NBC) untuk proses klasifikasi data mahasiswa berpotensi *drop out*.

Pada tahap ini dataset dibagi menjadi dua bagian, yaitu data *training* dan data *testing* dengan beberapa percobaan rasio yaitu 90% data *training* dan 10% data *testing*, 80% data *training* dan 20% data *testing*, 70% data *training* dan 30% data *testing*, berdasarkan penelitian Gormantara (2020). Tabel 4.1 adalah perbandingan data *training* dan data *testing* yang digunakan dalam penelitian.

**Tabel 4.2** Perbandingan Data *Train* dan Data *Test* Metode C4.5 dan NBC

<b>Data Training</b>	<b>Data Testing</b>
90%	10%
80%	20%
70%	30%

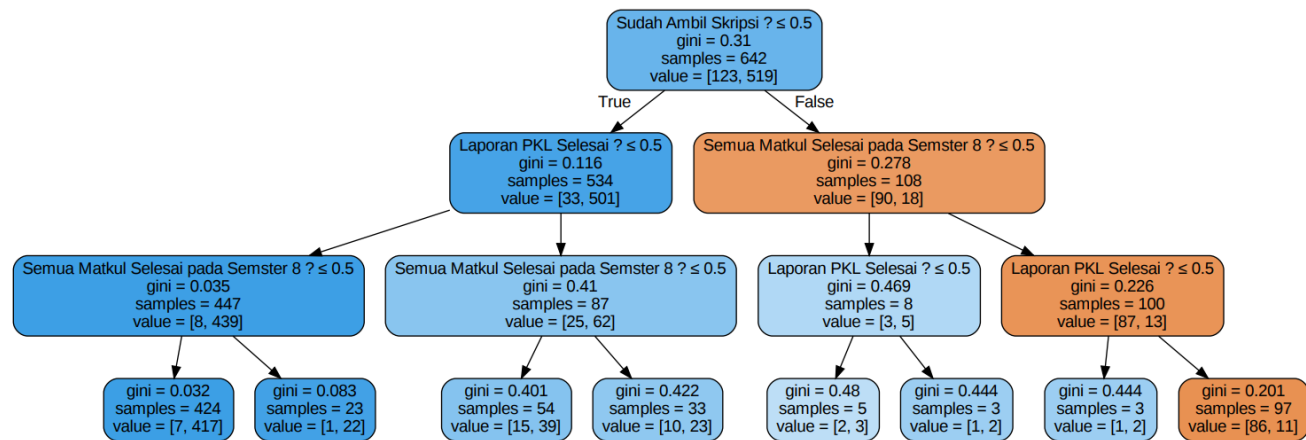
Berikut adalah penulisan dengan python yang berfungsi untuk membagi data *training* dan data *testing*.

```
x_train, x_test, y_train, y_test = train_test_split(x, y,
test_size=0.2, random_state=42)
```

Dari hasil perbandingan dataset yang ada, algoritma C4.5 diketahui mendapatkan *score* tertinggi dari perbandingan 90% data *training* dan 10% data *testing*. Diketahui jumlah mahasiswa prodi Sistem Informasi tahun masuk 2018 sebanyak 150 mahasiswa, terdapat 119 mahasiswa berpotensi lulus dan 31 mahasiswa berpotensi *drop out*.

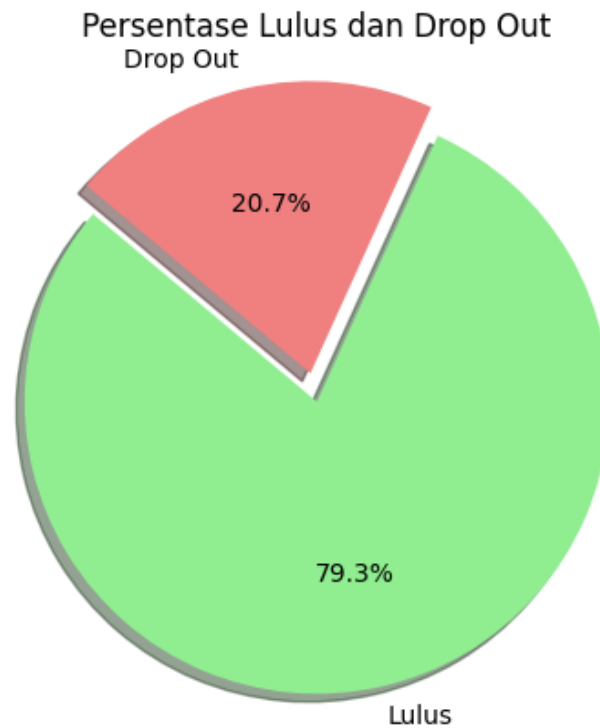
Pada algoritma C4.5, terdapat *decision tree* dalam proses pengaplikasiannya, *decision tree* yang dibuat sebagai gambaran proses dalam mengklasifikasi mahasiswa berpotensi *dropout*. Gambar 4. adalah *decision tree* dari algoritma C4.5.





Gambar 4.5 Decision Tree C4.5

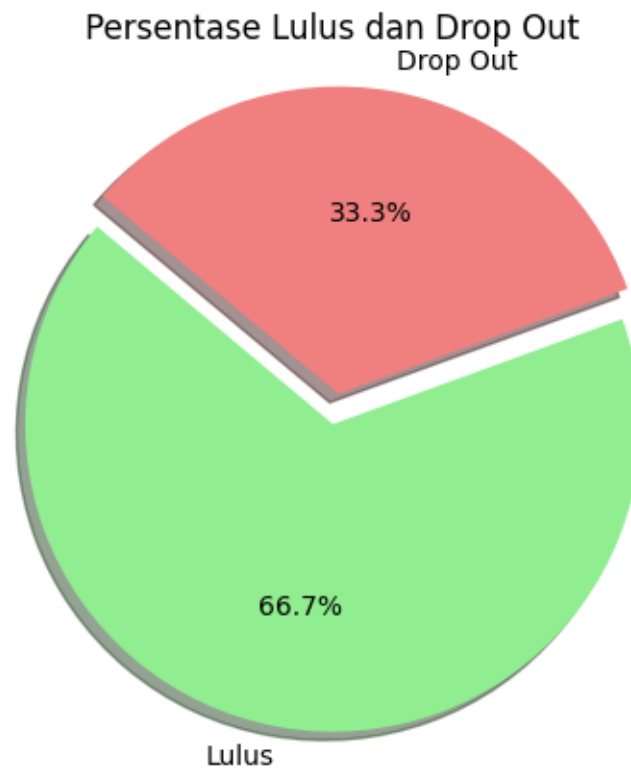
Hasil akhir dalam penelitian ini adalah pengklasifikasian kedua algoritma dalam mengklasifikasi mahasiswa berpotensi *drop out*. Gambar 4.5 adalah diagram lingkaran hasil akhir dari pengklasifikasian model algoritma C4.5 terdapat 79,3% mahasiswa masuk ke status pengklasifikasian lulus dan 20,7% masuk ke status pengklasifikasian *drop out*.



**Gambar 4.6** Diagram Lingkaran Hasil Akhir Algoritma C4.5

Sedangkan pada algoritma *Naïve Bayes Classification* (NBC), terdapat 66,7% mahasiswa masuk ke status pengklasifikasian lulus dan 33,3% masuk ke status pengklasifikasian *drop out*.





**Gambar 4.7** Diagram Lingkaran Hasil Akhir Algoritma NBC

Tabel 4.2 adalah contoh hasil dari model algoritma C4.5 menggunakan perbandingan dataset 90:10.

Tabel 4.3 Hasil dari model Algoritma C4.5

Nama	Semua Matkul Selesai pada Semster 8 ?	Laporan PKL Selesai ?	Sudah Ambil Skripsi ?	Probabilitas Drop Out	Probabilitas LULUS	STATUS
TSI	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
ARH	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
NH	IYA	TIDAK	IYA	27,78	72,22	Lulus
NAR	IYA	IYA	IYA	1,65	98,35	Lulus
RN	IYA	IYA	IYA	1,65	98,35	Lulus
SN	IYA	IYA	IYA	1,65	98,35	Lulus
RN	TIDAK	TIDAK	IYA	30,30	69,70	Lulus
SFR	TIDAK	TIDAK	IYA	30,30	69,70	Lulus
EI	TIDAK	IYA	IYA	4,35	95,65	Lulus
RI	IYA	IYA	IYA	1,65	98,35	Lulus
MFA	IYA	IYA	IYA	1,65	98,35	Lulus
FF	IYA	IYA	IYA	1,65	98,35	Lulus
CAR	IYA	TIDAK	IYA	27,78	72,22	Lulus
LKN	TIDAK	TIDAK	IYA	30,30	69,70	Lulus
ARF	IYA	TIDAK	IYA	27,78	72,22	Lulus
DMM	IYA	TIDAK	IYA	27,78	72,22	Lulus
VO	IYA	IYA	IYA	1,65	98,35	Lulus
AIP	IYA	TIDAK	IYA	27,78	72,22	Lulus
MFA	IYA	IYA	IYA	1,65	98,35	Lulus
AM	IYA	TIDAK	TIDAK	33,33	66,67	Lulus
SR	IYA	TIDAK	IYA	27,78	72,22	Lulus

NAS	IYA	IYA	IYA	1,65	98,35	Lulus
APP	IYA	TIDAK	IYA	27,78	72,22	Lulus
TRA	IYA	IYA	IYA	1,65	98,35	Lulus
KM	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
JRM	IYA	IYA	IYA	1,65	98,35	Lulus
OAP	IYA	TIDAK	IYA	27,78	72,22	Lulus
S	IYA	IYA	IYA	1,65	98,35	Lulus
MJAA	IYA	IYA	IYA	1,65	98,35	Lulus
MS	IYA	TIDAK	IYA	27,78	72,22	Lulus
PNF	IYA	IYA	IYA	1,65	98,35	Lulus
WA	IYA	TIDAK	IYA	27,78	72,22	Lulus
PRF	IYA	IYA	IYA	1,65	98,35	Lulus
SDA	IYA	IYA	IYA	1,65	98,35	Lulus
AS	IYA	TIDAK	IYA	27,78	72,22	Lulus
AB	IYA	TIDAK	IYA	27,78	72,22	Lulus
AJP	IYA	TIDAK	IYA	27,78	72,22	Lulus
DAL	IYA	TIDAK	IYA	27,78	72,22	Lulus
NU	IYA	TIDAK	IYA	27,78	72,22	Lulus
SDAS	IYA	TIDAK	TIDAK	33,33	66,67	Lulus
MEI	IYA	TIDAK	IYA	27,78	72,22	Lulus
AS	IYA	TIDAK	IYA	27,78	72,22	Lulus
KUNA	IYA	TIDAK	IYA	27,78	72,22	Lulus
DR	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
KA	IYA	IYA	IYA	1,65	98,35	Lulus
MJF	IYA	IYA	IYA	1,65	98,35	Lulus

FHS	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
AD	IYA	TIDAK	IYA	27,78	72,22	Lulus
YR	IYA	TIDAK	IYA	27,78	72,22	Lulus
DH	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
FA	IYA	TIDAK	IYA	27,78	72,22	Lulus
IAP	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
TH	IYA	IYA	IYA	1,65	98,35	Lulus
AF	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
AKA	IYA	IYA	IYA	1,65	98,35	Lulus
RF	IYA	TIDAK	IYA	27,78	72,22	Lulus
CDAAPC	IYA	TIDAK	IYA	27,78	72,22	Lulus
ISS	IYA	TIDAK	IYA	27,78	72,22	Lulus
APH	IYA	TIDAK	IYA	27,78	72,22	Lulus
FN	IYA	TIDAK	IYA	27,78	72,22	Lulus
VAS	IYA	TIDAK	IYA	27,78	72,22	Lulus
MHD	TIDAK	TIDAK	IYA	30,30	69,70	Lulus
GAF	IYA	TIDAK	IYA	27,78	72,22	Lulus
SA	IYA	IYA	IYA	1,65	98,35	Lulus
RII	IYA	TIDAK	IYA	27,78	72,22	Lulus
YAJ	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
NZR	TIDAK	TIDAK	IYA	30,30	69,70	Lulus
BP	IYA	TIDAK	IYA	27,78	72,22	Lulus
FHS	IYA	TIDAK	TIDAK	33,33	66,67	Lulus
MAH	TIDAK	IYA	IYA	4,35	95,65	Lulus
DGR	IYA	TIDAK	IYA	27,78	72,22	Lulus

MRFB	IYA	TIDAK	IYA	27,78	72,22	Lulus
EAR	IYA	IYA	IYA	1,65	98,35	Lulus
PA	IYA	IYA	IYA	1,65	98,35	Lulus
SMR	IYA	IYA	IYA	1,65	98,35	Lulus
MFK	IYA	IYA	IYA	1,65	98,35	Lulus
SSZ	IYA	TIDAK	IYA	27,78	72,22	Lulus
YAI	IYA	IYA	IYA	1,65	98,35	Lulus
MA	IYA	TIDAK	IYA	27,78	72,22	Lulus
MTPP	IYA	IYA	IYA	1,65	98,35	Lulus
APU	IYA	TIDAK	IYA	27,78	72,22	Lulus
AWF	IYA	IYA	IYA	1,65	98,35	Lulus
NAA	IYA	IYA	IYA	1,65	98,35	Lulus
MHA	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
MRF	TIDAK	TIDAK	IYA	30,30	69,70	Lulus
MESP	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
EP	IYA	TIDAK	IYA	27,78	72,22	Lulus
RA	IYA	IYA	IYA	1,65	98,35	Lulus
ABN	TIDAK	TIDAK	IYA	30,30	69,70	Lulus
ASA	IYA	TIDAK	IYA	27,78	72,22	Lulus
HIAL	IYA	TIDAK	IYA	27,78	72,22	Lulus
F	TIDAK	TIDAK	IYA	30,30	69,70	Lulus
ARA	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
DN	IYA	TIDAK	IYA	27,78	72,22	Lulus
AQNA	IYA	TIDAK	TIDAK	33,33	66,67	Lulus
MDAAY	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out

RA	IYA	TIDAK	IYA	27,78	72,22	Lulus
AALR	IYA	TIDAK	IYA	27,78	72,22	Lulus
ACA	IYA	TIDAK	IYA	27,78	72,22	Lulus
DAR	IYA	TIDAK	IYA	27,78	72,22	Lulus
MRA	TIDAK	TIDAK	IYA	30,30	69,70	Lulus
RIH	IYA	IYA	IYA	1,65	98,35	Lulus
MRPD	IYA	TIDAK	IYA	27,78	72,22	Lulus
AFFP	IYA	IYA	IYA	1,65	98,35	Lulus
SRF	TIDAK	TIDAK	IYA	30,30	69,70	Lulus
MA	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
JF	IYA	IYA	IYA	1,65	98,35	Lulus
ALK	IYA	TIDAK	IYA	27,78	72,22	Lulus
MRP	IYA	TIDAK	IYA	27,78	72,22	Lulus
EAR	IYA	IYA	IYA	1,65	98,35	Lulus
MG	IYA	TIDAK	IYA	27,78	72,22	Lulus
SBK	IYA	TIDAK	IYA	27,78	72,22	Lulus
MR	TIDAK	IYA	TIDAK	33,33	66,67	Lulus
MFAK	IYA	TIDAK	IYA	27,78	72,22	Lulus
AUH	IYA	IYA	IYA	1,65	98,35	Lulus
SF	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
LKS	IYA	TIDAK	TIDAK	33,33	66,67	Lulus
LMA	IYA	TIDAK	IYA	27,78	72,22	Lulus
RPPA	IYA	TIDAK	IYA	27,78	72,22	Lulus
NAAM	IYA	TIDAK	TIDAK	33,33	66,67	Lulus
AR	TIDAK	TIDAK	IYA	30,30	69,70	Lulus

AS	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
AV	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
NAR	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
AR	IYA	TIDAK	IYA	27,78	72,22	Lulus
RNM	IYA	TIDAK	IYA	27,78	72,22	Lulus
DK	IYA	IYA	IYA	1,65	98,35	Lulus
RAPS	IYA	IYA	IYA	1,65	98,35	Lulus
WFF	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
AR	IYA	TIDAK	IYA	27,78	72,22	Lulus
RV	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
SNR	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
FQ	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
HR	IYA	IYA	IYA	1,65	98,35	Lulus
VP	IYA	IYA	IYA	1,65	98,35	Lulus
FA	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
NS	IYA	IYA	IYA	1,65	98,35	Lulus
MR	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
MA	IYA	IYA	IYA	1,65	98,35	Lulus
AMI	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
RR	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
MH	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
HMA	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
RCW	IYA	TIDAK	IYA	27,78	72,22	Lulus
MR	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
RGP	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out

IH	IYA	IYA	IYA	1,65	98,35	Lulus
MQZZ	TIDAK	TIDAK	TIDAK	88,66	11,34	Drop Out
MO	IYA	TIDAK	IYA	27,78	72,22	Lulus
RRSM	TIDAK	TIDAK	IYA	30,30	69,70	Lulus

Sedangkan pada Algoritma *Naïve Bayes Classification* (NBC), terdapat 100 mahasiswa berpotensi lulus dan 50 mahasiswa berpotensi *drop out*. Tabel 4.3 adalah hasil dari model Algoritma *Naïve Bayes Classification* menggunakan perbandingan dataset 80:20.

Nama	Semua Matkul Selesai pada Semester 8 ?	Laporan PKL Selesai ?	Sudah Ambil Skripsi ?	Probabilitas Drop Out	Probabilitas Lulus	STATUS
TSI	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
ARH	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
NH	IYA	TIDAK	IYA	5,24	94,76	Lulus
NAR	IYA	IYA	IYA	0,00	100,00	Lulus
RN	IYA	IYA	IYA	0,00	100,00	Lulus
SN	IYA	IYA	IYA	0,00	100,00	Lulus
RN	TIDAK	TIDAK	IYA	92,72	7,28	Drop Out
SFR	TIDAK	TIDAK	IYA	92,72	7,28	Drop Out
EI	TIDAK	IYA	IYA	0,89	99,11	Lulus
RI	IYA	IYA	IYA	0,00	100,00	Lulus
MFA	IYA	IYA	IYA	0,00	100,00	Lulus
FF	IYA	IYA	IYA	0,00	100,00	Lulus



CAR	IYA	TIDAK	IYA	5,24	94,76	Lulus
LKN	TIDAK	TIDAK	IYA	92,72	7,28	Drop Out
ARF	IYA	TIDAK	IYA	5,24	94,76	Lulus
DMM	IYA	TIDAK	IYA	5,24	94,76	Lulus
VO	IYA	IYA	IYA	0,00	100,00	Lulus
AIP	IYA	TIDAK	IYA	5,24	94,76	Lulus
MFA	IYA	IYA	IYA	0,00	100,00	Lulus
AM	IYA	TIDAK	TIDAK	100,00	0,00	Drop Out
SR	IYA	TIDAK	IYA	5,24	94,76	Lulus
NAS	IYA	IYA	IYA	0,00	100,00	Lulus
APP	IYA	TIDAK	IYA	5,24	94,76	Lulus
TRA	IYA	IYA	IYA	0,00	100,00	Lulus
KM	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
JRM	IYA	IYA	IYA	0,00	100,00	Lulus
OAP	IYA	TIDAK	IYA	5,24	94,76	Lulus
S	IYA	IYA	IYA	0,00	100,00	Lulus
MJAA	IYA	IYA	IYA	0,00	100,00	Lulus
MS	IYA	TIDAK	IYA	5,24	94,76	Lulus
PNF	IYA	IYA	IYA	0,00	100,00	Lulus
WA	IYA	TIDAK	IYA	5,24	94,76	Lulus
PRF	IYA	IYA	IYA	0,00	100,00	Lulus
SDA	IYA	IYA	IYA	0,00	100,00	Lulus
AS	IYA	TIDAK	IYA	5,24	94,76	Lulus
AB	IYA	TIDAK	IYA	5,24	94,76	Lulus
AJP	IYA	TIDAK	IYA	5,24	94,76	Lulus

DAL	IYA	TIDAK	IYA	5,24	94,76	Lulus
NU	IYA	TIDAK	IYA	5,24	94,76	Lulus
SDAS	IYA	TIDAK	TIDAK	100,00	0,00	Drop Out
MEI	IYA	TIDAK	IYA	5,24	94,76	Lulus
AS	IYA	TIDAK	IYA	5,24	94,76	Lulus
KUNA	IYA	TIDAK	IYA	5,24	94,76	Lulus
DR	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
KA	IYA	IYA	IYA	0,00	100,00	Lulus
MJF	IYA	IYA	IYA	0,00	100,00	Lulus
FHS	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
AD	IYA	TIDAK	IYA	5,24	94,76	Lulus
YR	IYA	TIDAK	IYA	5,24	94,76	Lulus
DH	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
FA	IYA	TIDAK	IYA	5,24	94,76	Lulus
IAP	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
TH	IYA	IYA	IYA	0,00	100,00	Lulus
AF	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
AKA	IYA	IYA	IYA	0,00	100,00	Lulus
RF	IYA	TIDAK	IYA	5,24	94,76	Lulus
CDAAPC	IYA	TIDAK	IYA	5,24	94,76	Lulus
ISS	IYA	TIDAK	IYA	5,24	94,76	Lulus
APH	IYA	TIDAK	IYA	5,24	94,76	Lulus
FN	IYA	TIDAK	IYA	5,24	94,76	Lulus
VAS	IYA	TIDAK	IYA	5,24	94,76	Lulus
MHD	TIDAK	TIDAK	IYA	92,72	7,28	Drop Out

GAF	IYA	TIDAK	IYA	5,24	94,76	Lulus
SA	IYA	IYA	IYA	0,00	100,00	Lulus
RII	IYA	TIDAK	IYA	5,24	94,76	Lulus
YAJ	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
NZR	TIDAK	TIDAK	IYA	92,72	7,28	Drop Out
BP	IYA	TIDAK	IYA	5,24	94,76	Lulus
FHS	IYA	TIDAK	TIDAK	100,00	0,00	Drop Out
MAH	TIDAK	IYA	IYA	0,89	99,11	Lulus
DGR	IYA	TIDAK	IYA	5,24	94,76	Lulus
MRFB	IYA	TIDAK	IYA	5,24	94,76	Lulus
EAR	IYA	IYA	IYA	0,00	100,00	Lulus
PA	IYA	IYA	IYA	0,00	100,00	Lulus
SMR	IYA	IYA	IYA	0,00	100,00	Lulus
MFK	IYA	IYA	IYA	0,00	100,00	Lulus
SSZ	IYA	TIDAK	IYA	5,24	94,76	Lulus
YAI	IYA	IYA	IYA	0,00	100,00	Lulus
MA	IYA	TIDAK	IYA	5,24	94,76	Lulus
MTPP	IYA	IYA	IYA	0,00	100,00	Lulus
APU	IYA	TIDAK	IYA	5,24	94,76	Lulus
AWF	IYA	IYA	IYA	0,00	100,00	Lulus
NAA	IYA	IYA	IYA	0,00	100,00	Lulus
MHA	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
MRF	TIDAK	TIDAK	IYA	92,72	7,28	Drop Out
MESP	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
EP	IYA	TIDAK	IYA	5,24	94,76	Lulus

RA	IYA	IYA	IYA	0,00	100,00	Lulus
ABN	TIDAK	TIDAK	IYA	92,72	7,28	Drop Out
ASA	IYA	TIDAK	IYA	5,24	94,76	Lulus
HIAL	IYA	TIDAK	IYA	5,24	94,76	Lulus
F	TIDAK	TIDAK	IYA	92,72	7,28	Drop Out
ARA	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
DN	IYA	TIDAK	IYA	5,24	94,76	Lulus
AQNA	IYA	TIDAK	TIDAK	100,00	0,00	Drop Out
MDAAY	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
RA	IYA	TIDAK	IYA	5,24	94,76	Lulus
AALR	IYA	TIDAK	IYA	5,24	94,76	Lulus
ACA	IYA	TIDAK	IYA	5,24	94,76	Lulus
DAR	IYA	TIDAK	IYA	5,24	94,76	Lulus
MRA	TIDAK	TIDAK	IYA	92,72	7,28	Drop Out
RIH	IYA	IYA	IYA	0,00	100,00	Lulus
MRPD	IYA	TIDAK	IYA	5,24	94,76	Lulus
AFFP	IYA	IYA	IYA	0,00	100,00	Lulus
SRF	TIDAK	TIDAK	IYA	92,72	7,28	Drop Out
MA	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
JF	IYA	IYA	IYA	0,00	100,00	Lulus
ALK	IYA	TIDAK	IYA	5,24	94,76	Lulus
MRP	IYA	TIDAK	IYA	5,24	94,76	Lulus
EAR	IYA	IYA	IYA	0,00	100,00	Lulus
MG	IYA	TIDAK	IYA	5,24	94,76	Lulus
SBK	IYA	TIDAK	IYA	5,24	94,76	Lulus

MR	TIDAK	IYA	TIDAK	100,00	0,00	Drop Out
MFAK	IYA	TIDAK	IYA	5,24	94,76	Lulus
AUH	IYA	IYA	IYA	0,00	100,00	Lulus
SF	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
LKS	IYA	TIDAK	TIDAK	100,00	0,00	Drop Out
LMA	IYA	TIDAK	IYA	5,24	94,76	Lulus
RPPA	IYA	TIDAK	IYA	5,24	94,76	Lulus
NAAM	IYA	TIDAK	TIDAK	100,00	0,00	Drop Out
AR	TIDAK	TIDAK	IYA	92,72	7,28	Drop Out
AS	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
AV	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
NAR	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
AR	IYA	TIDAK	IYA	5,24	94,76	Lulus
RNM	IYA	TIDAK	IYA	5,24	94,76	Lulus
DK	IYA	IYA	IYA	0,00	100,00	Lulus
RAPS	IYA	IYA	IYA	0,00	100,00	Lulus
WFF	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
AR	IYA	TIDAK	IYA	5,24	94,76	Lulus
RV	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
SNR	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
FQ	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
HR	IYA	IYA	IYA	0,00	100,00	Lulus
VP	IYA	IYA	IYA	0,00	100,00	Lulus
FA	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
NS	IYA	IYA	IYA	0,00	100,00	Lulus

MR	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
MA	IYA	IYA	IYA	0,00	100,00	Lulus
AMI	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
RR	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
MH	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
HMA	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
RCW	IYA	TIDAK	IYA	5,24	94,76	Lulus
MR	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
RGP	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
IH	IYA	IYA	IYA	0,00	100,00	Lulus
MQZZ	TIDAK	TIDAK	TIDAK	100,00	0,00	Drop Out
MO	IYA	TIDAK	IYA	5,24	94,76	Lulus
RRSM	TIDAK	TIDAK	IYA	92,72	7,28	Drop Out

**Tabel 4.4** Hasil dari Model Algoritma *Naive Bayes Classification*

## 4.5 *Access*

Tahap *assess* merupakan tahap dimana dilakukan evaluasi dari model penelitian. Hasil evaluasi untuk metode C4.5 dan *Naïve Bayes Classification* (NBC) menggunakan *confusion matrix* yang berisi nilai *accuracy*, *precision*, dan *recall* dari data *test*.

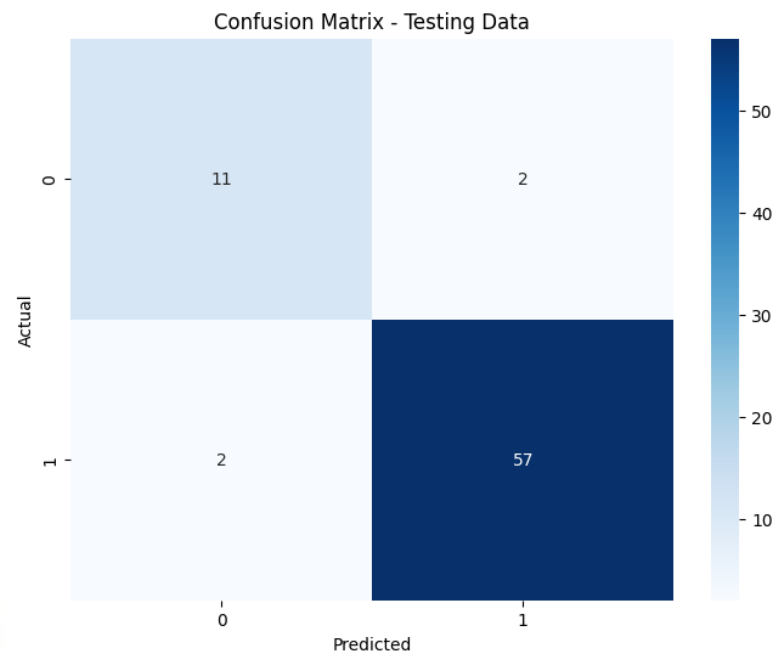
### 4.5.1 C4.5

Pada metode C4.5 digunakan beberapa skenario pembagian data uji seperti yang tertulis pada Tabel 4.1. Data *training* dan data *testing* selanjutnya diolah menggunakan metode C4.5. Hasil akurasi dengan pembagian data set 90% : 10% adalah 94,44% untuk pembagian data 80% : 20% akurasinya adalah 92,30% dan untuk pembagian data 70% : 30% menghasilkan akurasi sebesar 92,55%. Hasil akurasi terbaik adalah pada rasio dataset 90% : 10%. Tabel 4.4 adalah hasil *accuracy*, *precision*, dan *recall* dari beberapa rasio data set.

**Tabel 4.5** Hasil *Accuracy*, *Precision*, dan *Recall* pada metode C4.5

<i>Data Train : Data Test</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>
90 : 10	94,44%	94%	94%
80 : 20	92,30%	92%	92%
70 : 30	92,55%	93%	93%

Hasil evaluasi klasifikasi menggunakan algoritma C4.5 yang ditunjukkan pada Gambar 4.7.



**Gambar 4.8** *Confusion Matrix* pada Algoritma C4.5

Pada Gambar 4.5 didapat hasil evaluasi klasifikasi menggunakan *confussion matrix* dengan jumlah data TP yang dihasilkan sebesar 11 data, jumlah FN sebesar 2 data, FP berjumlah 2 data dan TN berjumlah 57 data. Untuk lebih jelasnya dapat dilihat pada Tabel 4.5.

**Tabel 4.6** *Confusion Matrix* pada Algoritma C4.5

Data Sebenarnya	Klasifikasi	
	<i>Drop Out</i>	Lulus
<i>Drop Out</i>	11 (TP)	2 (FN)
Lulus	2 (FP)	57 (TN)

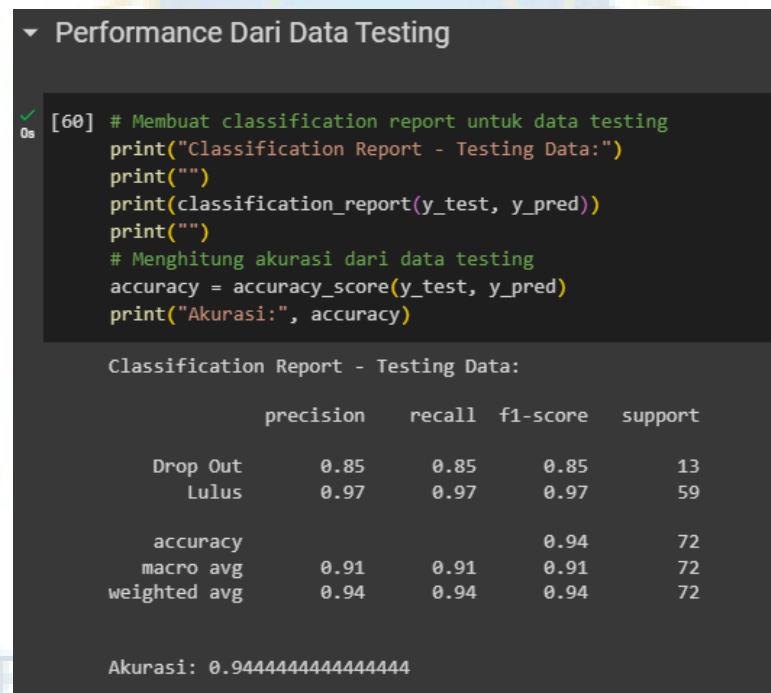
Dari hasil klasifikasi pada Tabel 4.5 dapat dilihat bahwa:

- True Positive* (TP) menjelaskan dimana data terklasifikasi *drop out*, memang benar *drop out*. Dalam hal ini jumlah data yang didapat sebanyak 11 data.



- b. *False Positife* (FP) menjelaskan bahwa data yang terklasifikasi *drop out*, ternyata tidak *drop out* / lulus. Jumlah data yang didapat sebesar 2 data.
- c. *False Negative* (FN) menjelaskan bahwa data yang terklasifikasi lulus, sebenarnya adalah *drop out*. Data yang dihasilkan berjumlah 2 data.
- d. *True Negative* (TN) menjelaskan dimana data yang terklasifikasi lulus, memang benar lulus. Jumlah data yang didapat sebanyak 57 data.

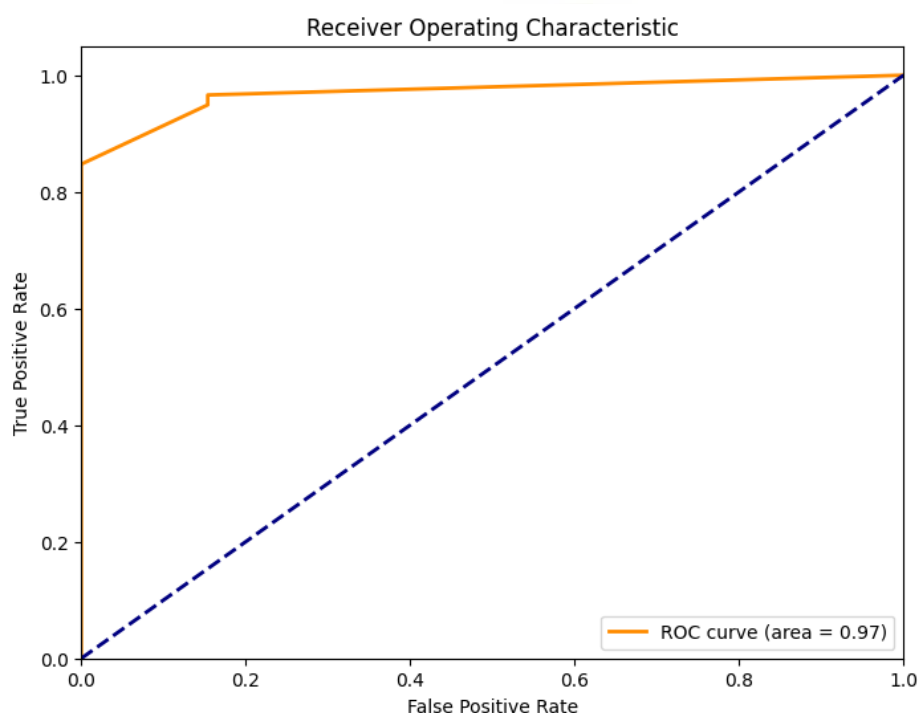
Gambar 4.8 adalah hasil dari *confussion matrix* yang telah diperoleh, dihasilkan nilai *accuracy*, *precision*, *recall* dan *f1-score*.



**Gambar 4.9** Hasil Pengukuran Kinerja menggunakan Algoritma C4.5

Hasil akurasi yang didapat sebesar 94,44% dengan nilai *precision* untuk data *Drop Out* dan Lulus masing-masing sebesar 85% dan 97%. Nilai *recall* dari data *Drop Out* dan Lulus yaitu 85% dan 97%. Dan nilai *f1-score* dari data *Drop Out* dan Lulus yaitu 85% dan 97%. Untuk nilai *precision*, *recall* dan *f1-score* secara

keseluruhan dapat dilihat pada nilai rata-rata macro (*macro average*) dari *precision*, *recall* dan *f1-score*. Nilai rata-rata yang dihasilkan yaitu *precision* sebesar 91%, *recall* sebesar 91% dan *f1-score* sebesar 91%. Gambar 4.8 adalah hasil dari ROC Curve yang telah diproses dengan Algoritma C4.5 dan menghasilkan *score accuracy* sebesar 97%.



**Gambar 4.7** ROC Curve pada Algoritma C4.5

#### 4.5.2 Naïve Bayes Classification

Pada metode *Naïve Bayes Classification* digunakan beberapa skenario pembagian data uji seperti yang tertulis pada Tabel 4.1. Data *training* dan data *testing* selanjutnya diolah menggunakan metode *Naïve Bayes Classification*. Hasil akurasi dengan pembagian data set 90% : 10% adalah 90,27% untuk pembagian data 80% : 20% akurasinya adalah 93,00% dan untuk pembagian data 70% : 30%

menghasilkan akurasi sebesar 92,55%. Hasil akurasi terbaik adalah pada rasio data set 80% : 20%. Tabel 4.6 adalah hasil *accuracy*, *precision*, dan *recall* dari beberapa rasio dataset.

**Tabel 4.7** Hasil *Accuracy*, *Precision*, dan *Recall* pada Algoritma NBC

<i>Data Train : Data Test</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>
90 : 10	90,27%	92%	90%
80 : 20	93,00%	94%	93%
70 : 30	92,55%	94%	93%



**Gambar 4.10** *Confusion Matrix* pada Algoritma NBC

Dari Gambar 4.8 didapat hasil evaluasi klasifikasi menggunakan *confussion matrix* dengan jumlah data TP yang dihasilkan sebesar 20 data, jumlah FN sebesar 8 data, FP berjumlah 2 data dan TN berjumlah 113 data. Untuk lebih jelasnya dapat dilihat pada Tabel 4.7.

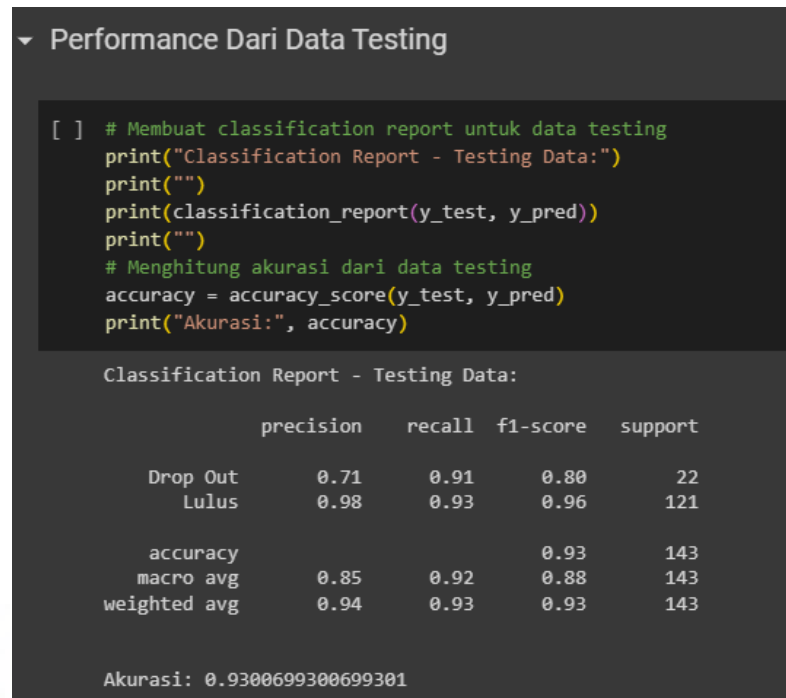
**Tabel 4.8** *Confusion Matrix* pada Algoritma NBC

Data Sebenarnya	Klasifikasi	
	<i>Drop Out</i>	Lulus
<i>Drop Out</i>	20 (TP)	8 (FN)
Lulus	2 (FP)	113 (TN)

Dari hasil klasifikasi pada Tabel 4.7, dapat dilihat bahwa:

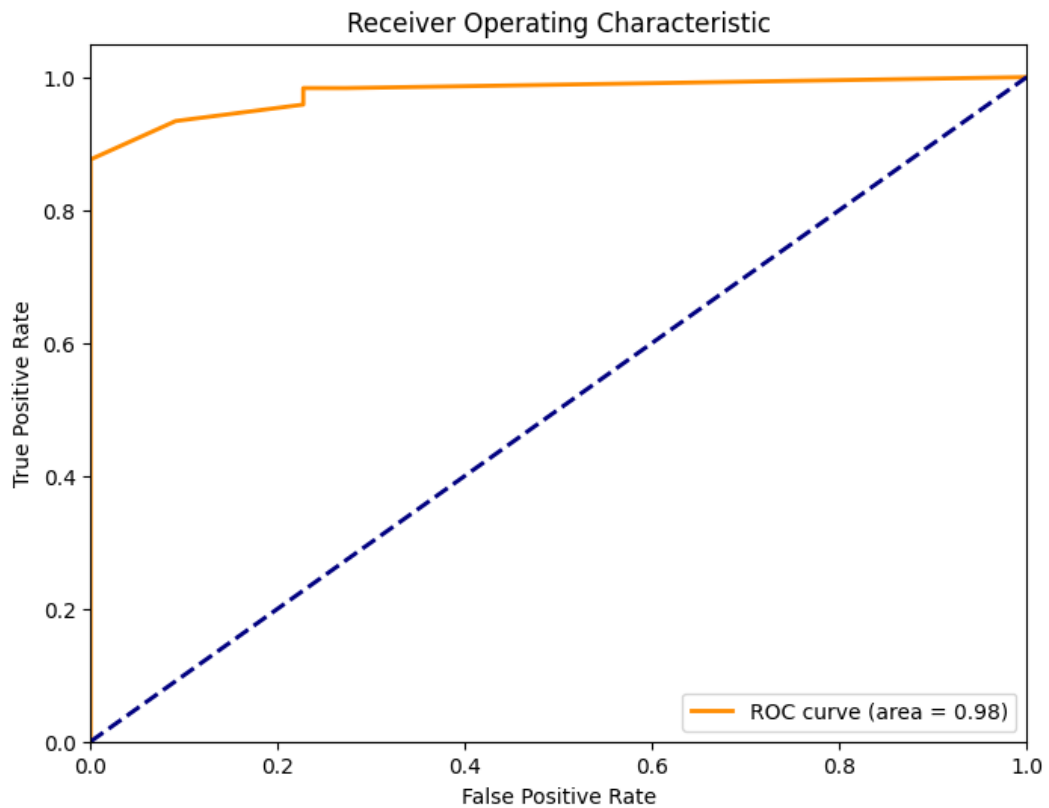
- True Positive* (TP) menjelaskan dimana data terklasifikasi *drop out*, memang benar *drop out*. Dalam hal ini jumlah data yang didapat sebanyak 20 data.
- False Positive* (FP) menjelaskan bahwa data yang terklasifikasi *drop out*, ternyata tidak *drop out* / lulus. Jumlah data yang di dapat sebesar 2 data.
- False Negative* (FN) menjelaskan bahwa data yang terklasifikasi lulus, sebenarnya adalah *drop out*. Data yang dihasilkan berjumlah 8 data.
- True Negative* (TN) menjelaskan dimana data yang terklasifikasi lulus, memang benar lulus. Jumlah data yang didapat sebanyak 113 data.

Gambar 4.9 adalah hasil *confussion matrix* yang telah diperoleh, dihasilkan nilai *accuracy*, *precision*, *recall* dan *f1-score*.



**Gambar 4.11** Hasil Pengukuran menggunakan Algoritma NBC

Berdasarkan Gambar 4.10, hasil akurasi yang didapat sebesar 93% dengan nilai *precision* untuk data *Drop Out* dan Lulus masing-masing sebesar 71% dan 98%. Nilai *recall* dari data *Drop Out* dan Lulus yaitu 91% dan 93%. Dan nilai *f1-score* dari data *Drop Out* dan Lulus yaitu 80% dan 96%. Untuk nilai *precision*, *recall* dan *f1-score* secara keseluruhan dapat dilihat pada nilai rata-rata macro (*macro average*) dari *precision*, *recall* dan *f1-score*. Nilai rata-rata yang dihasilkan yaitu *precision* sebesar 85%, *recall* sebesar 92% dan *f1-score* sebesar 88%. Gambar 4.10 adalah hasil dari ROC Curve yang telah diproses dengan Algoritma *Naïve Bayes Classification* dan menghasilkan *score accuracy* sebesar 98%.



Gambar 4. 12 ROC Curve pada Algoritma NBC

#### 4.5.3 Interpretasi Hasil

Berdasarkan proses pengujian klasifikasi dengan metode C4.5 dan *Naive Bayes Classification* didapatkan hasil akurasi dari kedua metode tersebut pada Tabel 4.8. Hasil dari penelitian menggunakan klasifikasi mahasiswa berpotensi *drop out* menggunakan metode C4.5 mendapatkan akurasi 94,44%. Sementara itu, akurasi yang didapatkan dengan menggunakan metode *Naive Bayes Classification* adalah 93,00%. Hal ini didapatkan karena rasio yang digunakan berbeda, peneliti menggunakan rasio terbaik pada setiap algoritma, didapatkan hasil bahwa algoritma C4.5 menggunakan rasio 90:10, sedangkan algoritma *Naive Bayes Classification* menggunakan rasio 80:20. Dari jumlah rasio bisa dilihat bahwa data *training* pada

algoritma C4.5 lebih banyak dibandingkan dengan algoritma *Naïve Bayes Classification* sehingga didapatkan *accuracy* yang lebih baik.

**Tabel 4.9** Hasil Perbandingan Kinerja

Rasio	C4.5			<i>Naïve Bayes Classification</i>		
	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>
90%:10%	94,44%	94%	94%	90,27%	92%	90%
80%:20%	92,30%	92%	92%	93,00%	94%	93%
70%:30%	92,55%	93%	93%	92,55%	94%	93%

#### 4.6 Evaluasi

Pada tahapan ini, peneliti akan melakukan evaluasi kinerja dan potensi dengan melihat penelitian sebelumnya dengan topik permasalahan yang sama yaitu pendidikan pada perguruan tinggi serta menggunakan Algoritma C4.5 atau *Naïve Bayes Classification* (NBC).

Pada penelitian sebelumnya Sinaga *et al.* (2021) menggunakan 98 data mahasiswa yang terdiri dari mahasiswa tahun masuk 2017-2018 yang telah melewati masa studi 4 semester, pada penelitian ini menggunakan beberapa faktor untuk mengklasifikasi mahasiswa berpotensi *drop out*, diantaranya jenis kelamin, umur, agama, tempat tinggal, IPS, disiplin, dan hutang. Setelah dilakukan pemodelan, penelitian ini mendapatkan hasil *accuracy* sebesar 90%, hasil dari *precision* 87,50 dan hasil dari *recall* sebesar 100%. Penelitian lainnya Iskandar *et al.* (2021) menggunakan 318 data mahasiswa yang terdiri dari mahasiswa tahun masuk 2019 jurusan Matematika, pada penelitian ini menggunakan beberapa faktor untuk mendapatkan *output* mahasiswa yang mendapatkan beasiswa bidikmisi, diantaranya pekerjaan orang tua, penghasilan orang tua, jumlah tanggungan, daya listrik (*watt*), dan nilai ujian nasional. Setelah dilakukan pemodelan, penelitian ini

mendapatkan hasil *accuracy* sebesar 79%. Penelitian milik Rahman *et al.* (2020) menggunakan data mahasiswa tahun lulus 2016 sebanyak 151 data mahasiswa, pada penelitian ini menggunakan beberapa faktor untuk mengklasifikasi kelulusan mahasiswa, di antaranya IPK, SKS, umur, dan jenis kelamin. Setelah dilakukan pemodelan, penelitian ini mendapatkan hasil *accuracy* 88,74%, pada penelitian ini memberikan saran bahwa jumlah *record* data yang digunakan untuk proses *training* ditingkatkan disebabkan jumlah data *training* mempengaruhi nilai *accuracy*.

Dari penelitian sebelumnya, maka peneliti melihat potensi model C4.5 dan *Naïve Bayes Classification* dalam mengklasifikasi mahasiswa berpotensi *drop out* dengan menggunakan 721 data mahasiswa yang terdiri atas 571 data *training* dan 150 data *testing*. Pada penelitian ini hasil yang didapatkan standar *accuracy* lebih tinggi dari penelitian sebelumnya, pada penelitian ini didapatkan hasil *accuracy* untuk algoritma C4.5 adalah 94,44% dan *accuracy* algoritma *Naïve Bayes Classification* sebesar 93%. Faktor-faktor yang digunakan dalam penelitian sebelumnya berbeda dengan faktor-faktor yang peneliti gunakan, sehingga perlu menambahkan beberapa faktor untuk meningkatkan kualitas pada penelitian ini.

Penggunaan model C4.5 dan *Naïve Bayes Classification* diminati untuk proses pengklasifikasian karena menghasilkan nilai *accuracy* yang baik.

Penggunaan bahasa pemrograman lain bisa dilakukan untuk mendapatkan hasil yang berbeda dari penelitian ini, peneliti menyarankan untuk menggunakan bahasa pemrograman R atau Java yang memiliki kemampuan yang baik untuk mengelola data yang besar sehingga dapat dibandingkan dengan bahasa pemrograman Python yang peneliti gunakan dalam penelitian.



## BAB 5

### PENUTUP

#### 5.1 Kesimpulan

Penelitian ini telah melakukan implementasi model klasifikasi mengenai mahasiswa berpotensi *drop out* dan mendapatkan nilai *accuracy*, *precision*, dan *recall* dari metode C4.5 dan *Naïve Bayes Classification*. Berdasarkan pembahasan sebelumnya, maka berikut adalah kesimpulan yang bisa ditarik pada penelitian ini:

- a. Setelah melakukan pencarian melalui studi pustaka dan wawancara, ditemukan bahwa indikator kelulusan mahasiswa Prodi Sistem Informasi, Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta mencakup tiga faktor utama, yaitu kelulusan dalam semua matakuliah, berhasil menyelesaikan laporan Praktik Kerja Lapangan (PKL), serta sedang dalam proses melakukan atau mengerjakan laporan skripsi.
- b. Penelitian ini melakukan klasifikasi mahasiswa berpotensi *drop out* dengan menggunakan metode C4.5 dan mencapai tingkat *accuracy* benar mengklasifikasi mahasiswa tersebut berpotensi lulus atau *drop out* sebesar 94,44%. Di samping itu, menggunakan metode *Naïve Bayes Classification* juga diuji, dan hasil *accuracy* benar mengklasifikasi mahasiswa berpotensi lulus atau *drop out* adalah sebesar 93,00%.
- c. Hasil prediksi total mahasiswa berpotensi *drop out* untuk tahun masuk 2018 menggunakan algoritma C4.5 adalah sebesar 20,67%, yang setara dengan 31 mahasiswa yang berpotensi *drop out*, sementara 79,33% atau 119 mahasiswa berpotensi lulus. Sementara itu, pada penggunaan algoritma *Naïve Bayes*

*Classification*, ditemukan hasil prediksi total mahasiswa sebesar 33,33%, yang mewakili 50 mahasiswa yang berpotensi *drop out*, sedangkan 66,67% atau 100 mahasiswa berpotensi lulus.

- d. Hasil kinerja algoritma C4.5 dalam mengklasifikasi mahasiswa berpotensi *drop out* adalah 94,44% *accuracy*, 94% *precision*, 94% *recall*, dan 97% ROC, sedangkan algoritma *Naïve Bayes Classification* mencapai 93% *accuracy*, 94% *precision*, 93% *recall*, dan 98% ROC.

## 5.2 Saran

Berdasarkan hasil dari penelitian yang dilakukan, peneliti memiliki beberapa saran yang bisa menjadi masukan dan bahan pertimbangan untuk penelitian selanjutnya sebagai berikut:

- a. Penelitian ini dapat dikembangkan dalam sebuah aplikasi/*dashboard* untuk kebutuhan program studi mendapatkan informasi mahasiswa berpotensi *drop out*, agar mahasiswa yang bersangkutan dapat lebih mempersiapkan diri.
- b. Penambahan data *training* untuk menambahkan kualitas dari *accuracy* algoritma C4.5 dan *Naïve Bayes Classification* dalam mengklasifikasi mahasiswa berpotensi *drop out*.

## DAFTAR PUSTAKA

- Alizahh Muhammad, D., Nugroho, A., Radiyah, U., & Gata, W. (2020). Sentimen Analisis Terkait Lockdown pada Sosial Media Twitter. *IJSE-Indonesian Journal on Software Engineering*, 6(2), 223–229.
- Anggreany, M. S. (2020, November 1). *Confusion Matrix*. Bina Nusantara.
- Annur, H. (2018). Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes. *ILKOM Jurnal Ilmiah*, 10(2), 160–165. <https://doi.org/10.33096/ilkom.v10i2.303.160-165>
- Arifin, O., & Sasongko, T. B. (2018). Analisa perbandingan tingkat performansi metode support vector machine dan naïve bayes classifier. *Seminar Nasional Teknologi Informasi Dan Multimedia 2018*, 6(1), 67–72.
- Azevedo, A., & Santos, M. F. (2008). KDD, semma and CRISP-DM: A parallel overview. *MCCSIS'08 - IADIS Multi Conference on Computer Science and Information Systems; Proceedings of Informatics 2008 and Data Mining 2008. MCCSIS*, 182–185.
- Azzahra Nasution, D., Khotimah, H. H., & Chamidah, N. (2019). Perbandingan Normalisasi Data Untuk Klasifikasi Wine Menggunakan Algoritma K-NN. *CESS Journal of Computer Engineering System and Science*, 4(1), 78–82.
- Budiman, I., Muliadi, & Ramadina, R. (2015). Penerapan Fungsi Data Mining Klasifikasi untuk Prediksi Masa Studi Mahasiswa Tepat Waktu pada Sistem Informasi Akademik Perguruan Tinggi. *Jurnal Jupiter*, 7, 39–50.
- Cahyanti, D., Rahmayani, A., & Husniar, S. A. (2020). Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara. *Indonesian Journal of Data and Science*, 1(2), 39–43. <https://doi.org/10.33096/ijodas.v1i2.13>
- David, F., & Defrianto. (2019). VISUALISASI DATA DALAM BENTUK 3 DIMENSI DENGAN MENGGUNAKAN BAHASA PEMROGRAMAN PYTHON. *Seminar Nasional Peranan Iptek Menuju Industri Masa Depan (PIMIMD-5)*, 1–6. <https://doi.org/10.21063/PIMIMD5.2019.1>
- Fadhilah, S., Arifin, Z., & Suandi, E. (2018). Pedoman Akademik 2018 UIN Jakarta. In *Pedoman Akademik 2018 UIN Jakarta*. UIN Syarif Hidayatullah Jakarta.
- Fatma Ayu Rahman, A., Wartulas, S., Raya Pagojengan, J. K., & Brebes, P. (2020). Prediksi Kelulusan Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus Di Universitas Peradaban). *Ade Fatma Ayu Rahman IJIR*, 1(2), 70–77.
- Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques*.

- Hermanto, K., Salim, D., Wu, B., Salim, O. R., & Gunadi, R. B. (2023). Penggunaan Python Untuk Menganalisis Pola Penyebaran Covid-19 Di Masa Pandemi. *Journal of Student Development Information System (JoSDIS)* , 3, 62–75.
- Indriani, A. (2014). Klasifikasi Data Forum dengan menggunakan Metode Naïve Bayes Classifier. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI) Yogyakarta*, 1(1), 21–2014. [www.bluefame.com](http://www.bluefame.com),
- Iskandar, J. W, Nataliani, Y. (2021). Perbandingan Naïve Bayes, SVM, dan k-NN untuk Analisis Sentimen Gadget Berbasis Aspek. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 5(6), 1120–1126. <https://doi.org/10.29207/resti.v5i6.3588>
- Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data* (D. T. Larose (ed.); Second). John Wiley & Sons.
- Lee Shin-Jye, Xu Zhaozhao, Li Tong, & Yang Yun. (2018). A Novel Bagging C4.5 Algorithm Based on Wrapper Feature Selection for Supporting Wise Clinical Decision Making. *Biomedical Informatics*, 78, 144–155.
- Melita, R., Amrizal, V., Suseno, H. B., & Dirjam, T. (2018). Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim). *Jurnal Teknik Informatika*, 11(2), 149–164. <https://doi.org/10.15408/jti.v11i2.8623>
- Miftah, S. (2021, July 8). *Belajar Python Mengenal Pandas dan Series untuk Meningkatkan Kompetensi Data*. DQLab.
- Mubarok, M. I. (2018, August 14). *Algoritma C4.5*. MIM.
- Manullang, N., Sembiring, R. W., Gunawan, I., Parlina, I., & Irawan. (2021). Implementasi Teknik Data Mining untuk Prediksi Peminatan Jurusan Siswa Menggunakan Algoritma C4.5. *JURNAL ILMU KOMPUTER DAN TEKNOLOGI*, 2(2), 1–5. <http://creativecommons.org/licenses/by/4.0/>
- Muhathir, M., & Santoso, M. H. (2020). Analysis Naïve Bayes In Classifying Fruit by Utilizing Hog Feature Extraction. *JITE (Journal of Informatics and Telecommunication Engineering)*, 4(1), 250–259. <https://doi.org/10.31289/jite.v4i1.3860>
- Munawir, M., & Iqbal, T. (2019). Prediksi Kelulusan Mahasiswa menggunakan Algoritma Naive Bayes (Studi Kasus 5 PTS di Banda Aceh). In *Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)* (Vol. 3, Issue 2, p. 59). <https://doi.org/10.35870/jtik.v3i2.77>

- Nu'man, H. S., Sofyan, Y., & Tahtawi, A. R. (2020). Pengendalian Robot Lengan Pemilah Benda Berdasarkan Bentuk Menggunakan Teknologi Computer Vision. *SEMNASTERA (Seminar Nasional Teknologi Dan Riset Terapan)*, 42–48.
- Rahman, A. F. A., Sorikhi, & Wartulas, S. (2020). *Prediksi Kelulusan Mahasiswa menggunakan Algoritma C4.5 dengan Studi Kasus Universitas Peradaban*. Universitas Peradaban.
- Santoso, A., & Ariyanto, G. (2007). Implementasi Deep Learning Berbasis Keras Untuk Pengenalan Wajah. *Jurnal Teknik Elektro*, 18(01), 15–21. <https://www.mathworks.com/discovery/convol>
- Septiani, W. D. (2017). Komparasi Metode Klasifikasi Data Mining Algoritma C4.5 dan Naive Bayes Untuk Prediksi Penyakit Hepatitis. *Jurnal Pilar Nusa Mandiri*, 13(1), 76–84. <http://archive.ics.uci.edu/ml/>.
- Setiawan Rony. (2021, October 30). *Apa itu Data Mining dan Bagaimana Metodenya?* Dicoding.
- Sinaga, D., Solaiman, E. J., & Kaunang, F. J. (2021). Penerapan Algoritma Decision Tree C4.5 Untuk Klasifikasi Mahasiswa Berpotensi Drop out Di Universitas Advent Indonesia. *TeIka*, 11(2), 167–173. <https://doi.org/10.36342/teika.v11i2.2613>
- Sudriyanto, S., Rizaldi, R., & Hariri, M. A. R. (2021). Implementasi Particle Swarm Optimization (PSO) untuk Optimisasi Algoritma Naive Bayes dalam Memprediksi Mahasiswa Lulus Tepat Waktu. *COREAI: Jurnal Kecerdasan Buatan, Komputasi Dan Teknologi Informasi*, 2(1), 62–68. <https://www.ejournal.unuja.ac.id/index.php/core/article/view/2181>
- Suherman, & Muzaky, I. (2018). Analisis Penjualan Barang Laris Dan Kurang Laris Terhadap Percetakan Awfa Digitl Printing Menggunakan Metode Decision Tree Dengan Optimasi Algoritma Genetika. *SIGMA - Jurnal Teknologi Pelita Bangsa*, 10, 118–130.
- Wang, X., Zhou, C., & Xu, X. (2019). Application of C4.5 decision tree for scholarship evaluations. *Procedia Computer Science*, 151, 179–184. <https://doi.org/10.1016/j.procs.2019.04.027>
- Widianto, M. H. (2019). *Algoritma Naive Bayes*. Bina Nusantara.
- Wilson, K. (2021, September 30). *Proses Data Mining SEMMA*. Bina Nusantara.
- Wiratmaja I Gede Harjumawan, Wijaya I Wayan Sukerta, Pramana I Dewa Agung, & Aditya I Komang Gede Ryan Aditya. (2021). Program Menghitung Banyak Bata pada Ruangan Menggunakan Bahasa Python. *TIERS Information Technology*, 2, 12–22.

Yuniarti, W. D., Faiz, A. N., & Setiawan, B. (2020). Identifikasi Potensi Keberhasilan Studi Menggunakan Naïve Bayes Classifier. *Walisongo Journal of Information Technology*, 2(1), 1. <https://doi.org/10.21580/wjit.2020.2.1.5204>



Universitas Islam Negeri  
AR-RANIRY  
JALAN KH. HUSAIN RANIRY  
KOTA AR-RANIRY, KABUPATEN AR-RANIRY, PROVINSI SUMATERA UTARA 22111

## LAMPIRAN

### Source Code

```
[ ] # Import Library
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns

from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import classification_report, accuracy_score, confusion_matrix
from sklearn.metrics import roc_curve, auc

from sklearn.metrics import roc_auc_score, roc_curve, auc

import pickle
```

```
# Import Data
dataset = pd.read_csv("DataSetSkripsi.csv", sep=',')
dataset
```

```
# Melihat Info Data
dataset.info()
```

```
# Melihat Balanced Label
dataset.STATUS.value_counts()
```

```
[ ] # Membuat objek LabelEncoder
label_encoder = LabelEncoder()

# Melakukan encoding pada semua fitur kecuali kolom "STATUS"
for column in dataset.columns[1:]:
    dataset[column] = label_encoder.fit_transform(dataset[column])
```

```
# Cek dataset yang sudah di encode
dataset
```

```
[ ] # Memisahkan fitur dan target
X = dataset.drop('STATUS', axis=1)
y = dataset['STATUS']

# Memisahkan data menjadi data latih dan data uji
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Melihat Jumlah Dataset Training dan Data Testing
X_train.shape, X_test.shape, y_train.shape, y_test.shape # Data Testing berjumlah 143 dan Jumlah Data Training berjumlah 571
```

```
[ ] # Membuat objek klasifikasi Naive Bayes
naive_bayes = GaussianNB()

# Melatih model dengan data latih
naive_bayes.fit(X_train, y_train)
```

```
▶ # Melakukan prediksi terhadap data uji
y_pred = naive_bayes.predict(X_test)
y_pred
```

```
▶ # Menghitung confusion matrix untuk data training
cm_train = confusion_matrix(y_train, y_train_pred)

# Membuat grafik confusion matrix untuk data training
plt.figure(figsize=(8, 6))
sns.heatmap(cm_train, annot=True, fmt='d', cmap='Blues')
plt.title('Confusion Matrix - Training Data')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()
```

```
▶ # Menghitung probabilitas prediksi untuk data pengujian
probs = naive_bayes.predict_proba(X_test)

# Probabilitas kelas positif (misalnya, "Lulus")
positive_probs = probs[:, 1]

# Menghitung skor AUC
auc_score = roc_auc_score(y_test, positive_probs)

# Mencetak skor AUC dengan 2 angka di belakang koma
print("Skor AUC: {:.2f}".format(auc_score))
```



## Wawancara

### Diskusi Skripsi Hamzah Aji Prat... [Leave](#)

Waiting for the host to start this meeting

**Meeting ID:** 881 1065 7738

**Date:** Sat, November 12

**Time:** 08:00

Transkrip wawancara dengan Sekretaris Prodi Sistem Informasi Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta periode 2015-2019.

Hari dan Tanggal : Sabtu, 12 November 2022

Waktu : 08.00

Lokasi : Google Meet

Narasumber : Meinarini Catur Utami, M.T.

Waktu : 45 menit

Hasil Wawancara :

Wawancara yang dilakukan adalah untuk mengetahui apa saja kriteria mahasiswa berpotensi *drop out* dari pihak prodi Sistem Informasi, Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta, diketahui bahwa kriteria mahasiswa berpotensi *drop out* diantaranya adalah:

1. Apakah mahasiswa saat semester 8 sudah menyelesaikan semua mata kuliah yang ada?
2. Apakah laporan PKL nya sudah diselesaikan?
3. Apakah laporan Skripsi nya sudah mulai dikerjakan?