Bidang Ilmu : Ilmu Komputer

LAPORAN PENELITIAN

Implementasi Data Mining Untuk Prediksi Kelulusan Tepat Waktu Mahasiswa STIMIK ESQ Menggunakan Decision Tree C4.5

Tim Peneliti:

Ketua : Desy Komalasari

Anggota : Mita Nurul Yatimah

PROGRAM STUDI ILMU KOMPUTER

LEMBAGA PENELITIAN DAN PENGABDIAN KEPADA MASYARAKAT
SEKOLAH TINGGI ILMU MANAJEMEN DAN ILMU KOMPUTER
(STIMIK ESQ)

2022

PENGESAHAN

1. Judul Penelitian : Implementasi Data Mining Untuk Prediksi Kelulusan Tepat Waktu Mahasiswa STIMIK ESQ Menggunakan Decision Tree C4.5
2. Peneliti :
   a. Nama Lengkap : Desy Komalasari
   b. Jenis Kelamin : Perempuan
   c. NIP/NIDN : 0322129202
   d. Jabatan Struktural : Dosen Tetap
   e. Jabatan fungsional : Tenaga Pengajar
   f. Pangkat / Golongan : -
   g. Fakultas/Program Studi : Ilmu Komputer
   h. Pusat Penelitian : STIMIK ESQ
   i. Alamat Institusi : Menara 165 Lt.18-19. Jl. TB Simatupang Kav 1 Cilandak
   j. Telpon/Faks/E-mail :
3. Jangka Waktu Penelitian : 6 bulan (1 semester)
4. Pembiayaan
   a. Jumlah biaya yang diajukan ke STIMIK ESQ : Rp. 3.000.000,00

Jakarta, 5 Februari 2022

Mengetahui,

Ketua Program Studi Ilmu Komputer	Ketua Peneliti
Ahlijati Nuraminah, S.Kom., M.T.I.	Desy Komalasari, S.Kom., M.Kom,
NIDN: 0317128404	NIDN: 0322129202

Kepala LPPM

Danang Indrajaya, S.Si., M.Si
NIDN: 0311118108

IDENTITAS PENELITIAN

1. Judul Penelitian : Implementasi Data Mining Untuk Prediksi Kelulusan Tepat Waktu Mahasiswa STIMIK ESQ Menggunakan Decision Tree C4.5
2. Peneliti

a. Nama Lengkap : Desy Komalasari
b. NIP/NIK : -
c. NIDN : 0322129202
d. Pangkat / Golongan : -
e. Jabatan Fungsional : Asisten Ahli
f. Fakultas/Prodi : Ilmu Komputer
g. Pusat Penelitian : LP2M – Menara 165 Lt.18-19
h. Alamat Institusi : Jl. TB Simatupang Kav.1 Cilandak Jakarta Selatan
i. Telpon/Faks/E-mail :

3. Anggota Peneliti :

NO	NAMA	KEAHLIAN	ALOKASI WAKTU
1	Mita Nurul Yatimah	Data Mining	3 bulan

4. Objek Penelitian : LMS STIMIK ESQ
5. Masa Penelitian

   Mulai : Oktober 2021
   Berakhir : Februari 2022

6. Anggaran yang diusulkan

   Anggran yang diusulkan : Rp. 3.000.000,-

7. Lokasi Penelitian : STIMIK ESQ
8. Hasil yang ditargetkan (temuan baru/paket teknologi/hasil lain), beri penjelasan :
9. Institusi lain yang terlibat : -

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Judul : Implementasi *Data mining* Untuk Prediksi Kelulusan Mahasiswa STIMIK ESQ Menggunakan *Decision Tree* C4.5

Kelulusan mahasiswa merupakan isu penting yang sering dibahas di perguruan tinggi. Hal ini berkaitan dengan penilaian akreditasi. Kualitas perguruan tinggi sangat dipengaruhi oleh proses penilaian akreditasi. Salah satu komponen penilaian akreditasi adalah kelulusan mahasiswa. Semakin banyak lulusan tepat waktu (empat tahun) maka akan semakin baik pula penilaian terhadap perguruan tinggi tersebut. Untuk itu penelitian ini akan melakukan prediksi secara dini kepada mahasiswa untuk melihat potensi kelulusan tepat waktu. Pada penelitian ini akan dilakukan di kampus STIMIK ESQ yang merupakan salah satu perguruan tingi yang ada di Jakarta. Hasil dari penelitian ini akan sangat bermanfaat bagi civitas akademik terutama para dosen untuk memperoleh hasil analisis yang sifatnya objektif, cepat dan terotomatisasi. Selain itu, hasil analisis ini dapat menjadi suatu informasi pendukung dalam memberikan penanganan-penanganan khusus terhadap mahasiswa. Penelitian ini menggunakan *Decision Tree* C4.5 dalam melakukan prediksi tingkat kelulusan. Hasil akurasi yang diberikan penelitian ini sebesar 90% dengan menggunakan parameter jenis kelamin, usia, prodi, IPS1, SKS1, IPK1, IPS2, SKS2, IPK2, IPS3, SKS3, IPK3, IPS4, SKS4, IPK4, dan masa studi.

Kata kunci : *Prediksi, kelulusan, Decision Tree* C4.5

BAB 1
PENDAHULUAN

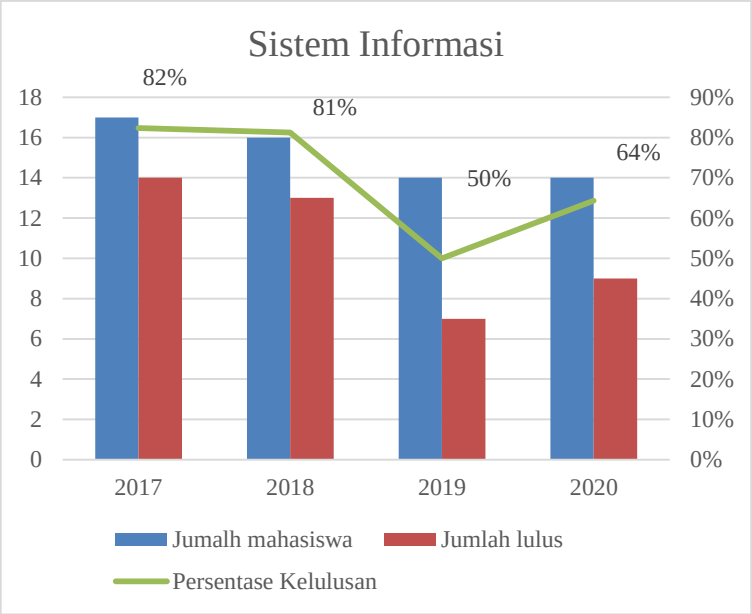
Pada bab ini berisi latar belakang penelitian, identifikasi masalah, rumusan masalah, tujuan penelitian, manfaat penelitian bagi beberapa pihak dan ruang lingkup penelitian.

1.1 Latar Belakang

Perkembangan dunia pendidikan di Indonesia telah memberikan dampak persaingan yang sangat ketat. Hal ini dipicu akibat semakin majunya pendidikan di perguruan tinggi. Salah satu dampak dari persaingan yaitu menghasilkan lulusan yang berkualitas. Adapun kriteria lulusan yang berkualitas diantaranya mampu menyelesaikan masa pembelajaran tepat waktu.

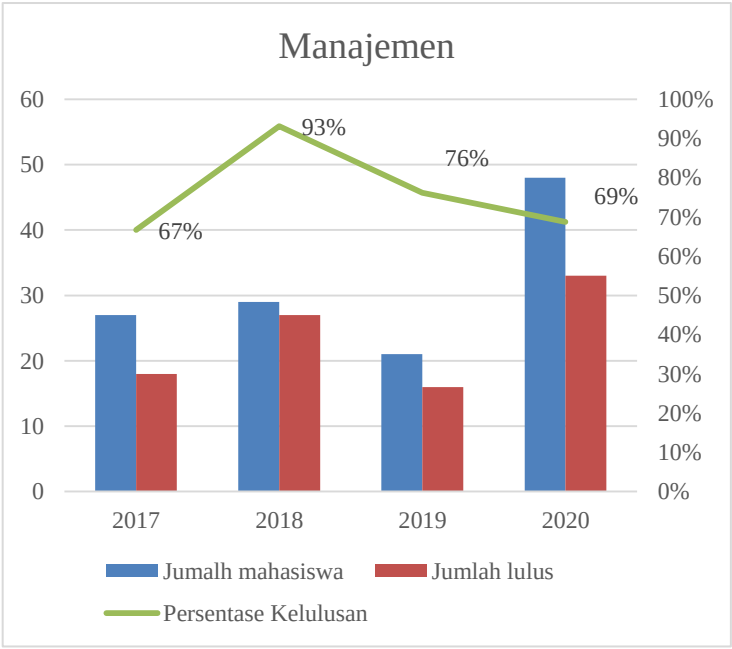
Masa pembelajaran tepat waktu sangat mempengaruhi kualitas dari perguruan tinggi. Kemampuan perguruan tinggi menghasilkan lulusan yang mampu menyelesaikan masa pembelajaran tepat waktu merupakan faktor yang mempengaruhi akreditasi perguruan tinggi. Hal ini sesuai dengan peraturan Badan Akreditasi Nasional Perguruan Tinggi Nomor 3 tahun 2019 tentang Instrumen Akreditasi Perguruan Tinggi yang menyatakan bahwa salah satu indikator penilaian akreditasi adalah persentase lulusan tepat waktu untuk setiap program dari perguruan tinggi (Badan Akreditasi Nasional Perguruan Tinggi, 2019). Untuk itu sangat penting mencari faktor yang dapat mempengaruhi kelulusan tepat waktu di perguruan tinggi.

Studi kasus pada penelitian ini adalah STIMIK (Sekolah Tinggi Ilmu Manajemen Ilmu Komputer) ESQ yang berada di Jakarta Selatan. Adapun yang menjadi objek pada penelitian ini merupakan mahasiswa dari prodi manajemen dan sistem informasi. Gambar 1.1 di bawah merupakan grafik kelulusan mahasiswa program studi informasi tahun 2017 sampai dengan 2020.



Gambar 1.1 Grafik Kelulusan Prodi Sistem Informasi

Gambar 1.2 di bawah merupakan grafik kelulusan mahasiswa program studi manajemen tahun 2017 sampai dengan 2020.



Gambar 1.2 Grafik Kelulusan Prodi Manajemen

Merujuk pada Gambar 1.1 dan Gambar 1.2 terlihat bahwa adanya pola kelulusan yang mengalami penurunan ataupun kenaikan yang sangat signifikan.

Hal ini terlihat pada prodi Sistem Informasi pada tahun 2018-2019 yang mengalami penurunan drastis sebanyak 31%. Hal ini juga terjadi pada prodi Manajemen pada tahun 2017-2018 yang menunjukan adanya kenaikan yang signifikan sebanyak 26%. Namun selanjutnya mengalami penurunan terus menerus hingga tahun 2020. Dengan demikian penelitian ini akan melakukan analisis terhadap faktor-faktor yang mempengaruhi kelulusan tepat waktu dengan menerapkan *data mining* untuk proses pengolahan data secara otomatis.

Analisis dan prediksi diharapkan mampu menemukan faktor-faktor yang mempengaruhi dan memprediksi kelulusan tepat waktu. Sehingga dapat dilakukan prediksi kelulusan mahasiswa lebih dini. Manfaat lainnya yaitu untuk memonitoring mahasiswa, menunjang data akreditasi dan meningkatkan sistem perguruan tinggi yang lebih terintegrasi.Data yang digunakan pada penelitian ini adalah data kelulusan mahasiswa STIMIK ESQ tahun 2017-2020 pada prodi Sistem Informasi dan Manajemen sebanyak 86 data. Pada penelitian ini, peneliti fokus pada faktor yang mempengaruhi kelulusan dari semester 1 sampai dengan semester 4 agar prediksi dapat dilakukan lebih dini. Pemilihan faktor tersebut mengacu pada Standar Nasional Pendidikan Tinggi (SN DIKTI).

Penelitian terhadap kelulusan mahasiswa telah banyak dilakukan, diantaranya penelitian yang berjudul "Prediksi Kelulusan Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus di Universitas Peradaban)" (Rahman dkk., 2020). Penelitian ini menggunakan metode *Decision Tree* C4.5, data yang digunakan adalah data kelulusan mahasiswa universitas Peradaban dengan atribut IPK, SKS, Umur, dan Jenis Kelamin. Tujuan dari penelitian tersebut untuk mengetahui tingkat akurasi algoritma C4.5. Hasil penelitian tersebut menunjukan bahwa algoritma C4.5 dapat memprediksi kelulusan mahasiswa Universitas Peradaban dengan tingkat akurasi 88,74%, presisi 91,7%, dan *recall* sebesar 95,34%.

Penelitian selanjutnya dengan judul "Prediksi Ketepatan Kelulusan Mahasiswa Diploma dengan Komparasi Algoritma Klasifikasi" (M. S. Maulana dkk., 2019). Penelitian ini mengkomparasi 5 algoritma *data mining* yaitu *Decision Tree* C4.5*, Naive Bayes,* K-NN, *rule Induction,* dan *random forest* dengan pengujian *T-test.* Tujuan dari penelitian ini adalah menentukan metode yang

paling akurat untuk menentukan ketepatan kelulusan mahasiswa. Data yang digunakan pada penelitian ini adalah data mahasiswa AMIK BSI Pontianak angkatan 2013/2017 prodi Informatika sebanyak 394 *record.* Atribut yang digunakan yaitu Nama, Jenis Kelamin, Pekerjaan Mahasiswa, Umur, Status Nikah, IPS 1-6, IPK, dan status kelulusan. Hasil dari penelitian ini didapatkan algoritma yang paling optimal dalam menentukan ketepatan kelulusan mahasiswa diploma AMIK BSI Pontianak adalah *Decision Tree* C4.5 dengan akurasi sebesar 90,85%.

Berdasarkan beberapa referensi penelitian terdahulu, *Decision Tree* memiliki nilai akurasi yang paling tinggi dibandingkan dengan metode *data mining* lainnya. Maka pada penelitian ini peneliti mencoba mencari faktor yang mempengaruhi kelulusan tepat waktu mahasiswa STIMIK ESQ menggunakan metode *Decision Tree*. Dengan menggunakan metode tersebut diharapkan mampu menemukan faktor yang mempengaruhi masa pembelajaran tepat waktu. Penerapan metode tersebut juga diharapkan mampu menghasilkan akurasi yang baik.

1.2 Identifikasi Masalah

Dari pernyataan yang telah diuraikan di latar belakang, dapat di identifikasi masalah-masalah sebagai berikut:

1. Tidak diketahuinya faktor yang mempengaruhi kelulusan tepat waktu di STIMIK ESQ
2. Belum ada pendekatan untuk prediksi kelulusan mahasiswa tepat waktu di STIMIK ESQ.

1.3 Rumusan Masalah

Berdasarkan identifikasi masalah yang telah peneliti pilih di atas maka dapat dirumuskan permasalahan penelitian sebagai berikut:

1. Faktor apa saja yang mempengaruhi kelulusan tepat waktu di STIMIK ESQ?
2. Bagaimana memprediksi kelulusan tepat waktu sedari dini di STIMIK ESQ dengan penerapan *data mining* ?

1.4 Tujuan Penelitian

Mengacu pada batasan masalah yang telah diuraikan di atas, maka tujuan penelitian ini adalah sebagai berikut:

1. Untuk mengetahui faktor yang mempengaruhi kelulusan di STIMIK ESQ
2. Untuk menerapkan *data mining* dalam memprediksi kelulusan tepat waktu sedari dini di STIMIK ESQ.

1.5 **Manfaat** Penelitian

Adapun manfaat dari penelitian ini terbagi kedalam tiga bagian yaitu manfaat bagi peneliti, pihak STIMIK ESQ, dan bagi akademisi.

1. Bagi Peneliti : Untuk memperoleh pengetahuan mengenai faktor yang mempengaruhi kelulusan tepat waktu di STIMIK ESQ dan penerapan *data mining* dalam prediksi kelulusan tepat waktu di STIMIK ESQ.
2. Bagi pihak STIMIK ESQ : Penelitian ini nantinya dapat digunakan sebagai hipotesis dalam memberikan suatu perlakuan kepada mahasiswa agar mahasiswa dapat lulus tepat waktu.

   Bagi akademisi : Penelitian ini juga diharapkan dapat menambah referensi bagi penelitian yang akan dilakukan selanjutnya.

1.6 Batasan Masalah

Batasan masalah digunakan untuk menghindari adanya pelebaran serta penyimpangan pokok masalah agar penelitian yang dilakukan lebih terarah dan memudahkan dalam pembahasan sehingga tujuan dari penelitian yang dilakukan dapat tercapai. Adapun batasan masalah dari penelitian ini adalah sebagai berikut :

1. Data penelitian yang digunakan hanya data mahasiswa lulusan tahun 2017 – 2020 prodi Sistem Informasi dan Manajemen
2. Data mahasiswa yang dikategorikan menjadi mahasiswa lulus tepat waktu adalah mampu menyelesaikan masa studi kurang atau sama dengan 4 tahun dan mahasiswa yang terlambat menyelesaikan masa studi adalah mahasiswa yang lulus lebih dari empat tahun.

   Standar kelulusan yang digunakan mengacu pada Standar Nasional Pendidikan Tinggi (SN DIKTI).

BAB 2
TINJAUAN PUSTAKA

2.1 Landasan Teori

Pada sub bab ini berisi landasan-landasan teori yang akan digunakan dalam penelitian ini yaitu, *Data Mining, Decision Tree, Regresi Logistik Biner,* tabel *krejcie*-Morgan, *Stratified Random Sampling,* dan *Confusion Matrix.*

2.1.1 *Data mining*

Data mining merupakan sebuah proses yang memanfaatkan ilmu statistik, *Artificial Intelligence* (AI) atau kecerdasan buatan, matematika, dan *machine learning* untuk mengidentifikasi dan mengekstraksi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar (Romadhona, A., Suprapedi , S. dan Himawan, 2017). *Data mining* menurut Nurdin dan Astika (Nurdin & Astika, 2015) *data mining* merupakan proses mengekstraksi informasi atau sesuatu yang menarik dari data yang berada di database, sehingga mampu menghasilkan suatu informasi yang sangat penting atau berharga. Pengertian tersebut juga sejalan dengan pengertian menurut Fajrin dan Maulana (A. Maulana & Fajrin, 2018) yang mendefinisikan bahwa *data mining* sebagai ekstraksi informasi potensial, implisit, dan tidak dikenal dari sekumpulan data kemudian mengubah hasilnya secara akurat menjadi informasi yang mudah dipahami (A. Maulana & Fajrin, 2018).

Data mining mewarisi banyak aspek dan teknik dari bidang-bidang ilmu yang sudah mapan terlebih dahulu, *data mining* bertujuan untuk memperbaiki teknik tradisional sehingga mampu menangani (Fajrin dkk., 2018):

1. Jumlah data yang sangat besar,
2. Data yang sangat heterogen,
3. Data yang memiliki perbedaan sifat dan data berdimensi tinggi.

Data mining dapat mengidentifikasi pola-pola yang valid, dapat dipahami secara mudah dan berpotensi memberikan manfaat (Firdaus, 2017). Serta berguna untuk penggalian nilai tambah dari sekumpulan data berupa pengetahuan yang tidak diketahui secara manual. Sehingga *data mining* mampu melakukan pengelompokan data seperti (Heryana, 2019):

1. Klasifikasi

   Klasifikasi adalah suatu teknik pengelompokan dengan target tertentu dalam bentuk kategori. Misalnya, IPK yang dapat dipisahkan dalam 3 kategori yaitu, memuaskan, cumlaude, dan summa cumlaude.

2. Klasterisasi

   klasterisasi merupakan suatu teknik pengelompokan, pengamatan atau memperhatikan data berdasarkan kemiripan tertentu antar objek yang kemudian disatukan dalam suatu kelompok tertentu.

3. Prediksi

   Sama dengan klasifikasi dan klasterisasi perbedaannya terletak pada nilai yang dihasilkan. Prediksi menghasilkan nilai yang akan ada pada masa mendatang atau dapat dikatakan sebagai suatu teknik memperkirakan nilai yang belum diketahui untuk masa yang akan datang.

4. Estimasi

   Mirip dengan klasifikasi namun variabel target lebih kearah numerik dari pada kategori.

5. Asosiasi

   Tugas dari asosiasi dalam *data mining* adalah untuk menemukan variabel yang muncul dalam satu waktu. Dapat dikatakan juga sebagai proses identifikasi hubungan antar berbagai peristiwa yang terjadi pada suatu waktu.

6. Deskripsi

   Data mining mampu menggambarkan atau menjelaskan pola-pola yang terdapat dalam data sehingga data dapat dianalisis lebih lanjut.

Menurut Rahayu dan kawan-kawan (Rahayu dkk., 2019) *data mining* memiliki tujuh tahapan, yaitu:

1. Data *Cleaning*

   Data *cleaning* merupakan tahapan pembersihan data dari yang tidak konsisten atau data yang tidak relevan, pada umumnya data mentah memiliki nilai-nilai yang hilang atau tidak valid. Pembersihan data sangat mempengaruhi performa dari *data mining*.

2. *Data Integration*

   Pada tahap ini dilakukan penggabungan data dari beberapa database ke dalam suatu database baru. Integrasi data perlu dilakukan dengan penuh ketelitian karena kesalahan pada proses integrasi data dapat menghasilkan hasil yang tidak relevan, menyimpang bahkan dapat menyesatkan pengambilan keputusan.

3. Seleksi Data (*Data Selection*)

   Data mentah yang didapatkan dari database seringkali hanya dipilih beberapa data memiliki pengaruh. Oleh karena itu hanya data yang sesuai untuk dianalisis yang dipilih dari database. Sebagai contoh, analisis faktor yang mempengaruhi kelulusan tepat waktu mahasiswa maka tidak perlu mengambil data NIM dan nama mahasiswa, cukup dengan data akademis seperti IPK dan IPS.

4. Transformasi Data (*Data Transformation*)

   Data diubah ke dalam format yang sesuai atau digabung berdasarkan ketentuan yang berlaku. Beberapa teknik *data mining* hanya mampu memproses data kategori seperti analisis asosiasi dan *clustering*. Oleh karena itu data dengan jenis numerik perlu dibagi-bagi ke dalam beberapa interval. Sebagai contoh data IPK dan IPS mahasiswa yang memiliki nilai numerik dari angka 0 sampai dengan 4 dapat dibagi ke dalam beberapa interval seperti IPK <2.00 = 0 dan IPK >=2.00 = 1.

5. Proses Penambangan (*Process Mining*)

   Proses penambangan merupakan proses yang paling utama saat menerapkan metode untuk menggali pengetahuan berharga yang tersembunyi

dalam data. Pada tahap ini akan diketahui berbagai informasi yang terdapat dalam data sesuai dengan tujuan yang diinginkan. Sebagai contoh prediksi kelulusan mahasiswa, maka pada tahap ini akan diketahui pola prediksi mahasiswa yang dapat lulus tepat waktu.

6. Evaluasi Pola (*pattern evaluation*)

   Pada tahap ini akan dilakukan evaluasi terhadap pola yang dihasilkan dari tahap penambangan (*process mining).* Proses ini akan menghasilkan nilai ketepatan atau akurasi dari pola yang dihasilkan dengan tujuan apakah penambangan yang dilakukan memang tercapai.

7. Interpretasi pengetahuan

   Bagaimana merepresentasikan informasi yang didapatkan ke dalam bentuk pengetahuan yang dapat dimengerti oleh semua orang. Penyajian pengetahuan dapat berupa visualisasi untuk membantu mengkomunikasikan hasil dari *data mining*.

2.1.2 *Decision Tree* C4.5

Decision Tree merupakan salah satu metode klasifikasi yang menggunakan representasi struktur pohon yang berisi alternatif-alternatif pemecahan dari suatu permasalahan (Hamidah dkk., 2019). Pengertian tersebut sejalan dengan pengertian *Decision Tree* menurut Kamal dan kawan-kawan (Kamal dkk., 2017) yang menyatakan bahwa *Decision Tree* merupakan salah satu metode klasifikasi yang direpresentasikan dengan struktur pohon (*Tree)* atribut direpresentasikan oleh node, nilai dari atribut direpresentasikan oleh cabang, kelas direpresentasikan oleh daun.

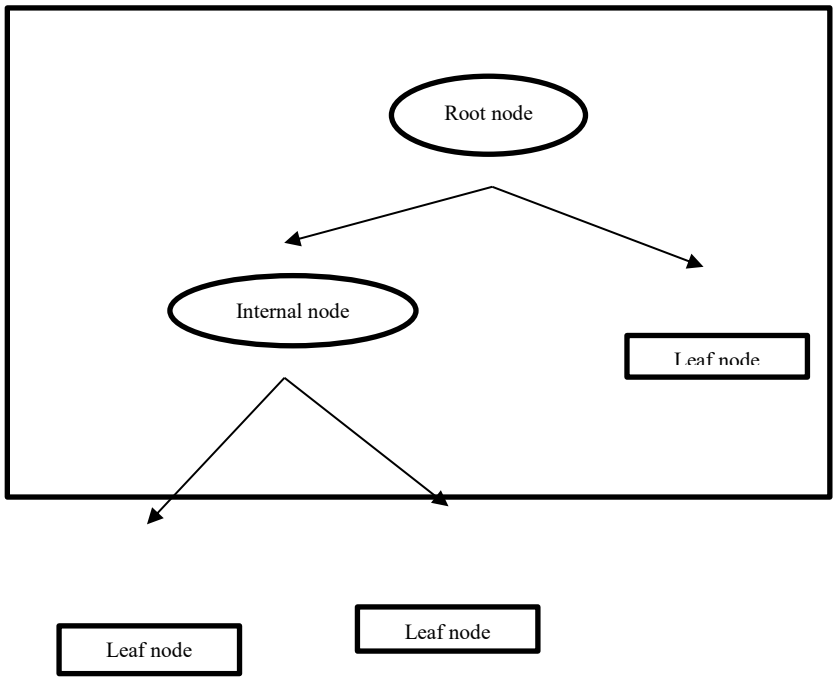
Sedangkan menurut Patami Kasih (Kasih, 2019) menyatakan bahwa *Decision Tree* merupakan diagram alir yang berbentuk seperti pohon (*Tree*) dimana node menyatakan pengujian terhadap atribut, *output* dari setiap pengujian tersebut dinyatakan dengan cabang dan kelas dinyatakan dengan *leaf node*. Node akar (*root)* merupakan level teratas yang biasanya berupa atribut yang memiliki pengaruh paling besar pada suatu kelas tertentu, konsep dari *Decision Tree* adalah mengubah data menjadi suatu model pohon keputusan yang kemudian diubah

menjadi sebuah *rule* (Setio dkk., 2020). *Decision Tree* memiliki 3 jenis node, yaitu:

1. *Root Node* (akar), merupakan node utama yang terletak paling atas dan paling memiliki pengaruh. Node ini tidak memiliki *input* namun dapat memiliki *output* lebih dari satu.
2. *Internal Node,* merupakan node yang terletak ditengah atau percabangan antara *root node* dan *leaf node.* Node ini memiliki satu input dan minimal mempunyai 2 *output*.

   Leaf Node atau terminal node, merupakan node yang terletak paling akhir, node ini hanya memiliki satu *input* dan tidak memiliki *output*.

Gambar **2.1** merupakan gambar struktur *Decision Tree*



Gambar 2.1 Struktur *Decision Tree*

Pohon keputusan dapat membantu manusia untuk mengidentifikasi dan melihat dengan mudah indikator yang mempengaruhi suatu permasalahan dan dapat menentukan pencarian solusi terbaik dengan memperhitungkan indikator

tersebut (Hermanto & SN, 2017)). Banyak algoritma yang dapat digunakan dalam membangun sebuah pohon keputusan salah satunya C4.5.

Algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang dikembangkan oleh J. Ross Quinlan pada akhir tahun 1970 hingga awal tahun 1980-an (Haidar dkk., 2019). Beberapa pengembangan pada algoritma C4.5 diantaranya dapat menangani *pruning, missing value,* dan *continu data* (Harman, 2018). Proses pertama yang dilakukan oleh algoritma C4.5 untuk membangun sebuah pohon keputusan yaitu dengan menentukan atribut sebagai akar dilanjutkan dengan pembuatan cabang untuk tiap-tiap nilai dalam akar tersebut. Yahya dan Jananto (Yahya & Jananto, 2019) menyatakan bahwa secara umum *Decision Tree* C4.5 memiliki tahapan algoritma sebagai berikut:

1. Menghitung nilai *Entropy,* dengan persamaan sebagai berikut:

$$Entropy\ (S) = \sum_{i=1}^{n} \_P_i \text{ x } log_2 P_i$$

   Keterangan:
   Pi : Proporsi data S dengan kelas *i*
   k : Jumlah kelas pada *output S.*
   n : Jumlah partisi atribut A
2. Menghitung nilai *information gain* untuk masing-masing atribut:

$$Gain(S,A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} x Entropy(S_i)$$

   Keterangan:
   S : Himpunan kasus
   A : Atribut
   n : Jumlah partisi atribut A
   |*Si*| : Jumlah kasus pada partisi ke i
   |*S*| : Jumlah kasus dalam S
3. Menghitung nilai *split info* untuk masing-masing atribut

$$Split\ Info\ (S,A) = -\sum_{i=1}^{n} \frac{Si}{S} log_2 \frac{Si}{S}$$

   Keterangan :
   S : Himpunan kasus
   A : Atribut
   n :Jumlah partisi atribut A
   Si : Jumlah sampel untuk atribut i

4. Menhitung nilai *gain ratio* untuk setiap atribut:

$$Gain\ Ratio\ (S,A) = \frac{Gain(S,A)}{Split\ Info\ (S,A)}$$

   Keterangan :

S : Himpunan kasus
A : Atribut
Gain(*S,A)* : *Info gain* pada atribut A
Split Info(S,A) : Split info pada atribut A

5. Akar (*root*) dipilih berdasarkan atribut yang memiliki *Gain Ratio* tertinggi dan atribut dengan *gain ratio* terendah dipilih menjadi cabang (*branch)*
6. Menghitng nilai *Gain Ratio* pada setiap atribut tanpa mengikut sertakan atribut yang telah terpilih pada tahap sebelumnya
7. Atribut dengan nilai *Gain Ratio* tertinggi dipilih sebagai branch
8. Mengulang langkah 6-7 sampai hasil *entropy* bernilai nol untuk setiap atribut yang tersisa.

2.1.3 Regresi Logistik Biner

Regresi Logistik biner merupakan suatu metode analisis statistik yang berguna untuk menganalisis hubungan antar suatu variabel respon dengan beberapa prediktor. Dengan variabel responnya berupa data kualitatif dikotomi yaitu bernilai 1 untuk menyatakan keberadaan suatu karakteristik dan 0 untuk menyatakan ketidakberadaan sebuah karakteristik (Tampil dkk., 2015).

Agresti dalam Fahmi (Fahmi & Khikmah, 2018) menyatakan bahwa regresi logistik biner digunakan untuk memodelkan suatu kejadian dengan variabel respon bertipe kategori dua pilihan yaitu sukses dan gagal. Menurut Hosmer dan Lemeshow dalam Manthovani (Misna dkk., 2018) menyatakan bahwa regresi logistik biner merupakan suatu metode analisis data yang dapat digunakan untuk mencari hubungan antar variabel y (terikat) yang bersifat dikotomus yang memiliki kategori biner dan variabel x (bebas) bersifat polikotomus.

2.1.4 Tabel *Krejcie-Morgan*

Tabel Krejcie-Morgan merupakan tabel yang digunakan untuk menentukan jumlah sampel dari n populasi dengan mengasumsikan tingkat kehandalannya sebesar 95% dan tingkat kesalahannya 5% (W. Wirawan dkk., 2019) sampel yang diperoleh memiliki tingkat kepercayaan sebesar 95% terhadap keseluruhan

populasi. Untuk memperoleh jumlah sampel menggunakan tabel *krejcie-morgan* diperoleh dengan cara melihat tabel *krejcie-morgan* seperti pada Tabel 2.1:

Tabel 2.1 Tabel Krejcie-Morgan

N	S	N	S	N	S	N	S
10	10	35	32	60	52	85	70
15	14	40	46	65	56	90	73
20	19	45	40	70	59	95	76
25	24	50	44	75	63	100	80
30	36	55	48	80	66	110	86

Sumber: diolah kembali dari (Mardisetosa dkk., 2020)

N merupakan jumlah populasi dan S merupakan jumlah sampel. Sebagai contoh jika jumlah populasi sebanyak 85 data, maka jumlah data yang dijadikan sebagai sampel adalah 70.

2.1.5 Stratified Random Sampling

Stratified Random Sampling Merupakan suatu proses pengambilan data yang akan dijadikan sebagai sampel melalui pembagian populasi ke dalam strata tertentu, memilih sampel secara acak dari setiap stratum dan kemudian digabungkan dalam sebuah data yang akan dijadikan sebagai sampel (Ulya dkk., 2018). Pemilihan data didasarkan pada angka random dan diperoleh sesuai dengan jumlah responden terpilih (sesuai dengan jumlah sampel yang telah ditentukan) (Arieska dkk., 2018). Meskipun tidak melibatkan semua data populasi, hasil *Stratified Random Sampling* mampu menggeneralisiasi sebagai representasi dari populasi.

Komposisi pengambilan sampel dilakukan secara acak dengan jumlah sampel bersifat proporsional sesuai dengan jumlah kelas yang ada untuk kemduian dijadikan sebagai total sampel penelitian (Arieska dkk., 2018). Cohcran dalam Ulya (Ulya dkk., 2018) menyatakan bahwa proses pengukuran dapat dilakukan dengan pengambilan jumlah sampel yang sedikit serta tidak melibatkan semua anggota populasi namun hasilnya dapat digeneralisasi sebagai representasi populasi.

2.1.6 *Confusion matrix*

Confusion matrix merupakan suatu metode yang berguna untuk mengevaluasi informasi dari sistem dengan menghitung akurasi sistem berdasarkan data training dan data uji (Tanjung dkk., 2016).

Metode ini melakukan proses evaluasi terhadap model klasifikasi berdasarkan proses perhitungan data latih dan data uji. Tabel 2.2 merupakan *tabel confusion matrix*:

Tabel 2.2 ***Confusion Matrix***

Klasifikasi	Prediksi	
	Prediksi = ya	Prediksi = tidak
Aktual = ya	Jumlah data positif yang teridentifikasi benar (*true positive* - TP)	Jumlah data negatif yang teridentifikasi salah (*false negative* - FN)
Aktual = tidak	Jumlah data Positif yang teridentifikasi salah (*false positive* - FP)	Jumlah data negatif yang teridentifikasi benar (*true negative* - TN)

Sumber (Galih, 2019)

Dari Tabel 2.2 dapat dilakukan perhitungan akurasi, presisi dan *recall*. Akurasi didefinisikan sebagai tingkat kecocokan antara nilai prediksi dan nilai aktual. Rumus yang digunakan untuk menghitung akurasi adalah

$$\text{Akurasi} = \frac{TP+TN}{TP+FP+TN+FN}$$

Akurasi dikatakan baik jika nilai prediksi dan aktual memiliki nilai yang sama atau hampir sama (memiliki tingkat kesalahan yang sangat rendah) dan dikatakan kurang baik jika akurasi yang dihasilkan berbeda dengan aktual (memiliki tingkat kesalahan yang tinggi). Akurasi merupakan seberapa dekat nilai hasil pengukuran dengan nilai sebenarnya (*true value*) atau nilai yang dianggap benar (Hanifah & Prastowo, 2016). Tabel 2.3 merupakan tabel nilai akurasi :

Tabel 2.3 Nilai Akurasi

Nilai Akurasi	**Hasil Prediksi**
Batas Nilai 0.90 - 1.00	Sangat baik
Batas Nilai 0.80 - 0.90	Baik
Batas Nilai 0.70 - 0.80	Sedang
Batas Nilai 0.60 - 0.70	Lemah
Batas Nilai 0.50 - 0.60	Sangat lemah

Sumber (Galih, 2019)

Presisi merupakan tingkat ketepatan antara informasi yang diberikan oleh sistem dengan permintaan user. Rumus yang digunakan untuk menghitung presisi adalah

$$\text{Presisi} = \frac{TP}{TP+FN}$$

Rumus yang digunakan untuk menghitung sensitivity adalah

$$\text{Sensitivity} = \frac{TP}{TP+FP}$$

Keterangan :

a) TP (*True Positive*) : Jumlah data positif yang teridentifikasi benar
b) TN (*True Negative*) : Jumlah data negatif yang teridentifikasi benar
c) FN (*False Negative*) : Jumlah data negatif yang teridentifikasi salah
d) FP (*False Positive*) : Jumlah data Positif yang teridentifikasi salah

2.2 Penelitian Terdahulu

Pada sub bab ini berisikan tentang penelitian-penelitian terdahulu yang berkaitan dengan penelitian yang akan dilakukan. Tabel 2.4 merupakan tabel yang berisi rangkuman dari penelitian-penelitian terdahulu.

Tabel 2.4 Penelitian Terdahulu

No	Judul	Peneliti	Tahun Terbit	Deskripsi	Metode	Data	HasilPenelitian
1	Teknik *Data mining* Menggunakan algoritma *Decision Tree* C4.5 untuk memprediksi tingkat kelulusan tepat waktu.	Chandra Wirawan	2020	Tingkat kelulusan dan mahasiswa baru tidak seimbang, maka akan mempengaruhi penilaian akreditasi pada Program Studi dan Universitas tersebut. Pada penelitian ini peneliti membahas prediksi tingkat kelulusan tepat waktu menggunakan teknik *data mining* algoritma C.4.5 dengan studi kasus UIN Syarif Hidayatullah Jakarta.	*Decision Tree* C4.5	Data kelulusan mahasiswa UIN Syarif Hidayatullah Jakarta.	Menggunakan teknik *data mining Decision Tree* algoritma C.4.5 dengan menggunakan tools RapidMiner dan diuji menggunakan confusion matrix menghasilkan nilai akurasi 89,82% dengan *Precision* 52,63% dan *Recall* 41,67%, dan nilai AUC sebesar 76,6% nilai 89,82% ini mengindikasikan bahwa performa keakuratan pada percobaan tersebut bernilai cukup baik.
2	Prediksi Kelulusan Mahasiswa Menggunakan Algoritma C4.5	Ade Fatma Ayu Rahman, Sorikhi dan Wartulas	2020	Tujuan dari penelitian ini yaitu untuk mengetahui tingkat akurasi algoritma C4.5 dalam prediksi kelulusan mahasiswa Universitas	*Decision Tree* C4.5	Data kelulusan mahasiswa universitas	Hasil penelitian tersebut menunjukan bahwa algoritma C4.5 dapat memprediksi kelulusan mahasiswa Universitas Peradaban dengan

No	Judul	Peneliti	Tahun Terbit	Deskripsi	Metode	Data	HasilPenelitian
	(Studi Kasus di Universitas Peradaban)			Peradaban. Atribut yang digunakan IPK, SKS, Umur, dan Jenis Kelamin sebanyak 151 data		Peradaban	tingkat akurasi 88,74%, presisi 91,7%, dan recall sebesar 95,34
3	Prediksi Ketepatan Kelulusan Mahasiswa Diploma dengan Komparasi Algoritma Klasifikasi	Maulana dan kawan-kawan	2019	Penelitian ini mengkomparasi 5 algoritma *data mining* yaitu *Decision Tree* C4.5, *Naive Bayes*, K-NN, rule Induction, dan random forest dengan pengujian T-test. Tujuan dari penelitian ini adalah menentukan metode yang paling akurat untuk menentukan ketepatan kelulusan mahasiswa. Atribut yang digunakan yaitu Nama, Jenis Kelamin. Pekerjaan Mahasiswa, Umur, Status Nikah, IPS 1-6, IPK, dan status kelulusan.	*Decision Tree* C4.5, *Naive Bayes*, K-NN, rule Induction, dan random forest	Data yang digunakan pada penelitian ini adalah data mahasiswa AMIK BSI Pontianak angkatan 2013/2017 prodi Informatika sebanyak 394 record.	Hasil dari penelitian ini didapatkan algoritma yang paling optimal dalam menentukan ketepatan kelulusan mahasiswa diploma AMIK BSI Pontianak adalah *Decision Tree* C4.5 dengan akurasi sebesar 90,85%.
4	Prediksi Kelulusan Mahasiswa Tepat Waktu Berdasarkan Usia, Jenis Kelamin,	Romadhona, Suprapedi,dan dan kawan-kawan	2017	Penelitian ini bertujuan untuk menemukan pola prediksi kelulusan mahasiswa tepat waktu dengan menggunakan teknik *data mining*	*Decision Tree* C4.5, CHAID, dan ID3	Data mahasiswa STIMIK Adhiguna	Hasil dari penelitian ini menunjukan bahwa dari seluruh atribut yang digunakan menunjukan bahwa indeks prestasi mendapatkan nilai

No	Judul	Peneliti	Tahun Terbit	Deskripsi	Metode	Data	HasilPenelitian
	Dan Indeks Prestasi Menggunakan Algoritma *Decision Tree*			dan model untuk memprediksi lama masa studi adalah algoritma *Decision Tree* C4.5 dibandingkan dengan algoritma ID3 dan CHAID menggunakan data uji untuk menentukan persentase presisi, recall.			gain tertinggi yaitu 0,340 dengan begitu atribut tersebut dapat dijadikan sebagai root. Algoritma. Decision tree memiliki kompleksitas yang cukup tinggi karena adanya penelusuran dan pemrosesan nilai pada setiap atributnya dengan tujuan untuk mendapatkan entropi. Akurasi yang dihasilkan dari penerapan Decision tree C4.5 yaitu sebesar 91,51%.
5	Implementasi *Data mining* Menggunakan Algoritma *Naive Bayes* Classifier dan C4.5 untuk Predksi Kelulusan Mahasiswa	Endang Etriyanti,	2020	Penelitian ini bertujuan untuk mengetahui kinerja dari algoritma *Naive Bayes*dan decsion tree C4.5 dengan tingkat akurasi yang besar untuk menyelesaikan permasalhan prediksi kelulusan mahasiswa STIMIK Bina Nusantara Jaya Lubuk Linggau.	*Naive Bayes* dan descion tree C4.5	Data mahasiswa STIMIK Bina Nusantara Jaya Lubulkinggau angkatan 2013 dan 2014 sebanyak 162	Hasil penelitian menunjukan bahwa metode *Decision Tree* C4.5 dapat digunakan untuk prediksi kelulusan mahasiswa STIMIK Bina Nusantara Jaya Lubuk Linggau dengan nilai akurasi sebesar 79,08%. dengan root-nya adalah IPK-S4. Sedangkan metode *Naive Bayes* classifier hanya mencapai akurasi 78,46%.

No	Judul	Peneliti	Tahun Terbit	Deskripsi	Metode	Data	HasilPenelitian
						data.	
6	*Data mining* di Bidang Pendidikan untuk Analisis Predikisi Kinerja Mahasiswa dengan Komparasi 2 Model Klasifikasi	Galih	2019	Penelitian ini mengkomparasi algoritma classifier yaitu *Decision Tree* C4.5 dan *Naive Bayes* untuk melihat tingkat akurasi yang dihasilkan dari kedua metode tersebut sehingga metode yang memiliki akurasi paling tinggi akan digunakan untuk melakukan prediksi kellusan mahasisswa.	*Decision Tree* C4.5 dan *Naive Bayes*	Data mahasiswa STIMIK JABAR sebanyak 836 *record*	Penelitian ini menunjkan bahwa *Decision Tree* memiliki tingkat akurasi yang lebih tinggi dibandingkan dengan *Naive Bayes.* Dengan akurasi sebesar 88,10% dan *Naive Bayes* 86,3%. Dapat disimpulkan bahwa *Decision Tree* C4.5 dapat digunakan untuk memprediki kelulusan mahasiswa.

Berikut merupakan penjelasan lengkap Tabel 2.4 :

1. Teknik *Data mining* Menggunakan Algoritma *Decision Tree* C4.5 Untuk Memprediksi Tingkat Kelulusan Tepat Waktu.

   Penelitian dilakukan untuk memprediksi tingkat kelulusan tepat waktu menggunakan teknik *data mining* algoritma C4.5 di UIN Syarif Hidayatullah Jakarta. Data yang digunakan sebanyak 754 data yang terdiri dari 30% *data testing* dan 70% *data training* yang terdiri dari data 10 atribut diaman 9 atribut merupakan atribut prediktor dan 1 atribut merupakan atribut hasil. Pengambilan jumlah sampel dilakukan dengan menggunakan metode *stratified random sampling*. Penelitian ini menggunakan *tools* RapidMiner dan di uji menggunakan *confusion matrix,* menghasilkan nilai akurasi 89,82% dengan *precision* 52,63% dan *Recall* 41,67%, dan nilai AUC sebesar 76,6%. Nilai 89,82% ini mengindikasikan bahwa performa keakuratan pada percobaan tersebut bernilai cukup baik. Dapat disimpulkan bahwa *Decision Tree* C4.5 dapat digunakan untuk memprediksi kelulusan mahasiswa (C. Wirawan, 2020).

   Relevansi dengan penelitian yang akan dilakukan yaitu terletak pada subjek penelitian dan metode pengambilan sampel yaitu menggunakan metode *stratified random sampling*, sehingga diharapkan hasil dari penelitian tersebut dapat memberikan sumbangsih ide terhadap penelitian yang akan dilakukan. Perbedaannya terletak pada pemilihan dan jumlah atribut. Dalam penelitian tersebut pemilihan atribut didasarkan pada opini peneliti sedangkan pada penelitian yang akan dilakukan pemilihan atribut didasarkan pada Omnibus *test* dan *variabel in the equation.*

2. Prediksi Kelulusan Mahasiswa Menggnakan Algoritma C4.5 (Studi Kasus di Universitas Peradaban)

Tujuan dari penelitian ini yaitu untuk mengetahui tingkat akurasi algoritma C4.5 dalam prediksi kelulusan mahasiswa Universitas Peradaban. Penelitian ini menggunakan metode *Decision Tree* C4.5, data yang digunakan adalah data kelulusan mahasiswa Universitas Peradaban dengan atribut IPK, SKS, Umur, dan Jenis Kelamin sebanyak 151 data. Hasil dari penelitian ini didapatkan atribut yang

menjadi *root* adalah IPK. Pengujian dilakukan dengan meggunakan *confusion matrix* dengan tingkat akurasi yang dihasilkan 88,74%, presisi 91,7%, dan *recall* sebesar 95.34. Dapat disimpulkan bahwa *Decision Tree* C4.5 dapat digunakan untuk memprediksi kelulusan mahasiswa (Rahman dkk., 2020).

Relevansi dengan penelitian yang akan dilakukan terletak pada objek penelitian, sehingga hasil dari penelitian tersebut diharapkan mampu memberikan sumbangsih ide terhadap penelitian yang akan dilakukan. Perbedaannya terletak pada tujuan penelitian. Penelitian tersebut bertujuan untuk mengukur tingkat akurasi metode *Decision Tree* C4.5 sedangkan pada penelitian ini bertujuan untuk menemukan prediktor apa saja yang mempengaruhi kelulusan di STIMIK ESQ. Perbedaan lainnya yaitu dari atribut yang digunakan dan teknik pengambilan sampel. Pada penelitian tersebut pemilihan atribut didasarkan pada opini peneliti sedangkan pada penelitian ini pemilihan atribut didasarkan pada Omnibus *test* dan *variabel in the equation.*

3. Prediksi Ketepatan Kelulusan Mahasiswa Diploma dengan Komparasi Algoritma Klasifikasi

Penelitian ini mengkomparasi 5 algoritma *data mining* yaitu *Decision Tree* C4.5*, Naive Bayes,* K-NN, *rule Induction* dan *random forest* dengan pengujian *T-test.* Tujuan dari penelitian ini adalah menentukan metode yang paling akurat untuk menentukan ketepatan kelulusan mahasiswa. Data yang digunakan pada penelitian ini adalah data mahasiswa AMIK BSI Pontianak angkatan 2013-2017 prodi Informatika sebanyak 394 *record.* Atribut yang digunakan yaitu Nama, Jenis Kelamin, Pekerjaan Mahasiswa, Umur, Status Nikah, IPS 1-6, IPK, dan status kelulusan. Hasil dari penelitian ini didapatkan algoritma yang paling optimal dalam menentukan ketepatan kelulusan mahasiswa diploma AMIK BSI Pontianak adalah *Decision Tree* C4.5 dengan akurasi sebesar 90,85%. Dapat disimpulkan bahwa *Decision Tree* C4.5 dapat digunakan untuk memprediksi kelulusan mahasiswa (M. S. Maulana dkk., 2019).

Relevansi dengan penelitian yang akan dilakukan terletak pada objek penelitian, sehingga hasil dari penelitian tersebut diharapkan dapat memberikan sumbangsih terhadap penelitian yang akan dilakukan. Perbedaannya terletak pada tujuan penelitian, penelitian tersebut bertujuan untuk mengkomparasi 5 algoritma *data mining* sedangkan pada penelitian ini bertujuan untuk menemukan prediktor apa saja yang mempengaruhi kelulusan di STIMIK ESQ. Perbedaan lainnya yaitu dari atribut yang digunakan dan teknik pengambilan sampel, pada penelitian tersebut pemilihan atribut didasarkan pada opini peneliti sedangkan pada penelitian ini pemilihan atribut didasarkan pada Omnibus *test* dan *variabel in the equation.*

4. Prediksi Kelulusan Mahasiswa Tepat Waktu Berdasarkan Usia, Jenis Kelamin, Dan Indeks Prestasi Menggunakan Algoritma *Decision Tree*

Penelitian ini melakukan komparasi antar algoritma yag dipakai dalam *Decision Tree* yaitu ID3, CHAID, dan *Decision Tree* C4.5. Tujuan dari penelitian ini yaitu untuk menemukan pola prediksi kelulusan mahasiswa tepat waktu dengan menggunakan teknik *data mining* dan model untuk memprediksi lama masa studi adalah algoritma *Decision Tree* C4.5 dibandingkan dengan algoritma ID3 dan CHAID menggunakan data uji untuk menentukan persentase *presisi, recall,* dan akurasi. Data yang digunakan pada penelitian ini adalah data mahasiswa STIMIK Adhiguna angkatan 2009 dengan atibut usia, jenis kelamin, dan indeks prestasi semester 1-4. Hasil penelitian menunjukan bahwa *Decision Tree* memiliki kompleksitas yang cukup tinggi karena adanya penelusuran dan pemrosesan nilai pada setiap atributnya dengan tujuan untuk mendapatkan *entropi.* Akurasi tertinggi yang dihasilkan dari penerapan metode *Decision Tree* C4.5, ID3, dan CHAID didapatkan oleh *Decision Tree* C4.5 yaitu sebesar 91,51%. Dapat disimpulkan bahwa *Decision Tree* C4.5 dapat digunakan untuk memprediksi kelulusan mahasiswa (Romadhona, A., Suprapedi , S. dan Himawan, 2017).

Relevansi dengan penelitian yang akan dilakukan terletak pada objek penelitian dan tujuan penelitian. Perbedaanya terletak pada atribut dan algoritma yang digunakan. Pada penelitian tersebut dilakukan perbandingan algoritma *Decision*

Tree yaitu C4.5, ID3, dan CHAID. Sedangkan pada penelitian yang akan dilakukan hanya menggunakan satu algoritma saja.

5. Implementasi *Data mining* Menggunakan Algoritma *Naive Bayes*

Pada penelitian ini dilakukan komparasi algoritma *Naive Bayes* dan *Decision Tree* C4.5. Tujuan penelitian ini yaitu untuk mengetahui kinerja dari algoritma *Naive Bayes* dan *Decision Tree* C4.5 dengan tingkat akurasi yang besar untuk menyelesaikan permasalahan prediksi kelulusan mahasiswa STIMIK Bina Nusantara Jaya Lubuk Linggau. Data yang digunakan pada penelitian ini yaitu data mahasiswa STIMIK Bina Nusantara Jaya Lubuk Linggau angkatan 2013 dan 2014 sebanyak 162 data. Atribut yang digunakan pada penelitian ini yaitu jenis kelamin, status sekolah, sekolah asal, IPS 1-4 dan status kelulusan. Hasil penelitian menunjukan bahwa metode *Decision Tree* C4.5 dapat digunakan untuk prediksi kelulusan mahasiswa STIMIK Bina Nusantara Jaya Lubuk Linggau dengan nilai akurasi sebesar 79,08%, sedangkan metode *Naive Bayes classifier* hanya mencapai akurasi 78,46%. Variabel yang menjadi *root* dalam prediksi kelulusan mahasiswa di STIMIK BINA Nusantara Jaya Lubuk Linggau adalah IPK-S4, jenis kelamin, dan asal sekolah. Dapat disimpulkan bahwa *Decision Tree* C4.5 dapat digunakan untuk memprediksi kelulusan mahasiswa (Etriyanti dkk., 2020).

Relevansi dengan penelitian yang akan dilakukan yaitu terletak pada permasalahan yang diangkat yaitu kelulusan mahasiswa sehingga hasil dari penelitian tersebut diharapkan dapat memberikan sumbangsih ide. Perbedaanya terletak pada penggunaan metode, pada penelitian tersebut digunakan dua metode *data mining* sedangkan pada penelitian yang akan dilakukan hanya menggunakan satu metode saja. Perbedaan lain yaitu terletak pada atribut yang digunakan.

6. *Data mining* di Bidang Pendidikan untuk Analisis Prediksi Kinerja Mahasiswa dengan Komparasi 2 Model Klasifikasi pada STIMIK Jabar

Penelitian ini mengkomparasi algoritma *classifier* yaitu *Decision Tree* C4.5 dan *Naive Bayes* untuk melihat tingkat akurasi yang dihasilkan dari kedua metode tersebut sehingga metode yang memiliki akurasi paling tinggi akan digunakan untuk melakukan prediksi kelulusan mahasiswa. Data yang digunakan pada penelitian ini adalah data mahasiswa STIMIK JABAR sebanyak 836 *record*

dengan atribut NIM, jenis kelamin, kategori asal sekolah, asal kota, jarak tempuh ke kampus, pekerjaan orang tua, IPK, dan keterangan lulus. Pengolahan data dilakukan dengan tools RapidMiner. Pengujian dilakukan dengan *confusion matrix* dan *cross validation* digunakan untuk memvalidasi hasil pengujian. Penelitian ini menunjukan bahwa *Decision Tree* memiliki tingkat akurasi yang lebih tinggi dibandingkan dengan *Naive Bayes.* Dengan akurasi sebesar 88,10% dan *Naive Bayes* 86,3%. Dapat disimpulkan bahwa *Decision Tree* C4.5 dapat digunakan untuk memprediksi kelulusan mahasiswa (Galih, 2019).

Relevansi dengan penelitian yang akan dilakukan yaitu prediksi kelulusan mahasiswa,sedangkan perbedaanya terletak pada metode yang digunakan. Pada penelitian ini digunakan dua metode kemudian dibandingkan untuk memilih metode yang memiliki nilai akurasi yang lebih tinggi, sedangkan pada penelitian yang akan dilakukan hanya menggunakan satu metode saja.

2.3 Metode Pemecahan Masalah

Tabel 2.5 merupakan tabel metode yang dapat digunakan untuk pemecahan masalah prediksi kelulusan mahasisswa

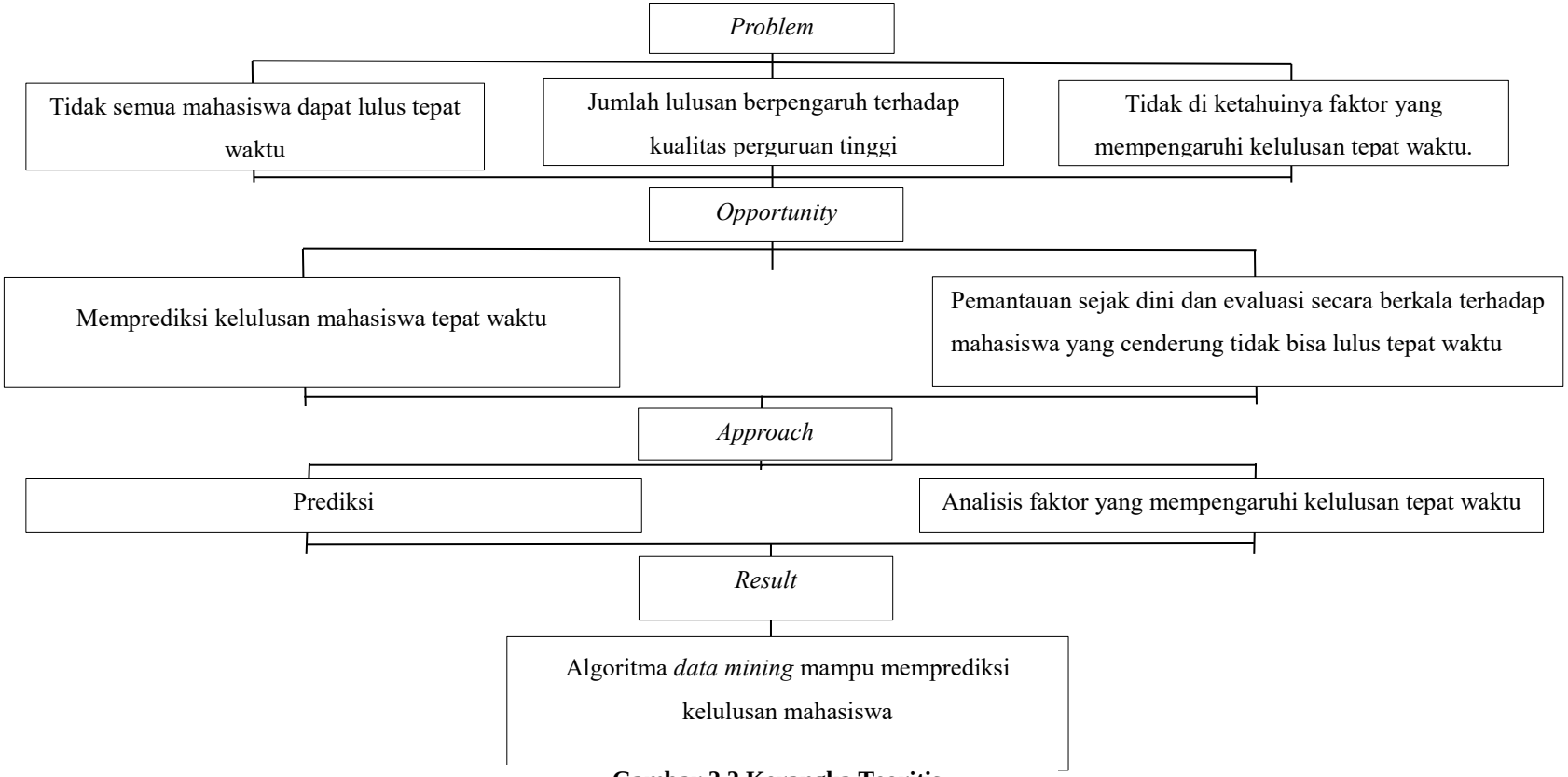
Tabel 2.5 Metode Pemecahan Masalah

No	Metode	Relevansi	Penelitian terdahulu	Pengujian
1	*Decision Tree*	Metode *Decision Tree* dapat digunakan untuk melakukan prediksi. *Decision Tree* dapat merepresentasikan setiap atribut berdasarkan atribut yang memiliki pengaruh tertinggi sampai atribut yang kurang berpengaruh. Sehingga sangat relevan jika digunakakan untuk memprediksi kelulusan mahasiswa berdasarkan faktor-faktor yang mempengaruhinya.	Teknik *Data mining* Menggunakan algoritma *Decision Tree* C4.5 untuk memprediksi tingkat kelulusan tepat waktu (C. Wirawan, 2020). Penelitian ini menghasilkan akurasi sebesar 89,82%.	Pengujian model menggunakan metode *confusion matrix* berupa nilai *accuracy,precision, recall,* dan diagram AUC
			Prediksi Kelulusan Mahasiswa Menggnakan Algoritma C4.5 (Studi Kasus di Universitas Peradaban) (Rahman dkk., 2020). Penelitian ini menghasilkan nilai akurasi sebesar 88,74%.	Pengujian dilakukan dengan menggunakan *confusion matrix*
			Prediksi Kelulusan Mahasiswa Tepat Waktu Berdasarkan Usia, Jenis Kelamin, Dan Indeks Prestasi Menggunakan Algoritma *Decision Tree* (Romadhona, 2017). Penelitian ini menghasilkan akurasi sebesar 91,51%.	Pengujian dilakukan dengan menggunakan *k-fold (number of validation)*
2	*Naive Bayes*	Metode *Naive Bayes* dapat digunakan untuk pengklasifikasian dengan menggunakan metode probabilitas untuk memprediksi peluang dimasa mendatang.	Prediki Kelulusan Mahasiswa Menggunakan *Naive Bayes* (Amelia dkk., 2017). Penelitian ini menghasilkan akurasi sebesar 88,96%	Pengujian dilakukan dengan menggunakan *k-fold cross validation 10-fold.*
			Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan *Naïve Bayes* : Studi Kasus UIN Syarif Hidayatullah Jakarta *Prediction of Timeliness Graduation of Students Using Naïve Bayes : A Case Study at Islamic State University Syarif Hidayatullah* (Salmu, S. & Solichin, 2017). Penelitian ini menghasilkan akurasi sebesar 80,72%	Pengujian dilakukan dengan *confusion matrix* yang terdiri dari *precission, recall, f-measure, dan akurasi*

No	Metode	Relevansi	Penelitian terdahulu	Pengujian
3	KNN (*K-Nearst Neighbor*)	Metode KNN merupakan metode klasifikasi yang melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek yang akan diuji dapat digunakan untuk prediksi.	Model Algoritma *K-Nearest Neighbor (K-NN)* Untuk Prediksi Kelulusan Mahasiswa (Rohman, 2015). Penelitian ini menghasilkan akurasi sebesar 85,15%	Kurva ROC (*Receiver Operating Characteristic)*

Berdasarkan Tabel 2.5, metode *Decision Tree* C4.5 memiliki nilai akurasi yang lebih tinggi jika dibandingkan dengan metode *Naive Bayes* dan KNN dalam melakukan prediksi kelulusan mahasiswa tepat waktu. Selain itu metode *Decision Tree* juga dapat merepresentasikan atribut yang memiliki pengaruh besar sampai atribut yang memiliki pengaruh rendah yang direpresentasikan dalam bentuk sebuah pohon keputusan. Maka pada penelitian ini akan digunakan metode *Decision Tree* C4.5 sebagai metode untuk memecahkan masalah.

2.4 Kerangka Teoritis

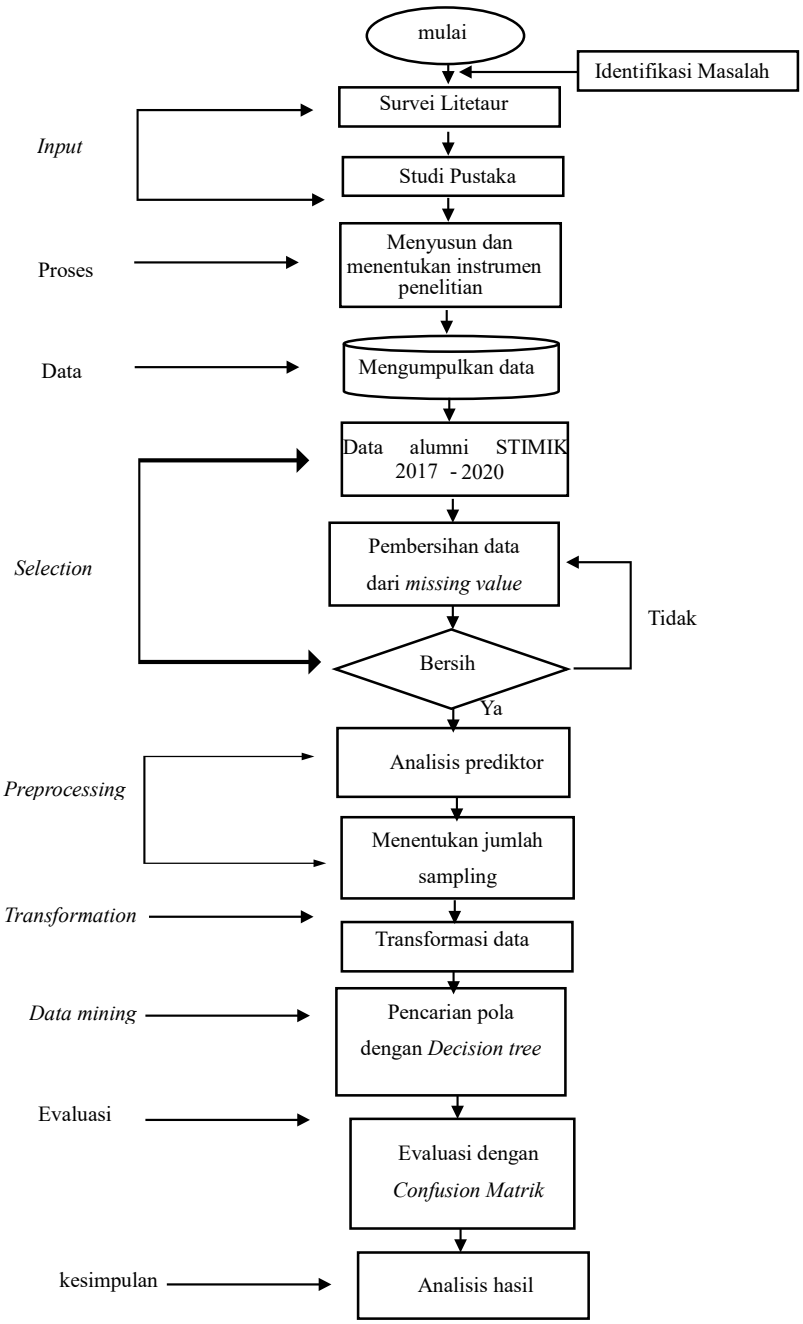


Gambar 2.2 Kerangka Teoritis

BAB 3
METODOLOGI PENELITIAN

3.1 Alur penelitian

Pada sub bab ini akan dijelaskan mengenai alur dari penelitian yang dilakukan



Gambar 3.1 Alur Penelitian

Tahapan penelitian pada Gambar 3.1 dapat dijelaskan sebagai berikut:

1. Identifikasi Masalah

   Identifikasi masalah merupakan tahapan paling awal yang dilakukan peneliti, salah satunya yaitu mengidentifikasi permasalahan terhadap kelulusan mahasiswa di STIMIK ESQ. Hasil dari identifikasi tersebut peneliti menemukan bahwa adanya permasalahan terhadap kelulusan mahasiswa di STIMIK ESQ di setiap tahunnya, adanya jumlah mahasiswa yang tidak lulus tepat waktu akan berdampak pada akreditasi STIMIK ESQ terlebih lagi STIMIK ESQ merupakan perguruan tinggi yang belum terakreditasi. Hal ini juga memberikan dampak kurang baik kepada mahasiswa yang tidak bisa lulus tepat waktu seperti adanya pembayaran tambahan untuk perkuliahan dan berkurangnya peluang untuk mencari pekerjaan serta tambahan waktu yang harus dikeluarkan untuk menyelesaikan masa perkuliahan.

2. Survei Literatur

   Tahapan ini merupakan tahapan pengumpulan bahan literatur dan informasi yang berkaitan dengan penelitian sebelumnya mengenai permasalahan kelulusan mahasiswa di perguruan tinggi serta pencarian metode yang terbaik untuk menanganinya.

3. Studi Pustaka

   Tahapan ini merupakan langkah awal dalam mengumpulkan data, dimana pengumpulan data tersebut diarahkan pada pencarian data dan informasi melalui dokumen-dokumen yang mampu mendukung dalam proses penelitian "Implementasi *Data mining* untuk Kelulusan Tepat Waktu Mahasiswa STIMIK ESQ Menggunakan *Decision Tree* C4.5".

4. Menentukan dan menyusun instrumen penelitian

   Tahapan ini merupakan penentuan instrumen penelitian yang akan digunakan pada penelitian "Implementasi *Data mining* untuk Kelulusan Tepat Waktu Mahasiswa STIMIK ESQ Menggunakan *Decision Tree* C4.5". Instrumen yang dilakukan pada penelitian ini adalah dokumentasi. Dengan dokumentasi, peneliti memperoleh data dan informasi dari

berbagai macam sumber, salah satunya yaitu dokumen sekunder berupa laporan penelitian dengan objek penelitian yang sama serta data yang dibutuhkan dari pihak akademik STIMIK ESQ.

5. Mengumpulkan Data

   Tahap ini dilakukan untuk mengumpulkan data dan informasi yang berkaitan dengan penelitian yang dilakukan, untuk memperoleh data tersebut dilakukan beberapa teknik yaitu teknik Perpustakaan (*Library Research*), yaitu penelitian yang dilakukan dengan cara pengumpulan data dengan membaca literatur untuk mendapatkan teori yang berhubungan dengan permasalahan yang sedang diteliti. Peneliti memperoleh data dengan *Secondary* data, dimana data yang diperoleh berasal dari sumber yang sudah ada, data yang diperoleh diantaranya berupa Nama Mahasiswa, NIM, Usia Saat Masuk, indeks prestasi semester 1 sampai dengan indeks prestasi semester 8, indeks prestasi kumulatif semester 1 sampai dengan semester 8, Jumlah SKS yang diambil setiap semesternya serta keterangan lulus tepat waktu atau tidak.

6. Seleksi Data

   Pada umumnya data yang diperoleh dari database memiliki isian yang tidak sesuai atau tidak sempurna seperti *missing value*, data yang tidak valid atau hanya sekedar salah penulisan. Maka pada tahap ini dilakukan pembersihan data dari *missing value* serta data yang tidak relevan. Pembersihan data perlu dilakukan agar data yang diolah memiliki performa yang baik. Pembersihan data dilakukan secara manual dengan menghapus data yang kosong ataupun data yang tidak relevan.

7. *Preprocessing*

   Pada tahap ini dilakukan pemilihan fitur menggunakan SPSS (*Statistica Product and Service Solutions*) dengan metode Regresi logistik biner pemilihan fitur bertujuan untuk menentukan fitur yang paling berpengaruh terhadap kelulusan mahasiswa tepat waktu. Pada tahap ini juga dilakukan penentuan fitur yang menjadi prediktor kemudian dilakukan penentuan jumlah sampel yang akan digunakan pada penelitian mengacu pada tabel

krejcie. Dengan populasi data sebanyak 188, didapatkan data yang akan dijadikan sebagai sampel sebanyak 86 yang terdiri dari 35 data mahasiswa lulus tepat waktu dan 35 data mahasiswa yang tidak lulus tepat waktu. Pengambilan anggota sampel dilakukan dengan menggunakan metode *random stratified sampling*.

8. Transformasi

   Pada tahap ini dilakukan transformasi data ke dalam bentuk numerik, adapun data yang ditransformasikan yaitu data pada atribut masa studi ke dalam data numerik, kelulusan tepat waktu dinotasikan dengan angka 0 dan kelulusan tidak tepat waktu dinotasikan dengan angka 1. Kemudian pada atribut Jenis kelamin laki-laki dinotasikan dengan angka 0 dan perempuan dinotasikan dengan angka 1. Pada atribut prodi, sistem informasi dinotasikan dengan angka 0 dan Manajemen dengan angka 1. Pada tahap ini juga dilakukan penskalaan pada atribut SKS, SKS dengan rentang 0-15 dinotasikan dengan angka 3, 16-18 sksk dinotasikan dengan angka 2, 19-21 dinotasikan dengan angka 1, dan 22-24 dinotasikan dengan angka 0, rentang nilai tersebut didasarkan pada panduan akademik tahun 2020. Selanjutnya pada nilai IPS dan IPK, jika nilai IPK <1.99 maka dinotasikan dengan angka 4, 2.00-2.75 dinotasikan dengan angka 3, 2.76-3.00 dinotasikan dengan angka 2, 3.00-3.50 dinotasikan dengan angka 1 dan rentang nilai IPS dan IPK dari 3.51-400 dinotasikan dengan angka 0.

9. *Data mining*

   Pengolahan data dilakukan dengan menggunakan bahasa pemrograman python. Adapun metode yang digunakan adalah *Decision Tree* C4.5, Dengan tujuan untuk mendapatkan *rules* yang tersembunyi dari data lulusan mahasiswa STIMIK ESQ tahun 2017-2020.

10. Evaluasi

    Setelah dilakukan pengolahan pada tahap sebelumnya kemudian dilakukan evaluasi terhadap metode yang digunakan, untuk mendapatkan nilai akurasi yang paling tinggi dan nilai error yang paling rendah. Evaluasi dilakukan dengan menggunakan *confusion matrix.*

11. Interpretasi

    Tahap ini merupakan tahap paling terakhir dilakukan, yaitu interpretasi dari *rule* yang dihasilkan dari tahap tahap sebelumnya.

3.2 Instrumen Penelitian

Instrumen penelitian menurut sugiyono dalam sujadijaya (Trimo, 2017) menyatakan bahwa instrumen penelitian adalah suatu alat yang digunakan untuk mengukur fenomena alam maupun sosial yang diamati. Berdasarkan permasalahan

yang telah dipaparkan pada bab sebelumnya, maka instrumen dalam penelitian ini terbagi dua yaitu bahan dan peralatan yang dijelaskan sebagai berikut:

3.2.1 Bahan

Dalam penelitian ini bahan yang diperlukan adalah data akademik dan profil kelulusan mahasiswa STIMIK ESQ tahun 2017-2020.

3.2.2 Alat

Peralatan yang dibutuhkan dalam penelitian ini meliputi kebutuhan perangkat keras dan perangkat lunak. Kebutuhan perangkat lunak:

1. SPSS *(Statistica Product and Service Solutions)*

   SPSS merupakan sebuah program aplikasi yang mampu melakukan analisis statistik serta manajemen data pada lingkungan grafis yang dilengkapi dengan menu-menu deskriptif dan kotak dialog yang sederhana sehingga memudahakan pengoprasiannya. SPSS *(Statistica Product and Service Solutions).* Dari berbagai jenis perangkat lunak yang ada, SPSS dapat dijadikan pilihan karena fasilitas dan kemudahan dalam mengoperasikannya.

   Pada penelitain ini SPSS digunakan sebagai *tools* untuk melakukan perhitungan *regresi binary logistic*. Perhitungan *regresi binary logistic* bertujuan untuk melihat hubungan variabel bebas dengan variabel terikat, serta untuk melihat variabel bebas yang memiliki pengaruh terbesar.

2. *Google Colab*

   G*oogle colab (Google colaboratory)* merupakan sebuah *tools* yang dirilis oleh google. *Google colab* memberikan fasilitas untuk pengolahan data

menggunakan *machine learning* maupun *deep learning*. Penggunaan *google colab* hampir sama dengan *jupyter nootebook* karena *google colab* dibuat di atas *environment* jupyter perbedaannya terletak pada hal penyimpanannya. Media penyimpanan pada *google colab* adalah google drive dan berjalan pada sistem *cloud. Google colab* menyediakan berbagai macam *library* seperti keras, Tensorflow, OpenCV maupun Pytorch. Pada penelitian ini G*oogle colab* digunakan sebagai *tools* untuk melakukan proses *data mining* dengan menggunakan bahasa pemrograman *python.*

3.3 Data

Pada penelitian ini, data yang digunakan merupakan data kelulusan mahasiswa STIMIK ESQ tahun 2017-2020 sebanyak 186 yang terdiri dari 2 prodi yaitu prodi sistem informasi dan Manajemen. Tabel 3.1 merupakan jumlah mahasiswa berdasarkan prodi

Tabel 3.1 Jumlah Mahasiswa Berdasarkan Prodi

Prodi	Jumlah
Sistem Informasi	61
Manajemen	125
Total	**186**

Pada penelitian ini, kelulusan mahasiswa dibagi menjadi dua kategori yaitu lulus tepat waktu dan lulus tidak tepat waktu atau terlambat. Mahasiswa yang lulus tepat waktu merupakan mahasiswa yang mampu menyelesaikan masa pembelajaran kurang dari atau sama dengan 4 tahun ($\leq$ 4 tahun) dan mahasiswa yang tidak lulus tepat waktu atau terlambat merupakan mahasiswa yang menyelesaikan masa studi lebih dari 4 tahun (>4 tahun). Tabel 3.2 merupakan jumlah dan persentase kelulusan mahasiswa STIMIK ESQ tahun 2017-2020:

Tabel 3.2 Jumlah Mahasiswa Berdasarkan Kelulusan

Kelulusan	Frekuensi	Persentase
lulus tepat waktu ($\leq$ 4 tahun)	137	74%
lulus tidak tepat waktu (>4 tahun)	49	26%
Total	**186**	**100%)**

Tabel 3.2 menunjukan bahwa mahasiswa yang mampu lulus tepat waktu berjumlah 137 mahasiswa atau setara dengan 74% dari total mahasiswa dan mahasiswa yang tidak lulus tepat waktu berjumlah 49 orang atau setara dengan 26% dari total mahasiswa.

Jenis data yang digunakan pada penelitian ini adalah data sekunder, data sekunder menurut Harry Hermawan (Hermawan, 2019) merupakan data yang didapatkan secara tidak langsung oleh peneliti sendiri, dapat melalui berbagai media yang diperoleh dan dipublikasikan oleh pihak lain, data tersebut dapat berupa laporan, catatan, dokumen, dan studi pustaka. Pada penelitian ini data diperoleh dari pihak kedua yaitu dari Biro Administrasi Akademik (BAA) STIMIK ESQ, data tersebut berupa data akademik dan profil mahasiswa STIMIK ESQ lulusan tahun 2017-2020.

Metode pengumpulan data pada penelitian ini menggunakan studi dokumentasi, dengan mengumpulkan dokumen-dokumen yang dibutuhkan. Dokumen merupakan catatan atau peristiwa yang sudah terjadi, yang dapat diperoleh dengan dokumen dalam bentuk teks, foto, gambar atau arsip suatu instansi yang memiliki relevansi dengan penelitian yang akan dilakukan. Pada penelitian ini pengumpulan data diperoleh dari dokumen STIMIK ESQ.

3.4 Sampling

Penentuan besarnya jumlah sampel mengacu pada tabel krejcie, dari total populasi 86 data didapatkan jumlah data sampel sebanyak 70 data. Penentuan anggota sampel menggunakan metode *stratified random sampling* dimana 35 data terdiri dari mahasiswa yang lulus tepat waktu dan 35 mahasiswa yang tidak lulus tepaat waktu atau terlambat.

3.4.1 SPSS *(Statistica Product and Service Solutions)*

Pada penelitian ini SPSS digunakan untuk melakukan analisis variabel yang mempengaruhi kelulusan mahasiswa. Analisis data dilakukan dengan tujuan sebagai berikut:

1. Untuk melihat variabel bebas (Usia, Prodi, IPS1, SKS1, IPK1, IPS2, SKS2, IPK2, IPS3, SKS3, IPK3, IPK3, IPS4, SKS4, IPK4) secara simultan (bersama-sama) mempengaruhi model atau tidak.
2. Untuk melihat variabel bebas (Usia, Prodi, IPS1, SKS1, IPK1, IPS2, SKS2, IPK2, IPS3, SKS3, IPK3, IPK3, IPS4, SKS4, IPK4) secara parisal mempengaruhi model atau tidak, dan
3. Untuk melihat apakah model yang dibuat fit atau tidak.

Untuk dapat menjawab tujuan digunakan tes sebagai berikut:

1. *Omnibus test* : untuk melihat variabel bebas (Usia, Prodi, IPS1, SKS1, IPK1, IPS2, SKS2, IPK2, IPS3, SKS3, IPK3, IPK3, IPS4, SKS4, IPK4) secara bersama-sama (simultan) mempengaruhi model atau tidak.
2. Variabel *in the Equation* : digunakan untuk melihat Apakah variabel bebas (Usia, Prodi, IPS1, SKS1, IPK1, IPS2, SKS2, IPK2, IPS3, SKS3, IPK3, IPK3, IPS4, SKS4, IPK4) secara parisal (sendiri-sendiri) mempengaruhi model?
3. *Hosmer and Lomeshow test* : digunakan untuk melihat model yang dibuat fit atau tidak.

Adapun variabel yang diujikan disajikan pada Tabel 3.3:

Tabel 3.3 Variabel Penelitian

No	Variabel	Tipe Data	Nilai
1	Usia	Integer	Rentang nilai 17-24
2	Jenis kelamin	Kategori	Perempuan dan laki-laki
3	Prodi	Kategori	Sistem Informasi
4	IPS1	Real	0.00-4.00
5	SKS1	Integer	0-24
6	IPK1	Real	0.00-4.00
7	IPS2	Real	0.00-4.00
8	SKS2	Integer	0-24
9	IPK2	Real	0.00-4.00
10	IPS3	Real	0.00-4.00
11	SKS3	Integer	0-24
12	IPK3	Real	0.00-4.00
13	IPS4	Real	0.00-4.00
14	SKS4	Integer	0-24
15	IPK4	Real	0.00-4.00
16	Masa studi	Kategori	Tepat dan terlambat

Dan data yang digunakan untuk melakukan analisis prediktor sebanyak jumlah sampel yaitu 70 data, adapun contoh datanya pada Tabel 3.4:

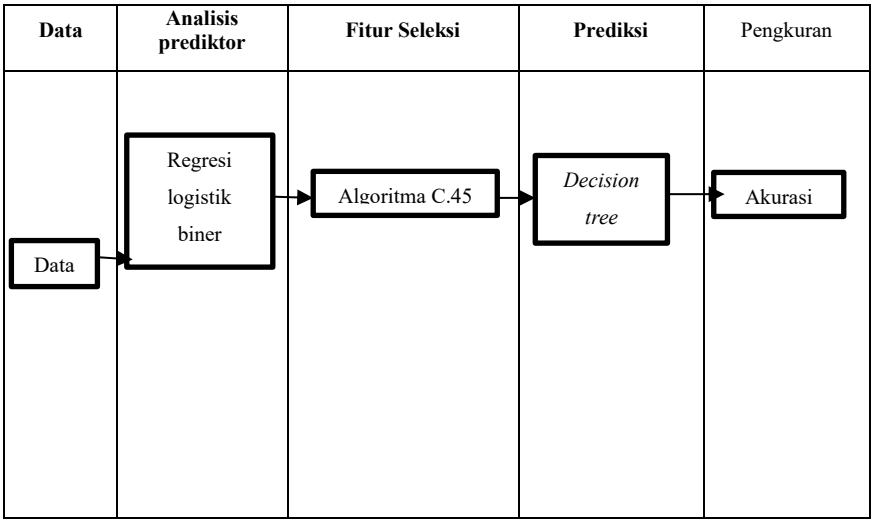
Tabel 3.4 Data Sampel

No	JK	Usia	Prodi	IPS1	SKS1	IPK1	IPK4	Masa Studi
1	P	21	Manajemen	3.86	21	3.86	3.5	terlambat
2	L	18	Sistem Informasi	3.05	19	3.05	3.58	terlambat
3	P	18	Sistem Informasi	3.11	19	3.11	2.71	terlambat
4	L	18	Sistem Informasi	1.79	19	1.79	2.07	terlambat
68	P	17	Manajemen	3.33	21	3.33	3.52	Tepat
69	P	18	Manajemen	3.48	21	3.48	3.24	Tepat
70	L	18	Manajemen	3.33	21	3.33	3.4	Tepat

Dari pengolahan data dengan metode regresi logistik biner didapatkan nilai pengaruh variabel bebas terhadap variabel terikat secara bersama-sama (parsial) mempengaruhi model dengan nilai sig 0,00 yang artinya nilai sig < 0.05 variabel bebas secara bersama sama mempengaruhi model. Sedangkan untuk pengujian pengaruh variabel bebas secara sendiri-sendiri (parsial) tidak mempengaruhi model dengan nilai sig > 0.05.

3.5 Prediksi Kelulusan

Prediksi kelulusan mahasiswa dilakukan dengan menggunakan metode *Decision Tree* C.45. Konstruksi pada *Decision Tree* dilakukan secara *top down* (dari atas kebawah) untuk solusinya. Gambar 3.2 merupakan proses prediksi kelulusan mahasiswa



Gambar 3.2 Proses Prediksi Kelulusan

Merujuk pada Gambar 3.2 data kelulusan mahasiswa STIMIK ESQ terlebih dahulu dilakukan analisis fitur menggunakan regresi logistik biner dengan tujuan untuk melihat tingkat pengaruh variabel bebas terhadap variabel terikat. Data yang digunakan sebanyak 70 data yang terdiri dari 35 data mahasiswa lulus tepat waktu dan 35 mahasiswa tidak lulus tepat waktu atau terlambat. Dari proses tersebut didapat nilai pengaruh dari tiap variabel bebas terhadap variabel terkait baik secara parsial maupun simultan.

Kemudian dilakukan seleksi fitur/variabel dengan menggunakan algoritma C4.5. Pada *Decision Tree* C4.5 pencarian atribut sebagai akar (*root*) dilakukan dengan mencari nilai gain ratio tertinggi. Pada proses pengklasifikasian *Decision Tree* semua data akan diuji dengan melakukan pelacakan pada jalur dari node akar (*root*) sampai dengan daun (*leaf)* dan kemudian akan diprediksi kelas yang dimiliki oleh data baru tersebut (Kasih, 2019).

Pengambilan *root node* didasarkan pada atribut yang memiliki nilai gain tertinggi, dengan cara perhitungan nilai *entropy* pada seluruh *case*, dan perhitungan *entropy* beserta gain pada setiap atribut. Perhitungan internal nodes juga didasarkan pada atribut yang memiliki nilai *gain* tertinggi, namun tanpa melibatkan atribut yang menjadi *root node.* Atribut yang menjadi *Leaf nodes* akan diambil dari atribut dengan nilai gain tertinggi yang nilai atributnya hanya masuk ke dalam satu kondisi. Berdasarkan perhitungan *Decision Tree* algoritma C4.5 didapatkan atribut yang menjadi root adalah IPK.

3.6 Evaluasi

Evaluasi terhadap model yang dibuat dilakukan dengan menggunakan *confusion matrix*. Evaluasi dilakukan dengan tujuan untuk melihat keberhasilan dari model yang digunakan. Akurasi didefinisikan sebagai tingkat kecocokan antara nilai prediksi, dan nilai aktual. Rumus yang digunakan untuk menghitung akurasi adalah

$$\text{Akurasi} = \frac{TP+TN}{TP+FP+TN+FN}$$

Keterangan :

e) TP (*True Positive*) : Jumlah data positif yang teridentifikasi benar
f) TN (*True Negative*) : Jumlah data negatif yang teridentifikasi benar
g) FN (*False Negative*) : Jumlah data negatif yang teridentifikasi salah
h) FP (*False Positive*) : Jumlah data Positif yang teridentifikasi salah

BAB 4
PEMBAHASAN DAN HASIL

4.1 Hasil

Pada sub bab ini akan dijelaskan pengumpulan data, seleksi data, *preprocessing, transformasi data,* proses data mining dan interpretasi.

4.1.1 Pengumpulan Data

Penelitian ini, data yang digunakan merupakan data kelulusan mahasiswa STIMIK ESQ tahun 2017-2020 sebanyak 186 yang terdiri dari 44 data mahasiswa lulusan tahun 2017, 45 data mahasiswa lulusan tahun 2018, 35 data mahasiswa lulusan tahun 2019, dan 62 data mahasiswa lulusan tahun 2020. Tabel 4.1 merupakan tabel jumlah data kelulusan mahasiswa STIMIK ESQ tahun 2017-2020.

Tabel 4.1 Jumlah Kelulusan Mahasiswa STIMIK ESQ 2017-2020

Tahun	prodi		Jumlah
	Sistem Informasi	Manajemen	
2017	17	27	44
2018	16	29	45
2019	14	21	35
2020	14	48	62
Jumlah	61	125	**186**

Data tersebut terdiri dari banyak atribut seperti pada Tabel 4.2:

Tabel 4.2 Atribut Data Kelulusan Mahasiswa STIMIK ESQ 2017-2020

No	Atribut	Keterangan
1	NIM	Nomor induk mahasiswa
2	Prodi	Predikat kelulusan
3	Nama	Nama mahasiswa
4	Jenis Kelamin	Jenis kelamin mahasiswa
5	Tempat, Tanggal lahir	Tempat dan tanggal lahir mahasiswa
6	Tahun Masuk	Tahun masuk mahasiswa
7	Jumlah SKS Lulus	Jumlah dan Satuan Kredit Semester Lulus
8	Predikat	Predikat kelulusan
9	Tanggal kelulusan	Tanggal kelulusan Mahasiswa
10	Tanggal Ijazah dan Transkrip	Tanggal Ijazah dan Transkip mahasiswa

No	Atribut	Keterangan
11	IPS 1-8	Indeks prestasi semester 1 - 8
12	SKS 1 - 8	Satuan kredit semester 1 - 8
13	IPK 1 - 8	Indeks prestasi kumulatif semester 1 - 8

Tabel 4.3 merupakan contoh data mahasiswa STIMIK ESQ lulusan tahun 2017-2020

Tabel 4.3 Contoh Data Mahasiswa

No	NIM	Prodi	Jenis Kelamin	Tempat, tanggal lahir	SKS 8	IPK 8
1	1410120003	Sistem Informasi	L	Tangerang, 19 September 1996	9	3,42
2	1410120018	Sistem Informasi	L	Jakarta, 12 April 2020	9	3,75
3	1410120019	Sistem Informasi	L	Jakarta, 9 Agustus 1996	17	2,81
4	1410120020	Sistem Informasi	L	Selangor, Malaysia, 1 Juli 1996	19	2,55
5	1410110001	Manajemen	P	Jakarta, 24 April 1996	21	2,79
6	1410110004	Manajemen	L	Serang, 7 Juni 1995	12	2,53
186	1410120014	Sarika Fitri	P	Pekanbaru, 19 Juni 1996	12	3,89

4.1.2 Seleksi Data

Pada tahap ini dilakukan seleksi pada data yang akan digunakan. Seleksi tersebut bertujuan untuk memilih variabel yang mempengaruhi kelulusan tepat waktu. analisis dilakukan dengan menggunakan bantuan aplikasi SPSS dengan teknik Regresi Logistik biner. Regresi Logistik Biner merupakan suatu metode analisis statistik yang berguna untuk menganalisis hubungan antar suatu variabel respon dengan beberapa prediktor. Dengan variabel responnya berupa data kualitatif dikotomi yaitu bernilai 1 untuk menyatakan keberadaan suatu karakteristik dan 0 untuk menyatakan ketidakberadaan sebuah karakteristik (Tampil dkk., 2015). Pada penelitian ini angka 0 menyatakan lulus tepat waktu dan 1 menyatakan terlambat.

Adapun variabel yang diuji dengan menggunakan metode regresi logistik biner adalah Jenis Kelamin, Usia saat masuk yang diketahui dari pengurangan tahun masuk dengan tahun lahir, Prodi, Indeks prestasi semester satu sampai dengan semester 4, Satuan Kredit semester 1 sampai dengan semester 4, Indeks prestasi kumulatif semester 1 sampai dengan semester 4 dan Masa studi sebagai label.

Tabel 4.4

No	Variabel	Keterangan	Status
1	Jenis Kelamin	Jenis kelamin mahasiswa	Bebas
2	Usia	Usia saat masuk (didapatkan dari pengurangan tahun masuk dan tahun lahir)	Bebas
3	Prodi	Prodi (Sistem Informasi dan Manajemen)	Bebas
4	IPS1	Indeks Prestasi Semester 1	Bebas
5	SKS1	Satuan kredit semester 1	Bebas
6	IPK1	Indeks Prestasi Kumulatif semester 1	Bebas
7	IPS2	Indeks Prestasi Semester 2	Bebas
8	SKS2	Satuan kredit semester 2	Bebas
9	IPK2	Indeks Prestasi Kumulatif semester 2	Bebas
10	IPS3	Indeks Prestasi Semester 3	Bebas
11	SKS3	Satuan kredit semester 3	Bebas
12	IPK3	Indeks Prestasi Semester 3	Bebas
13	IPS4	Indeks Prestasi Semester 4	Bebas
14	SKS4	Satuan kredit semester 4	Bebas
15	IPK4	Indeks Prestasi Kumulatif semester 4	Bebas
16	Masa Studi	Lama masa studi	Target

Atribut Prediktor Kelulusan Mahasiswa STIMIK ESQ

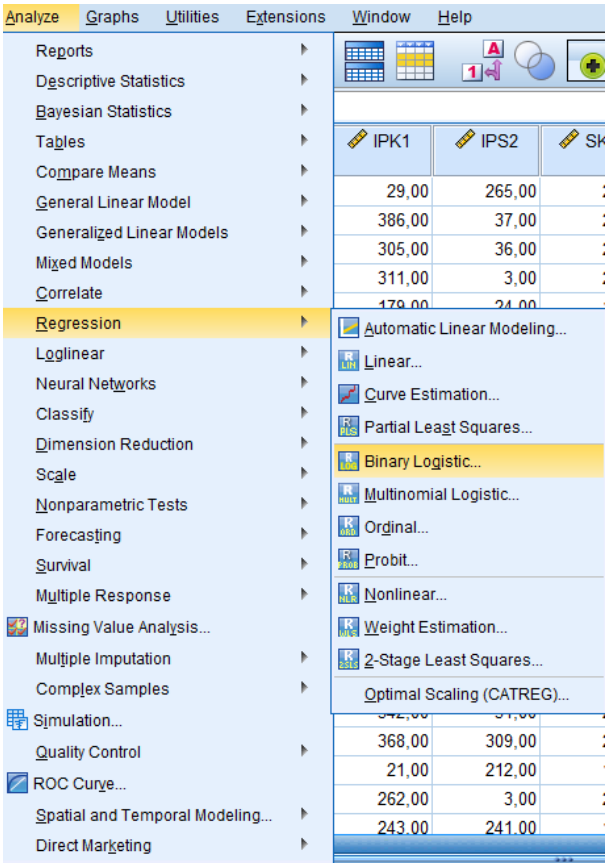
Enam belas variabel tersebut diuji untuk menentukan atribut yang paling berpengaruh terhadap kelulusan tepat waktu, pengujian didasarkan pada nilai signifikasi, atribut dikatakan berpengaruh jika memiliki nilai signifikasi <=0,05. Adapun tahapan pengujiannya ditunjukan pada prototype sebagai berikut:

1. Import data penelitian yang akan diolah.
2. Menentukan nama dan tipe data atribut

Tabel 4.5 Nama dan Tipe data Atribut

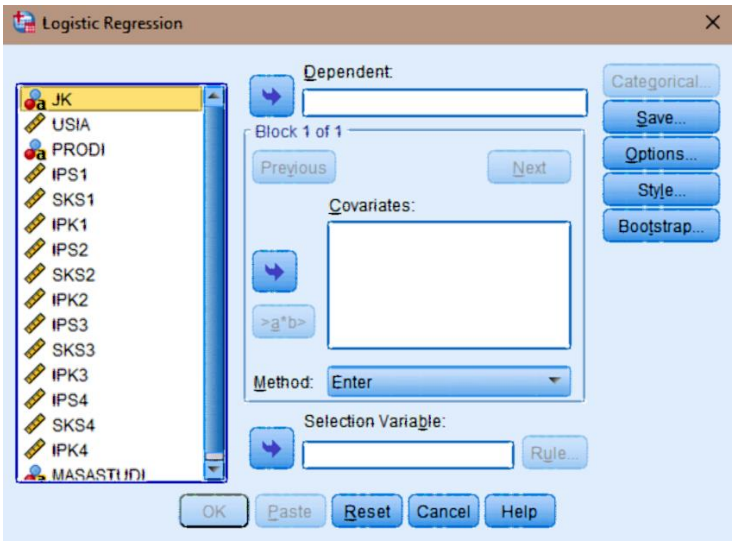
Nama Atriut	Tipe Data
Jenis Kelamin	String (laki-laki dan perempuan)
Usia	Integer
Prodi	String (manajemen dan sistem informasi)
IPS1	Real
SKS1	Integer
IPK1	Real Real
IPS2	Real
SKS2	Integer
IPK2	Real
IPS3	Real
SKS3	Integer
IPK3	Real
IPS4	Real
SKS4	Integer
IPK4	Real
Masa Studi	String (lulus dan terlambat)

Kemudian pilih analyze dan pilih regresi logistik biner



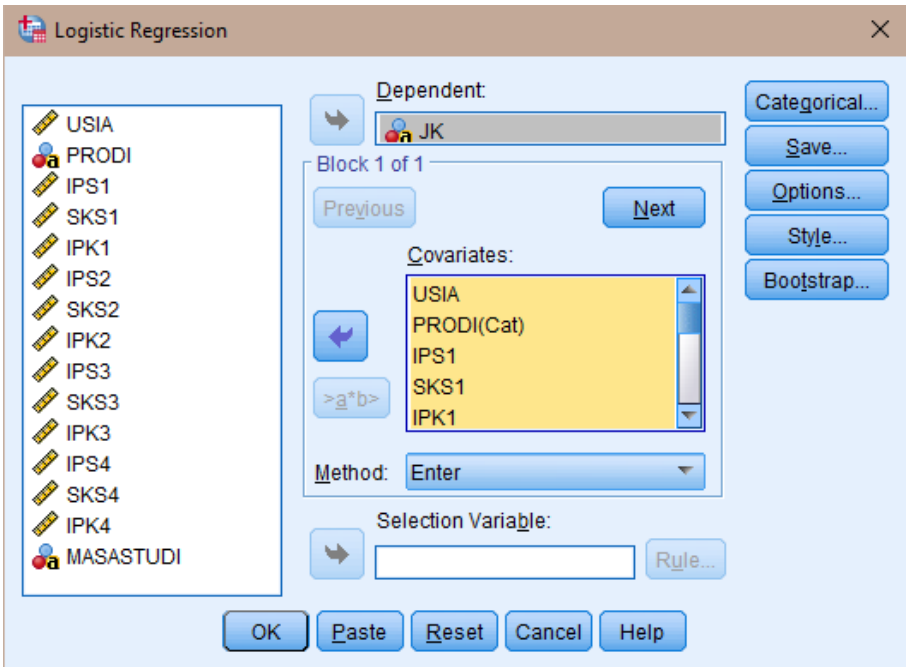
Gambar 4.1 Regression *Binary Logistic*

Kemudian akan muncul tampilan seperti pada Gambar 4.1:



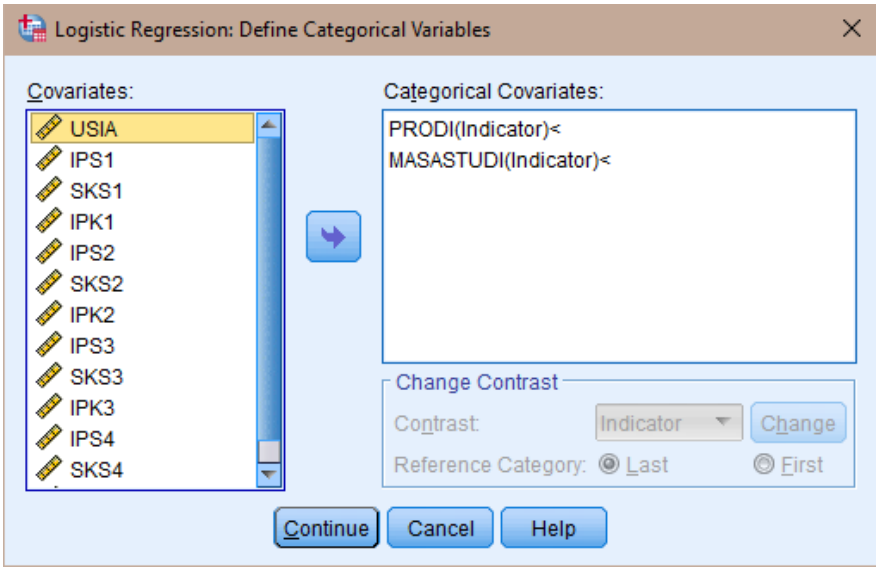
Gambar 4.2 Pemilihan Variabel Dependen dan Independen

3. Isikan variabel masa studi ke dalam kotak `dependent` dan variabel JK (Jenis Kelamin), Usia, Prodi, IPS1, SKS1, IPK1, IPS2, SKS2, IPK2, IPS3, SKS3, IPK3, IPK3, IPS4, SKS4, IPK4 ke dalam kotak `Covariates`.



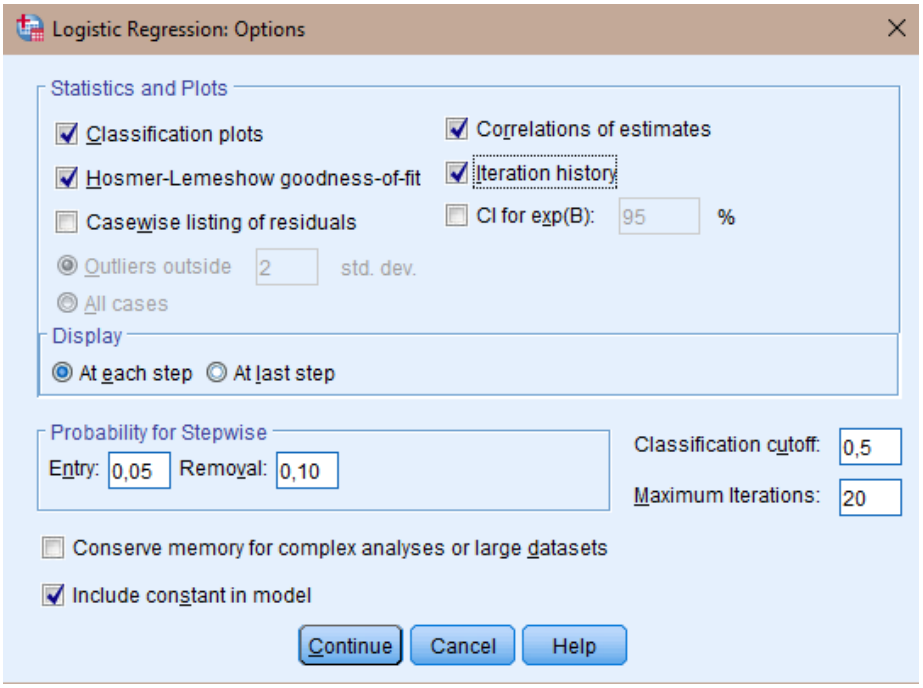
Gambar 4.3 Variabel Dependen dan Dependen

4. Klik tombol `categorical,` kemudian masukan variabel JK (Jenis Kelamin) dan Prodi ke dalam kotak *`categorical covariates`* kemudian klik *`continue.`*



Gambar 4.4 Variabel Kategori

5. Klik tombol *Options* kemudian centang pada opsi classification plots, Hosmer and Lomeshow test, correlation of estimates, iteration history kemudian klik continue.



Gambar 4.5 *Statistic and Plots*

6. SPSS menampilkan *output* yang berisi beberapa informasi berikut:
   1) Tabel Omnibus *test*

Tabel 4.6 Omnibus *test*

Omnibus *Tests* of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	55,209	14	,000
	Block	55,209	14	,000
	Model	55,209	14	,000

Pada Tabel 4.6 dapat dilihat bahwa nilai sig <0,05 yang artinya secara bersama-sama variabel bebas terbukti mempengaruhi model.

2) Tabel Variabel *in the Equation*

Tabel 4.7 *Variabel in the Equation*

Variables in the Equation	
	Sig.
JK(1)	,653
USIA	,369
PRODI(1)	,570
IPS1	,895
IPK1	,033
IPS2	,630
SKS2	,754
IPK2	,898
IPS3	,427
SKS3	,027
IPK3	,818
IPS4	,532
SKS4	,008
IPK4	,269
Constant	,069

Pada Tabel 4.7 dapat dilihat nilai sig >= 0,05 yang artinya secara parsial variabel bebas tidak mempengaruhi model. Dari ke-15 variabel tersebut yang mendekati nilai <=0,05 adalah variabel SKS4, SKS3, IPK1, dan seterusnya. Hal tersebut menunjukan bahwa SKS4 memiliki nilai yang paling berpengaruh terhadap kelulusan mahasiswa.

3) Tabel Hosmer and Lomeshow *test*

Tabel 4.8 Hosmer and Lomeshow test

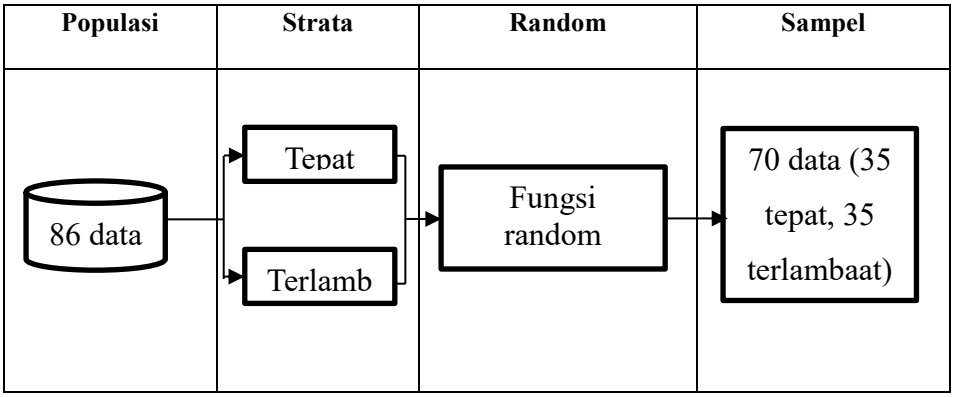
Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	5,266	8	,729

Pada Tabel 4.8 terlihat bahwa nilai sig > 0.05 yang artinya model yang dibuat terbukti fit.

4.1.3 *Preprocessing*

Data yang didapatkan dari pihak BAA (Biro Administrasi Akademik) STIMIK ESQ sebanyak 186 data namun tidak semua data dapat digunakan dalam penelitian. Data yang dilibatkan dalam penelitian hanya data yang bersih

dari *missing value,* dan data yang tidak relevan. Dari total 186 data didapatkan 86 data yang bersih dari *missing value* dan data yang tidak relevan. pembersihan data dilakukan secara manual dengan mengeliminasi data yang memiliki *missing value* dan data yang tidak relevan. Dari 86 data tersebut akan diambil sebagaian data sebagai sampel, proses pengambilan sampel dilakukan dengan menggunakan metode *stratified random sampling.* Gambar 4.6 merupakan



Gambar **4.6** ***Prototype Stratified Random Sampling*** *prototype stratified random sampling.*

Berdasarkan Gambar 4.6, populasi yang digunakan sebanyak 86 data, dimana data tersebut terbagi kedalam dua strata berdasarkan masa studi yaitu tepat dan terlambat. Pengolahan proses *stratified random sampling* dilakukan dengan bantuan aplikasi Microsoft Excel. Pengambilan anggota sampel dilakukan secara acak (*random)* dengan menggunakan fungsi random yang telah tersedia di Microsoft Excel. Penentuan jumlah sampel mengacu pada tabel Krejcie-Morgan, Tabel 4.9 merupakan tabel krejcie morgan:

Tabel 4.9 Krejcie-Morgan

N	S	N	S	N	S	N	S
10	10	35	32	60	52	85	70
15	14	40	46	65	56	90	73
20	19	45	40	70	59	95	76
25	24	50	44	75	63	100	80
30	36	55	48	80	66	110	86

Berdasarkan Tabel 4.9 dari jumlah populasi 86 data didapatkan sampel sebanyak 70 data, dimana data tersebut terdiri dari 35 data mahasiswa lulus tepat waktu dan 35 mahasiswa lulus tidak tepat waktu atau terlambat.

4.1.4 Transformasi Data

Pada tahap ini data diubah ke dalam format yang sesuai untuk diproses. Beberapa metode seperti clustering dan asosiasi hanya mampu mengolah data dalam bentuk kategorikal. Maka pada tahap ini data dari setiap atribut ditransformasikan ke dalam bentuk kategori agar data dapat diproses dan data berada di rentang nilai yang sama, kemudian dinotasikan dengan rentang angka 0-4. Adapun transformasi yang dilakukan adalah sebagai berikut:

1. Jenis Kelamin, atribut ini memiliki dua nilai yaitu laki-laki dan perempuan yang ditransformasikan sebagai berikut:

Tabel 4.10 Transformasi Atribut Jenis Kelamin

Jenis Kelamin	Transformasi
Laki-laki	0
Perempuan	1

2. Prodi, atribut ini memiliki dua nilai yaitu sistem informasi yang ditransformasikan sebagai berikut:

Tabel 4.11 Transformasi Atribut Prodi

Prodi	Transformasi
Manajemen	1
Sistem Informasi	0

3. Usia, atribut usia ditransformasikan sebagai berikut:

Tabel 4.12 Transformasi Atribut Usia

Usia	Transformasi
<=15	3
16-18	2
19-21	1
>=21	0

4. Indeks Prestasi Semester dan Indeks Prestasi Kumulatif, atribut ini memiliki rentang nilai dari 0.00 sampai dengan 4.00. agar atribut ini memiliki nilai dengan rentang nilai yang sama maka dilakukan transformasi dengan ketentuan seperti pada Tabel 4.13:

Tabel 4.13 Transformasi Atribut IPS (Indeks Prestasi Semester)

Indeks Prestasi	Notasi
≥3.51	0
3.01-3.50	1
2.76-3.00	2
2.00-2.75	3
≥1.99	4

Rentang nilai tersebut didasarkan pada panduan akademik STIMIK ESQ tahun 2020 seperti pada Tabel 4.14:

Tabel 4.14 Predikat Kelulusan

Indeks Prestasi	Predikat kelulusan
≥3,51	Pujian (Cum laude)
3,01≥3,50	Sangat memuaskan
2.76≥3,00	Memuaskan
≥2,00	Lulus

5. Satuan Kredit Semester (SKS), atribut ini memiliki rentang nilai antara 0 sampai dengan 24. Agar atribut SKS memiliki nilai dengan rentang yang sama dengan atribut lain maka dilakukan transformasi dengan ketentuan sebagai berikut :

Tabel 4.15 Transformasi SKS (Sistem Kredit Semester)

Jumlah SKS	Notasi
22-24	0
19-21	1
16-18	2
≤15	3

Rentang nilai tersebut didasarkan pada ketentuan akademik tahun 2020.

6. Masa Studi, atribut ini memiliki dua nilai yaitu “Tepat” yang ditransformasikan sebagai berikut:

Tabel 4.16 Transformasi Atribut Masa Studi

Masa Studi	Transformasi
Tepat (≤ 4 Tahun)	1
Terlambat (>4 tahun)	0

Tabel 4.17 merupakan tabel data sebelum dilakukan transformasi

Tabel 4.17 Data Mahasiswa Sebelum Di Tranformasi

No	JK	Usia	Prodi	IPS1	SKS1	IPK1	IPK4	Masa Studi
1	P	21	Manajemen	3,86	21	3,86	3,5	terlambat
2	L	18	Sistem Informasi	3,05	19	3,05	3,58	terlambat
3	P	18	Sistem Informasi	3,11	19	3,11	2,71	terlambat
4	L	18	Sistem Informasi	1,79	19	1,79	2,07	terlambat
5	L	18	Manajemen	2,86	21	2,86	2,46	terlambat
185	P	18	Manajemen	3,48	21	3,48	3,24	Tepat
186	L	18	Manajemen	3,33	21	3,33	3,4	Tepat

Tabel 4.18 data kelulusan mahasiswa STIMIK ESQ 2017-2020 setelah dilakukan transformasi

Tabel 4.18 Data Mahasiswa Setelah Ditransformasi

No	JK	Usia	Prodi	IPS1	SKS1	IPK1	IPK4	Masa Studi
1	1	1	1	0	1	0	1	0
2	0	2	0	1	1	1	0	0
3	1	2	0	1	1	1	3	0
4	0	2	0	4	1	4	3	0
5	0	2	1	2	0	2	3	0
185	1	18	1	1	0	1	1	1
186	0	18	1	1	0	1	1	1

4.1.5 Proses *Data mining*

Berdasarkan data dan atribut yang telah didapat dari proses sebelumnya, proses selanjutnya adalah melakukan pengolahan data dengan metode *Decision Tree*. Untuk melakukan prediksi dibutuhkan data latih dan data uji. Data latih berfungsi sebagai data pembelajaran dan data uji berfungsi untuk menguji model yang dihasilkan dari data latih.

Pengujian data kelulusan mahasiswa untuk memprediksi kelulusan mahasiswa tepat waktu dengan metode *Decision Tree* dilakukan dengan menggunakan bahasa pemrograman python pada *tools google colab*. Hasil dari pengujian akan didapatkan nilai akurasi metode *Decision Tree* untuk memprediksi kelulusan tepat waktu mahasiswa STIMIK ESQ. Adapun langkahnya adalah sebagai berikut:

1. `Import library` yang dibutuhkan

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
```

Source code tersebut merupakan perintah untuk memanggil `library` yang akan digunakan agar perintah yang akan dijalankan dapat berjalan dan pembuatan model dapat dieksekusi.

2. `Import` data yang akan digunakan

Setelah melakukan pemanggilan `library`, selanjutnya memanggil data yang akan digunakan, adapun *source* yang adalah sebagai berikut:

```
df= pd.read_csv('DATA.csv', delimiter=";", names = [ 'JenisKelamin',
 'usia', 'Prodi', 'IPS1', 'SKS1', 'IPK1', 'IPS2', 'SKS2', 'IPK2','IP
S3', 'SKS3', 'IPK3', 'IPS4', 'SKS4', 'IPK4', 'MasaStudi'])
```

Pada *source code* tersebut menunjukan pemanggilan data yang akan digunakan, data yang digunakan yaitu DATA.csv yang sebelumnya telah diupload. Data yang digunakan memiliki format csv dengan pemisah “;”. Dengan fitur yang digunakan yaitu JenisKelamin, usia, Prodi, IPS1, SKS1, IPK1, IPS2, SKS2, IPK2, IPS3, SKS3, IPK3, IPS4, SKS4, IPK4, MasaStudi.

3. Melihat bentuk data yang digunakan

```
df.shape
```

Output dari *source code* di atas adalah sebagai berikut:

```
(70, 16)
```

Output di atas menunjukan bahwa data yang digunakan sebanyak 70 data dengan atribut sebanyak 16. Dimana 15 atribut merupakan atribut prediktor dan 1 atribut merupakan atribut target.

4. Menentukan atribut target, menentukan jumlah data *test*ing dan training
   Proses selanjutnya adalah menentukan atribut target, membagi data kedalam data *training* (data latih) dan data *test*ing (data uji). Berikut *source code* yang digunakan

```
X = df.drop(['MasaStudi'], axis=1)
y = df['MasaStudi']
X_train, X_test, y_train, y_test = train_test_split(
X, y, test_size = 0.30, random_state = 42)
X_train.shape, X_test.shape
```

   Source code tersebut menunjukan bahwa atribut target adalah atribut 'MasaStudi', dengan jumlah data *testing* sebanyak 30% atau setara dengan 21 data yang diambil secara random.

5. Pemodelan dengan *Decision Tree*
   Pada tahap ini dilakukan pemodelan dengan menggunakan *Decision Tree* C4.5, dengan criterion 'entropy' dan maksimal kedalaman tree adalah 7, dengan random statenya adalah 0. Berikut *source code* yang digunakan

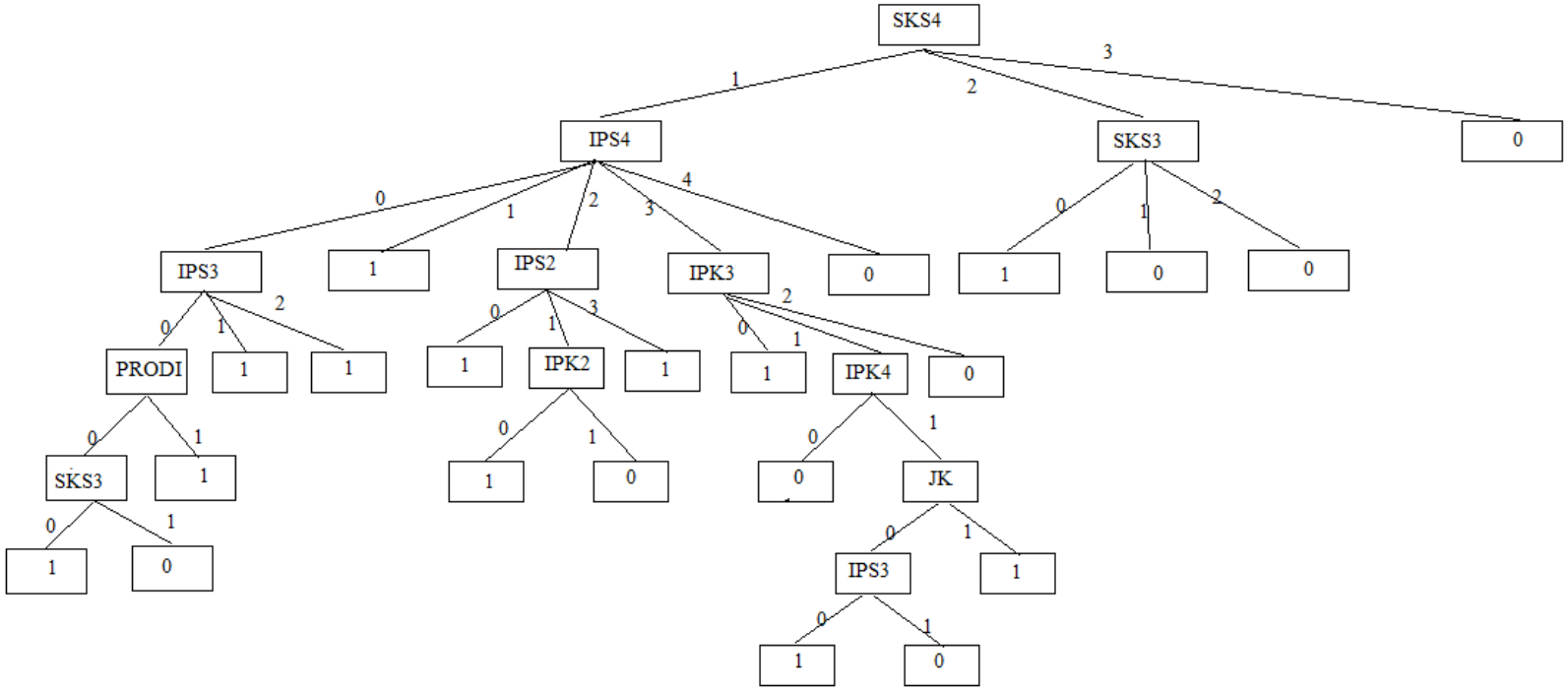
```
clf_en = DecisionTreeClassifier(criterion='entropy',
 max_depth=7, random_state=0)
clf_gini.fit(X_train, y_train)
```

6. Visualisasi model *Decision Tree* C4.5

```
import graphviz
dot_data = tree.export_graphviz(clf_en, out_file=No
ne,feature_names=X_train.columns,class_names=y_trai
n,                          filled=True, rounded=True,
special_characters=True
graph = graphviz.Source(dot_data)
graph
```

   Untuk mempermudah pembacaan *output* yang dihasilkan, maka diperlukan visualisasi kedalam sebuah bentuk pohon keputusan. Adapun *source code* yang digunakan untuk pembentukan pohon keputusan adalah sebagai berikut:

Output dari *source code* tersebut adalah sebagai berikut:



Gambar 4.7 Pohon Keputusan Prediksi Kelulusan Mahasiswa

Keterangan :

JK (Jenis Kelamin)	:	0 (Laki-laki) dan 1 (Perempuan)
Usia	:	0 (>=21), 1 (19-21), 2 (16-21), dan 3 (<=15)
Prodi	:	0 (Sistem informasi) dan 1 (Manajemen)
IPK dan IPS (1-4)	:	0 (>=3.51), 1 (3.01-3.50), 2 (2.76-3.00), 3 (2.00-2.75), dan 4 (<=1.99)
SKS (1-4)	:	0 (22-24), 1 (19-21), 2 (16-18), dan 3 (<=15)
Masa Studi	:	0 (Terlambat) dan 1 (Tepat)

Gambar 4.7 menunjukan bahwa atribut SKS4 menjadi *root node* sehingga menjadi atribut yang paling berpengaruh dalam prediksi kelulusan mahasiswa.

7. Pengujian model dengan *confusion matrix*

Tahapan terakhir dalam prediksi kelulusan mahasiswa STIMIK ESQ adalah pengujian model prediksi. Pengujian dilakukan dengan tujuan untuk melihat data

```
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred_en)
print('Confusion matrix\n\n', cm)
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred_en))
```

Output dari kode tersebut adalah sebagai berikut

Tabel 4.19 *Confusion Matrix*

Prediction value	**Actual Value**	
	Positive class	**Negative class**
Positive class	10	2
Negative class	0	9

Berikut merupakan penjelasan Tabel 4.19:

a. Jumlah data positif yang teridentifikasi benar (*True Positive)* sebanyak 10 data
b. Jumlah data negatif yang teridentifikasi benar (*True Negative*) 0 data.

c. Jumlah data negatif yang teridentifikasi salah (*False Negative*) sebanyak 9 data

d. Jumlah data Positif yang teridentifikasi salah (*False Positive*) hanya 2 data

```
from sklearn.metrics import classification_report

print(classification_report(y_test, y_pred_en))
```

```
              precision    recall  f1-score   support

           0       1.00      0.83      0.91        12
           1       0.82      1.00      0.90         9

    accuracy                           0.90        21
   macro avg       0.91      0.92      0.90        21
weighted avg       0.92      0.90      0.91        21
```

Output di atas menunjukan bahwa tingkat akurasi dari model yang dibuat adalah 90%, akurasi tersebut menunjukan bahwa model yang dibuat memiliki tingkat akurasi yang sangat baik sehingga dapat digunakan untuk meakukan prediksi kelulusan maha.

4.1.6 Interpretasi

Berdasarkan pemodelan yang dilakukan dengan *Decision Tree* C4.5 dengan data sebanyak 70 data. Dimana data uji yang digunakan sebanyak 30% atau setara dengan 21 data dan data latih sebanyak 70% atau setaa dengan 49 data. Model tersebut menghasilkan akurasi sebesar 90% dengan SKS4 sebagai *root* atau sebagai atribut yang paling berpengaruh. Adapun interpretasi dari model yang telah dibuat disajikan pada Tabel 4.20 berikut:

Tabel 4.20 Interpretasi

Program Studi	Rules
Sistem Informasi	1. Jika SKS4 sebanyak 19-21 SKS, IPS4 lebih besar atau sama dengan dari 3.51, IPS3 lebih besar atau sama dengan 3.51, Prodi Sistem Informasi, SKS3 sebanyak 22-24 maka diprdiksi “lulus tepat waktu”
Manajemen	1. Jika SKS4 sebanyak 19-21 SKS, IPS4 lebih besar atau sama dengan dari 3.51, IPS3 lebih besar atau sama dengan 3.51, Prodi

	Sistem Informasi, SKS3 sebanyak 19-21 maka diprediksi “tidak lulus tepat waktu atau terlambat”
Sistem Informasi dan Manajemen	1. Jika SKS4 sebanyak 19-21 SKS, IPS4 lebih besar atau sama dengan dari 3.51, IPS3 lebih besar atau sama dengan 3.51, Prodi Manajemen maka diprediksi “lulus tepat waktu” 2. Jika SKS4 sebanyak 19-21 SKS, IPS4 lebih besar atau sama dengan dari 3.51, IPS3 berada diantara rentang nilai 3.01-3.50 maka diprediksi “lulus tepat waktu” 3. Jika SKS4 sebanyak 19-21 SKS, IPS4 lebih besar atau sama dengan dari 3.51, IPS3 berada diantara rentang nilai 2.76-3.00 maka diprediksi “lulus tepat waktu” 4. Jika SKS4 sebanyak 19-21 SKS, IPS4 berada pada rentang nilai 3.01-3.50 maka diprediksi “lulus tepat waktu” 5. Jika SKS4 sebanyak 19-21 SKS, IPS4 berada pada rentang nilai 2.76-3.00, dan IPS2 lebih dari atau sama dengan 3.51 maka diprediksi “lulus tepat waktu” 6. Jika SKS4 sebanyak 19-21 SKS, IPS4 berada pada rentang nilai 2.76-3.00, dan IPS2 lebih dari atau sama dengan 3.51, IPK2 lebih dari 3.51 maka diprediksi “lulus tepat waktu” 7. Jika SKS4 sebanyak 19-21 SKS, IPS4 berada pada rentang nilai 2.76-3.00, dan IPS2 lebih dari atau sama dengan 3.51, IPK2 berada pada rentang nilai 3.01-3.50 21 maka diprediksi “tidak lulus tepat waktu atau terlambat” 8. Jika SKS4 sebanyak 19-21 SKS, IPS4 berada pada rentang nilai 2.76-3.00, dan IPS2 berada pada rentang niali 2.00-2.75 maka diprediksi “lulus tepat waktu” 9. Jika SKS4 sebanyak 19-21 SKS, IPS4 berada pada rentang nilai 2.00-2.75, IPK3 lebih dari atau sama dengan 3.51 maka diprediksi “lulus tepat waktu” 10. Jika SKS4 sebanyak 19-21 SKS, IPS4 berada pada rentang nilai 2.00-2.75, IPK3 berada pada rentang nilai 3.01-3.50 ,IPK4 lebih dari atau sama dengan 3.51 maka diprediksi “lulus tepat waktu” 11. Jika SKS4 sebanyak 19-21 SKS, IPS4 berada pada rentang nilai 2.00-2.75, IPK3 berada pada rentang nilai 3.01-3.50 , IPK4 berada pada rentang nilai 3.01-3.50, JK (Jenis Kelamin) laki-laki, IPS3 lebih dari 3.51 maka diprediksi “lulus tepat waktu” 12. Jika SKS4 sebanyak 19-21 SKS, IPS4 berada pada rentang nilai 2.00-2.75, IPK3 berada pada rentang nilai 3.01-3.50 , IPK4 berada

	pada rentang nilai 3.01-3.50, JK (Jenis Kelamin) laki-laki, IPS3 berada pada rentang nilai 3.01-3.51 maka diprediksi “tidak lulus tepat waktu atau terlambat” 13. Jika SKS4 sebanyak 19-21 SKS, IPS4 berada pada rentang nilai 2.00-2.75, IPK3 berada pada rentang nilai 3.01-3.50 , IPK4 berada pada rentang nilai 3.01-3.50, JK (Jenis Kelamin) maka diprediksi “lulus tepat waktu” 14. Jika SKS4 sebanyak 19-21 SKS, IPS4 berada pada rentang nilai 2.00-2.75, IPK3 berada pada rentang nilai 2.76-3.00 maka diprediksi “lulus tepat waktu” 15. Jika SKS4 sebanyak 19-21 SKS, IPS4 lebih kecil dari atau sama denga 1.99 maka diprediksi “tidak lulus tepat waktu atau terlambat” 16. Jika SKS4 sebanyak 19-21 SKS, SKS3 berada pada rentang 22-24 SKS maka diprediksi “lulus tepat waktu” 17. Jika SKS4 sebanyak 19-21 SKS, SKS3 berada pada rentang 19-21 “tidak lulus tepat waktu atau terlambat” 18. Jika SKS4 sebanyak 19-21 SKS, SKS3 berada pada rentang 16-18 maka diprediksi “tidak lulus tepat waktu atau terlambat” 19. Jika SKS4 kurang atau sama dengan 15 maka diprediksi “tidak lulus tepat waktu atau terlambat”

BAB 5
PENUTUP

5.1 Kesimpulan

Berdasarkan seluruh hasil dari tahapan penelitian yang telah dilakukan pada penerapan metode *Decision Tree* C4.5 untuk prediksi kelulusan mahasiswa STIMIK ESQ dapat ditarik kesimpulan sebagai berikut:

1. Faktor ayng mempengaruhi kelulusan tepat waktu di STIMIK ESQ berdasarkan pohon keputusan *decision tree C4.5* yaitu SKS4, IPS4, SKS3, IPS3, IPS2, IPK3, Prodi, IPK2, IPK4, dan Jenis Kelamin.
2. Prediksi kelulusan mahasiswa dapat dilakukan dengan menggunakan pendekatan *data mining* dengan metode *decision tree* C4.5 yang menghasilkan 21 *rules* dengan tingkat akurasi sebesar 90.

5.2 Saran

Berdasarkan hasil pengamatan dan analisa selama melakukan penelitian prediksi kelulusan mahasiswa STIMIK ESQ sarannya adalah sebagai berikut :

1. Penelitian selanjutnya lebih baik menggunakan data yang lebih banyak untuk menghasilkan peraturan dengan nilai akurasi yang jauh lebih baik.
2. Untuk menemukan metode yang lebih cocok pada penelitian selanjutnya lebih baik menggunakan metode yang lain.
3. Analisia yang dilakukan pada penelitian ini merupakan analisis yang sangat mendasar, maka perlu dikembangkan kembali. Akan lebih bagus jika pada penelitian selanjutnya prediksi Dikembangkan dalam bentuk sistem agar semakin mempermudah prediksi kelulusan STIMIK ESQ.

DAFTAR PUSTAKA

Amelia, M. winny, Lumenta, A. S. ., & Jacobus, A. (2017). Prediksi Masa Studi Mahasiswa dengan Menggunakan Algoritma Naïve Bayes. *Jurnal Teknik Informatika*, *11*(1). https://doi.org/10.35793/jti.11.1.2017.17652

Arieska, P. K., Herdiani, N., Sampling, S., & Relatif, E. (2018). *PEMILIHAN TEKNIK SAMPLING BERDASARKAN*. *6*(2).

Badan Akreditasi Nasional Perguruan Tinggi. (2019). *Naskah IAPT 3.0*. *April*, 7–9.

Etriyanti, E., Syamsuar, D., & Kunang, N. (2020). *Implementasi Data Mining Menggunakan Algoritme Naive Bayes Classifier dan C4 . 5 untuk Memprediksi Kelulusan Mahasiswa*. *13*(1), 56–67.

Fahmi, F., & Khikmah, L. (2018). *Analisis Regresi Logistik Terhadap Faktor yang Mempengaruhi Penggunaan Kontrasepsi pada Survey Demografi Kesehatan Indonesia 2012*. *52*, 122–130.

Fajrin, A. A., Maulana, A., Informatika, T., Batam, U. P., & Soeprapto, J. R. (2018). *PENERAPAN DATA MINING UNTUK ANALISIS POLA PEMBELIAN KONSUMEN DENGAN ALGORITMA FP- GROWTH PADA DATA TRANSAKSI PENJUALAN*. *05*(01), 27–36.

Firdaus, D. (2017). Penggunaan Data Mining dalam Kegiatan Sistem Pembelajaran Berbantuan Komputer. *Jurnal Format*, *6*(2), 91–97.

Galih, G. (2019). Data Mining di Bidang Pendidikan untuk Analisa Prediksi Kinerja Mahasiswa dengan Komparasi 2 Model Klasifikasi pada STMIK Jabar. *Jurnal Teknologi Sistem Informasi dan Aplikasi*, *2*(1), 23. https://doi.org/10.32493/jtsi.v2i1.2643

Haidar, L. R., Sediyono, E., & Irani, A. (2019). *ANALISA PREDIKSI MAHASISWA DROP OUT MENGGUNAKAN METODE DECISION TREE DENGAN*. *17*(2), 97–106.

Hamidah, M., Fitriyah, H., & Arwani, I. (2019). *Implementasi Decision Tree pada Penentuan Kondisi Ruang Berasap Menggunakan Multi-Sensor Berbasis Arduino Uno*. *3*(4), 3845–3854. http://j-ptiik.ub.ac.id

Hanifah, I., & Prastowo, B. N. (2016). Uji GPS Tracking Dalam Skala Transportasi

Antar Kota. *IJEIS (Indonesian Journal of Electronics and Instrumentation Systems)*, *6*(2), 175. https://doi.org/10.22146/ijeis.15257

Harman, R. (2018). *Jurnal Ilmiah Informatika (JIF)*.

Hermanto, B., & SN, A. (2017). Klasifikasi Nilai Kelayakan Calon Debitur Baru Menggunakan Decision Tree C4.5. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *11*(1), 43. https://doi.org/10.22146/ijccs.15946

Hermawan, H. (2019). *Riset Hospitalitas Metode Kuantitatif untuk Riset Bidang Kepariwisataan*. https://doi.org/10.31227/osf.io/fcnzh

Heryana, D. (2019). *Data mining*.

Kamal, I. M., Hendro, T., & Ilyas, R. (2017). Prediksi Penjualan Buku Menggunakan Data Mining Di Pt. Niaga Swadaya. *Seminar Nasional Teknologi Informasi & Multimedia*, *02*(February), 49–54. http://ojs.amikom.ac.id/index.php/semnasteknomedia/article/view/1712

Kasih, P. (2019). Pemodelan Data Mining Decision Tree Dengan Classification Error Untuk Seleksi Calon Anggota Tim Paduan Suara. *Innovation in Research of Informatics (INNOVATICS)*, *1*(2), 63–69. https://doi.org/10.37058/innovatics.v1i2.918

Mardisetosa, B., Khusaini, K., & Gumelar Widia Asmoro. (2020). Personality, Gender, Culture, and Entrepreneurial Intentions of Undergraduate Student: Binary Logistic Regression. *Jurnal Pendidikan Ekonomi Dan (JPEB)*, *8*(2), 128–143. https://doi.org/10.21009/jpeb.008.2.5

Maulana, A., & Fajrin, A. A. (2018). Penerapan Data Mining Untuk Analisis Pola Pembelian Konsumen Dengan Algoritma Fp-Growth Pada Data Transaksi Penjualan Spare Part Motor. *Klik - Kumpulan Jurnal Ilmu Komputer*, *5*(1), 27. https://doi.org/10.20527/klik.v5i1.100

Maulana, M. S., Sabarudin, R., & Nugraha, W. (2019). Prediksi Ketepatan Kelulusan Mahasiswa Diploma dengan Komparasi Algoritma Klasifikasi. *Jurnal Sistem dan Teknologi Informasi (JUSTIN)*, *7*(3), 202. https://doi.org/10.26418/justin.v7i3.33316

Misna, Rais, & Utami, I. T. (2018). Analisis Regresi Logistik Biner Untuk Mengklasifikasi Penderita Hipertensi Berdasarkan Kebiasaan Merokok Di RSU Mokopido Toli-Toli. *Natural Science: Journal of Science and Technology*, *7*(3),

341–348.

Nurdin, & Astika, D. (2015). *Penerapan Data Mining Untuk Menganalisis Penjualan Barang Dengan Pada Supermarket Sejahtera Lhokseumawe*. *6*(1), 134–155. https://doi.org/10.29103/TECHSI.V7I1.184

Rahayu, S., Purnama, J. J., Nawawi, H. M., & Nugraha, F. S. (2019). *Algoritma Naïve Bayes Classifier Untuk Memprediksi Gejala Autism Spectrum Disorders Pada Anak-Anak*. *November*.

Rahman, A. F. A., Sorikhi, & Wartulas, S. (2020). *Prediksi kelulusan mahasiswa menggunakan algoritma c4.5 (studi kasus di universitas peradaban)*. *1*(2), 70–77.

Rohman, A. (2015). Model Algoritma K-Nearest Neighbor (K-Nn) Untuk Prediksi Kelulusan Mahasiswa. *Neo Teknika*, *1*(1). https://doi.org/10.37760/neoteknika.v1i1.350

Romadhona, A., Suprapedi , S. dan Himawan, H. (2017). Prediksi Kelulusan Mahasiswa Tepat Waktu Berdasarkan Usia, Jenis Kelamin, dan Indeks Prestasi Menggunakan Algoritma Decision Tree. *Jurnal Teknologi Informasi*, *13*, 69–83.

Salmu, S., D., & Solichin, A. (2017). Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naïve Bayes : Studi Kasus UIN Syarif Hidayatullah Jakarta Prediction of Timeliness Graduation of Students Using Naïve Bayes : A Case Study at Islamic State University Syarif Hidayatullah Jakarta. *Prosiding Seminar Nasional Multidisiplin Ilmu*, *April*, 701–709.

Setio, P. B. N., Saputro, D. R. S., & Bowo Winarno. (2020). Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5. *PRISMA, Prosiding Seminar Nasional Matematika*, *3*, 64–71.

Tampil, Y. A., Komalig, H., & Langi, Y. (2015). *Analisis Regresi Logistik Untuk Menentukan Faktor-Faktor Yang Mempengaruhi Indeks Prestasi Kumulatif (IPK) Mahasiswa FMIPA Universitas Sam Ratulangi Manado Logistic Regression Analysis To Determine Factors Affecting The Grade Point Average (GPA) Of FM*.

Tanjung, Y. P., Sentinuwo, S. R., & Jacobus, A. (2016). Penentuan Daya Listrik Rumah Tangga Menggunakan Metode Decision Tree. *Jurnal Teknik Informatika*, *9*(1). https://doi.org/10.35793/jti.9.1.2016.14141

Ulya, S. F., Sukestiyono, Y., & Hendikawati, P. (2018). *RANDOM SAMPLING CONFIDENCE INTERVAL*. *7*(1), 108–119.

Wirawan, C. (2020). *Teknik Data Mining Menggunakan Algoritma Decision Tree C4 . 5 untuk Memprediksi Tingkat Kelulusan Tepat Waktu*. *3*(1), 47–52.

Wirawan, W., Aghastya, A., & Lailya, A. L. (2019). *No Title*. *III*, 55–61.

Yahya, N., & Jananto, A. (2019). Komparasi Kinerja Algoritma C.45 dan Naive Bayes Untuk Prediksi Kegiatan Penerimaanmahasiswa Baru (Studi Kasus : Universitas Stikubank Semarang). *Prosiding SENDI*, *2014*, 978–979. https://www.unisbank.ac.id/ojs/index.php/sendi_u/article/view/7389/2369