PREDIKSI KELULUSAN MAHASISWA FASILKOM UNSIKA MENGGUNAKAN ALGORITMA C4.5

Zahra Ammellia Putri, Betha Nurina Sari, Iqbal Maulana
Informatika, Universitas Singaperbangsa Karawang
Jl. HS.Ronggo Waluyo, Puseurjaya, Telukjambe Timur, Karawang, Indonesia
zahraammellia2@gmail.com

ABSTRAK

Perguruan tinggi adalah lembaga yang menyelenggarakan pendidikan akademik untuk mahasiswa dengan harapan dapat memberikan pendidikan berkualitas. Salah satu keberhasilan pihak perguruan tinggi dilihat dari banyaknya lulusan tepat waktu yang dihasilkan. Oleh karena itu, banyak lembaga pendidikan yang memberikan perhatian khusus terhadap kelulusan tepat waktu mahasiswa. Penelitian ini bertujuan untuk menerapkan algoritma C4.5 pada *decision tree* untuk memprediksi kelulusan mahasiswa Fasilkom Unsika. Data yang digunakan didapatkan melalui arsip tata usaha Fasilkom Unsika dan kuisioner *google form* yang meliputi data lulus tepat waktu, gender, asal sekolah, jalur masuk, pernah mengikuti organisasi, kuliah sambil bekerja dan pernah mengikuti sertifikasi. Penelitian ini mengimplementasikan proses *data mining* menggunakan tahapan tahapan dari KDD dengan menggunakan variasi 3 skenario perbandingan rasio *data training* dan *data testing* 90:10, 80:20, dan 70:30. Penelitian ini juga akan menggunakan *postprune* dengan *cost complexity* untuk memastikan bahwa hasil analisis yang diperoleh adalah yang paling optimal. Hasil *pruning decision tree* dari penelitian ini mendapatkan akurasi yang paling tinggi 74% dimana merupakan perbandingan *data training* dan *data testing* 80:20 dengan *precision* sebesar 80% dan *recall* sebesar 80%. Model klasifikasi tersebut bahwa atribut "mengikuti sertifikasi" menjadi *root node* karena mempunyai nilai gain tertinggi dibanding atribut lainnya yaitu sebesar 0.1689.

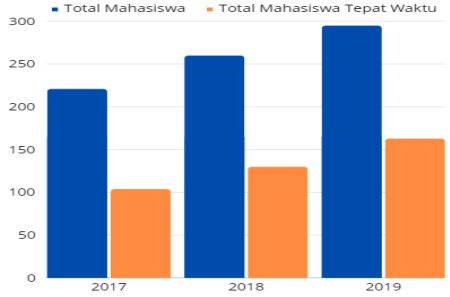
Kata kunci : *Decision Tree, Algoritma C4.5, Post Pruning, Cost Complexity*

1. PENDAHULUAN

Perguruan tinggi adalah lembaga yang menyelenggarakan pendidikan akademik untuk mahasiswa dengan harapan dapat memberikan pendidikan berkualitas guna menghasilkan individu yang berilmu dan mampu berkontribusi pada pembangunan nasional. Keberhasilan pendidikan tergantung pada pencapaian tujuan pendidikan nasional dan efektivitas serta efisiensi proses belajar mengajar. Evaluasi kesuksesan perguruan tinggi seringkali didasarkan pada prestasi belajar mahasiswa, sementara kegagalan atau rendahnya kualitas mahasiswa mencerminkan kurangnya kemampuan perguruan tinggi dalam menyelenggarakan pendidikan tinggi[1].

Salah satu keberhasilan pihak perguruan tinggi dapat dilihat dari banyaknya lulusan yang dihasilkan[2]. Lulus kuliah tepat waktu merupakan tujuan utama bagi setiap mahasiswa. Semua mahasiswa pastinya berharap dapat lulus secara tepat waktu dan tidak menambah semester. Lulus tepat waktu berarti mahasiswa telah menyelesaikan seluruh tahapan studi, termasuk seminar proposal, kolokium, dan yudisium dalam kurun waktu yang telah ditentukan. Lulus tepat waktu sangat penting bagi mahasiswa karena dapat mempermudah proses mencari pekerjaan dan menghemat biaya kuliah. Selain itu, kelulusan tepat waktu juga merupakan salah satu kriteria penting dalam menilai kualitas pendidikan di suatu lembaga. Oleh karena itu, banyak lembaga pendidikan yang memberikan perhatian khusus terhadap prediksi kelulusan tepat waktu mahasiswa.

Universitas Singaperbangsa Karawang adalah universitas negeri yang berada di Kabupaten Karawang, Provinsi Jawa Barat. Universitas Singaperbangsa Karawang memiliki 9 Fakultas yang terdiri dari 26 program studi di dalamnya. Salah satunya adalah Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang (Fasilkom Unsika). Fakultas ini memiliki dua program studi yaitu Informatika dan Sistem Informasi. Dikatakan pada salah satu misi dari Fasilkom Unsika, yaitu meningkatkan aksesibilitas dan kualitas pelayanan pendidikan tinggi di bidang ilmu komputer bagi masyarakat. Lulusan Fasilkom Unsika diharapkan mencapai hasil prestasi yang lebih baik untuk menghadapi prestasi yang lebih ketat kedepannya. Dikarenakan pencapaian kelulusan tepat waktu mahasiswa Fasilkom Unsika belum stabil dan relatif rendah. Maka hal ini menjadi tembok besar tersendiri bagi fakultas ilmu komputer untuk lebih semangat dan konsisten dalam meningkatkan lagi prestasi mahasiswanya. Berikut data kelulusan mahasiswa Fasilkom Unsika tahun 2017-2019.



Gambar 1. Data kelulusan mahasiswa fasilkom

Berdasarkan Gambar 1 di atas mengenai total mahasiswa yang lulus dan mahasiswa terdaftar pertahunnya. Data tersebut diperoleh dari Tata Usaha Fasilkom tahun 2023. Data yang lulus tepat waktu angkatan 2017 adalah 104 dari 221 mahasiswa yang terdaftar (47%). Angkatan tahun 2018 sebanyak 130 mahasiswa lulus tepat waktu dari 260 mahasiswa yang terdaftar (50%). Sedangkan angkatan 2019, sebanyak 163 mahasiswa yang lulus tepat waktu dari 295 yang terdaftar (55%). Berdasarkan dari data tersebut, dapat disimpulkan bahwa jumlah mahasiswa yang lulus tepat waktu mengalami kenaikan walaupun masih tergolong relatif rendah, menunjukkan bahwa lembaga pendidikan belum berhasil dalam meningkatkan kualitas pendidikan yang diberikan.

Oleh karena itu, prediksi kelulusan tepat waktu menjadi perhatian utama bagi lembaga pendidikan, termasuk Fakultas Ilmu Komputer (Fasilkom) Universitas Singaperbangsa Karawang (Unsika). Dengan adanya prediksi ini, diharapkan lembaga pendidikan dapat mengidentifikasi faktor-faktor yang mempengaruhi kelulusan tepat waktu dan mengambil langkah-langkah yang diperlukan untuk meningkatkan tingkat kelulusan tepat waktu mahasiswanya. Faktor faktor yang dibahas adalah gender, asal sekolah, jalur masuk, mengikuti organisasi, kuliah sambil bekerja dan mengikuti sertifikasi.

Berdasarkan penelitian [3] mengenai Data Mining untuk Memprediksi Kelulusan Mahasiswa Jurusan Teknik Informatika UIN Syarif Hidayatullah Jakarta menggunakan Metode Klasifikasi C4.5 tahun 2022, gender merupakan salah satu faktor yang mempengaruhi perilaku dan prestasi mahasiswa dalam proses pendidikan. Asal sekolah mahasiswa juga mempengaruhi kelulusan mereka. Sekolah yang memiliki kualitas tinggi dan curriculum yang baik dapat mempengaruhi prestasi mahasiswa dalam proses pendidikan. Jalur masuk yang berbeda mungkin memiliki perbedaan dalam tingkat kesediaan mahasiswa untuk menghadapi tantangan di universitas, sehingga mempengaruhi kelulusan mereka dalam proses pendidikan. Keaktifan berorganisasi juga dapat mempengaruhi kinerja akademik mahasiswa[4]. Dan kuliah sambil bekerja dapat mempengaruhi kinerja akademik mahasiswa.

Prediksi kelulusan mahasiswa lulus tepat waktu dapat dilakukan menggunakan algoritma prediksi *data mining*. Dimana pada prediksi ini menggunakan algoritma C4.5 pada *decision tree* sehingga dari *role* pada *decision tree* menghasilkan keputusan berupa klasifikasi lulus tepat waktu waktu atau tidaknya mahasiswa. Tujuan dari penelitian ini adalah menerapkan algoritma C4.5 untuk memprediksi kelulusan mahasiswa Fasilkom Unsika berdasarkan faktor-faktor yang telah dijabarkan sebelumnya. Kemudian untuk mengetahui performa dan kinerja dari model yang sudah dibangun algoritma C4.5 dalam memprediksi kelulusan mahasiswa Fasilkom Unsika.

2. TINJAUAN PUSTAKA

2.1. *Data Mining*

Data mining adalah sebuah proses panjang untuk mencari informasi yang nantinya akan menjadi ilmu pengetahuan baru yang didapatkan dari sekian banyak fakta fakta tak terarah[5]. Dalam melakukan data mining statitiska, matematika, *artifical intellegence* dan *machine learning* tidak dapat dipungkiri, sangat terlibat. Metode yang dapat diterapkan dalam *data mining* diantaranya yaitu: *classification*, *clustering*, *association*, *regression*, dan *forecasting*.

2.2. *Decision Tree*

Pohon keputusan adalah algoritma pembelajaran terawasi non-parametrik, yang digunakan untuk tugas klasifikasi dan regresi. Ini memiliki hierarki, struktur pohon, yang terdiri dari simpul akar, cabang, simpul internal dan simpul daun. Pohon keputusan dimulai dengan simpul akar, yang tidak memiliki cabang masuk. Cabang keluar dari simpul akar kemudian masuk ke simpul internal, juga dikenal sebagai simpul keputusan. Berdasarkan fitur yang tersedia, kedua tipe simpul melakukan evaluasi untuk membentuk himpunan bagian yang homogen, yang dilambangkan dengan simpul daun, atau simpul terminal[3].

2.3. Klasifikasi

Klasifikasi merupakan satu dari beberapa pengelompokkan teknik dari data mining yang yang biasa digunakan untuk memprediksi atau meramalkan sesuatu yang mungkin terjadi di masa depan maka dari itu klasifikasi termasuk ke dalam kategori *predictive mining*[6]. Pada pelaksanaannya teknik ini mengklasifikasikan data ke dalam kelas target berupa binary option seperti ya dan tidak atau 1 dan 0 yang mana hal tersebut diketahui dari pembuatan sebuah model atau pola hasil dari training data terhadap atribut-atribut yang paling berpengaruh dan kemudian diuji validitasnya menggunakan teknik uji tertentu.

2.4. Algoritma C4.5

Algoritma C4.5 biasanya digunakan dalam penambangan data sebagai pengklasifikasi pohon keputusan yang digunakan untuk mendapatkan hasil berupa keputusan, yang didasari suatu sampel data tertentu (prediktor univariat atau multivariat). Pohon keputusan yang dibangun Algoritma C4.5 dari sekumpulan data pelatihan dengan cara yang sama seperti ID3, yang menggunakan konsep entropi informasi. Dalam penerapannya, algoritma ini menggunakan rasio perolehan atau *gain ratio*.

2.5. *Knowledge Discovery in Database*

Knowledge Discovery in Database(KDD) adalah proses mengekstraksi informasi dari sejumlah besar data. Keseluruhan proses KDD[7] dapat dijelaskan sebagai berikut. Tahap pertama adalah *Data Selection* dimana informasi lebih lanjut tentang domain masalah dikumpulkan. Pada tahap *Data Preprocessing* menghasilkan database operasional yang telah dibersihkan. Tahap berikutnya yaitu *Data*

Transformation dimana data dilakukan pemilihan metode seperti klasifikasi, clustering, regresi, asosiasi, dll. Tahap keempat yaitu *Data Mining*, pola diekstraksi dengan menggunakan tugas dan algoritma penambangan data yang dipilih. Selain itu, model dievaluasi pada langkah ini. Selanjutnya *Pattern Evaluation* didefinisikan sebagai mengidentifikasi pola yang meningkat secara ketat yang mewakili pengetahuan berdasarkan ukuran yang diberikan. Dan terakhir *Knowledge Presentation* Ini melibatkan penyajian pengetahuan yang telah diekstraksi dari data dengan cara yang dapat dimengerti dan berguna bagi pengguna akhir.

2.6. *Python*

Python adalah bahasa pemrograman yang dikenal sebagai bahasa pemrograman tingkat tinggi yang dapat menangani berbagai tugas pemrograman seperti pengolahan data, komputasi numerik, web pengembangan, pemrograman basis data, pemrograman jaringan, parallel pemrosesan, dll[8].

2.7. *Confusion Matrix*

Confusion Matrix adalah alat pengukuran yang digunakan untuk kinerja klasifikasi model, seperti yang digunakan dalam pembelajaran mesin. Confusion Matrix membantu dalam menyalakan kinerja model klasifikasi dengan menunjukkan tingkat benar, ketidakbenaran, dan kelasifikasi yang salah. Dalam mesin pembelajaran, Confusion Matrix digunakan untuk membandingkan nilai aktual dengan nilai prediksi, sehingga dapat menyebarkan kinerja model dan membuat penyesuaian jika diperlukan. Dengan menggunakan Confusion Matrix, kita dapat menampilkan kinerja klasifikasi model dan membandingkannya dengan model lainnya untuk memilih model terbaik yang sesuai dengan permintaan.

3. METODE PENELITIAN

3.1. Objek Penelitian

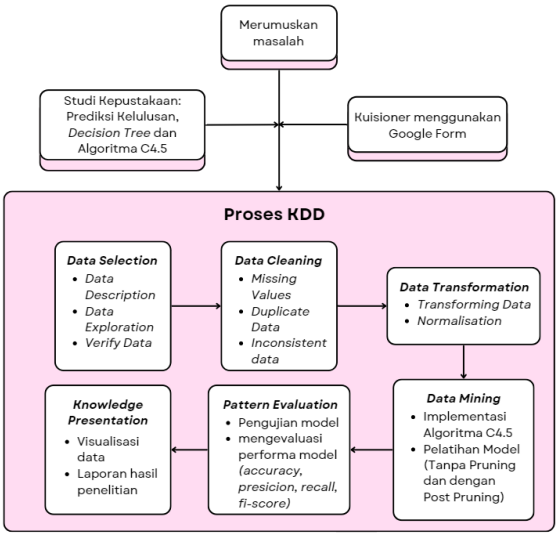
Objek pada penelitian ini adalah data mahasiswa Universitas Singaperbangsa Karawang angkatan tahun 2017 sampai dengan 2019. Data-data yang digunakan meliputi data lulus tepat waktu, gender, asal sekolah, jalur masuk, pernah mengikuti organisasi, kuliah sambil bekerja dan pernah mengikuti sertifikasi. Penelitian ini dilakukan dengan mengisi kuisioner dan data yang didapat dari tata usaha Fasilkom Unsika.

3.2. Metodologi Penelitian

Penelitian ini akan mengimplementasikan proses data mining menggunakan tahapan tahapan dari KDD dalam menentukan pola faktor yang memengaruhi kelulusan tepat waktu mahasiswa Universitas Singaperbangsa Karawang angkatan tahun 2017 sampai 2019. Adapun metode penelitian yang dipilih oleh peneliti adalah metode *decision tree* menggunakan Algoritma C4.5. Untuk mengetahui bagaimana variasi skenario dapat mempengaruhi hasil analisis, penelitian ini akan menggunakan tiga skenario yaitu; 90:10, 80:20 dan 70:30. Selain itu, penelitian ini juga akan menggunakan postprune dengan cost complexity untuk memastikan bahwa hasil analisis yang diperoleh adalah yang paling optimal dan sesuai dengan tujuan penelitian.

3.3. Rancangan Penelitian

Pada rancangan penelitian ini, peneliti akan membuat rincian tahapan yang akan dilakukan sampai dengan penelitian selesai. Dimulai dari pengumpulan data, sampai nanti diperoleh *output* yang diharapkan. Sehingga dari *output* tersebut peneliti mendapatkan sebuah keputusan yang berguna bagi perguruan tinggi dan mahasiswa khususnya Fasilkom Unsika dalam memprediksi kelulusan mahasiswa tepat waktu. Berikut gambaran dari rancangan penelitian yang akan dilakukan:



Gambar 2. Rancangan Penelitian

4. HASIL DAN PEMBAHASAN

4.1. *Data Selection*

Pada proses ini dilakukan pengumpulan data, dimana data yang dikumpulkan berasal dari arsip tata usaha Unsika dan hasil dari kuisioner *google form*. Data yang berasal dari arsip tata usaha Unsika adalah data kelulusan mahasiswa 2017-2019, dimana data tahun 2017 dan 2018 dari program studi Informatika dan di tahun 2019 data program studi Informatika dan Sistem Informasi. Dalam penelitian ini menggunakan Teknik pengambilan sampel yang akan berhubungan dengan penentuan jumlah sampel dengan menggunakan pendekatan rumus *slovin*. Menurut[9] rumus slovin dapat dirumuskan sebagai berikut:

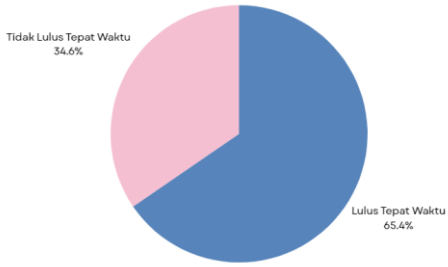
$$n = \frac{\mathrm{N}}{1 + \mathrm{N}e^2} \quad (1)$$

Dimana n adalah jumlah sampel, N adalah total populasi dan e adalah tingkat kesalahan dalam pengambilan sampel. Diketahui jumlah keseluruhan mahasiswa Fasilkom Unsika tahun 2017-2019 sebanyak 776 orang. Dalam penulisan penelitian ini, peneliti menggunakan tingkat kepercayaan sebesar 95% dengan kesalahan maksimum 5%[9]. Maka

perhitungan pengambilan sampel menggunakan rumus *slovin* adalah sebagai berikut:

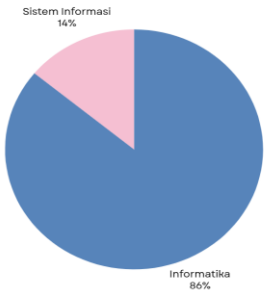
$$n = \frac{776}{1 + 776(0.05)^2}$$

Dari hasil rumus *slovin* tersebut menghasilkan angka jumlah sampel sebesar 263.945 yang dibulatkan menjadi 264 data sampel mahasiswa Fasilkom Unsika 2017-2019. Dari hasil penyebaran kuisioner dimulai dari tanggal 3 April 2024 hingga 11 Juni 2024. Dari 70 hari tersebut peneliti berhasil mengumpulkan data sebanyak 299 data dimana jumlah tersebut sudah memumpuni dari jumlah minimal sampel yang telah dihitung dengan rumus *slovin* yaitu 264 data.

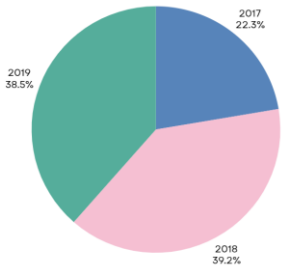


Gambar 3. *Pie chart* perbandingan kelulusan mahasiswa tepat waktu

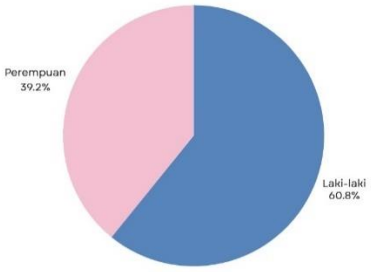
Dari 299 data kuisioner tersebut, terdapat 195 mahasiswa yang lulus tepat waktu dan 103 mahasiswa yang tidak lulus tepat waktu yang digambarkan pada gambar 4.1 *pie chart* persentase perbandingan kelulusan mahasiswa tepat waktu dan tidak.



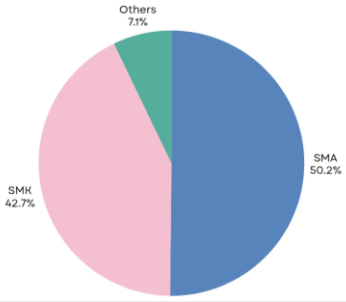
Gambar 4. Perbandingan isi dataset berdasarkan program studi



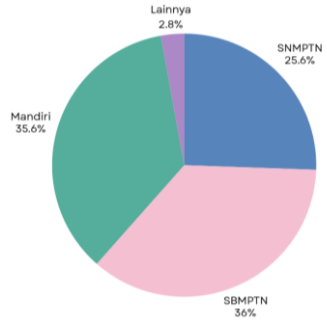
Gambar 5. Perbandingan isi dataset berdasarkan tahun masuk



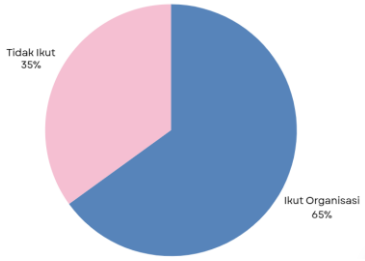
Gambar 6. Perbandingan isi dataset berdasarkan gender



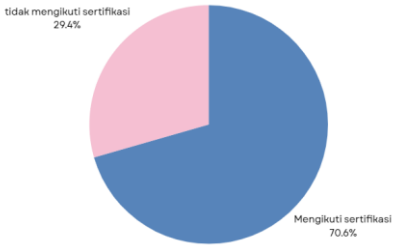
Gambar 7. Perbandingan isi dataset berdasarkan asal sekolah



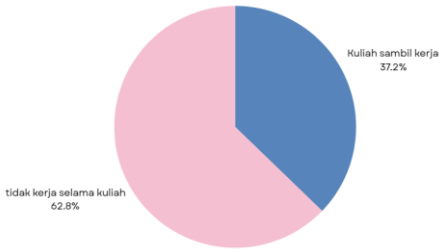
Gambar 8. Perbandingan isi dataset berdasarkan jalur masuk



Gambar 9. Perbandingan isi dataset berdasarkan keikutsertaan organisasi



Gambar 10. Perbandingan isi dataset berdasarkan keikutsertaan sertifikasi



Gambar 11. Perbandingan isi dataset berdasarkan kuliah sambil kerja

Berdasarkan Gambar 3 sampai Gambar 11, dari 299 data yang dikumpulkan, dapat diketahui bahwa 257 orang mengisi program studi informatika, sedangkan 42 orang mengisi program sistem informasi. Data tahunan menunjukkan bahwa pada tahun 2017, 69 orang mengisi kuisioner, pada tahun 2018, 121 orang, dan pada tahun 2019, 119 orang. Dalam hal gender, 188 orang yang mengisi kuisioner adalah laki-laki, sedangkan 121 orang adalah perempuan. Dalam hal asal sekolah, mayoritas mahasiswa berasal dari SMA, yaitu 155 orang, diikuti oleh SMK sebanyak 132 orang, dan sisanya berasal dari sekolah lain sebanyak 22 orang. Jalur masuk Fasilkom Unsika menunjukkan bahwa 81 orang masuk melalui SNMPTN, 114 orang melalui SBMPTN, 113 orang melalui jalur mandiri, dan 9 orang melalui jalur prestasi.Dalam hal keaktifan organisasi, 201 orang mengikuti organisasi, sedangkan 108 orang tidak. Dalam hal pekerjaan sambil kuliah, 115 orang bekerja, sedangkan 194 orang tidak. Dalam hal sertifikasi, 218 orang mengikuti sertifikasi, sedangkan 91 orang tidak mengikuti sertifikasi selama kuliah.

Data yang terdiri dari 299 dataset dengan 11 atribut awal termasuk profil mahasiswa dan kelas target. 11 atribut itu terdiri dari NIM, Nama, Program studi, Angkatan tahun, gender, Asal sekolah, Jalur masuk universitas, keikutsertaan organisasi, kuliah sambal bekerja, keikutsertaan sertifikasi dan waktu kelulusan. Beberapa dataset akan diseleksi untuk kebutuhan penelitian dikarenakan hanya memerlukan parameter yang penting saja menurut[10]. Tabel 1 merupakan dataset awal yang belum ada diseleksi.

Tabel 1. Dataset sebelum diseleksi

Attributes	Types
NIM	int
Nama	object
Program studi	object
Angkatan tahun	int
Gender	object
Asal sekolah	object
Jalur masuk universitas	object
Mengikuti organisasi saat kuliah	object
Kuliah sambil kerja	object
Mengikuti sertifikasi object	object
Waktu kelulusan	object

Data yang diseleksi adalah NIM, Nama, Program studi dan Angkatan tahun karena tidak dibutuhkan dalam proses data mining nantinya. Setelah diseleksi, atribut dataset menjadi berjumlah 7 atribut seperti yang ditampilkan pada Gambar 12 di bawah ini.

```
# Selection Data
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df = pd.read_excel("data_terbaru.xlsx")
df = df.drop(columns=['NPM', 'Nama Lengkap', 'Program Studi', 'Angkatan Tahun'])
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 301 entries, 0 to 300
Data columns (total 8 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Timestamp                     301 non-null    datetime64[ns]
 1   Email Address                 66 non-null     object
 2   Gender                        301 non-null    object
 3   Asal Sekolah                  301 non-null    object
 4   Jalur masuk Universitas       301 non-null    object
 5   Mengikuti organisasi saat kuliah  301 non-null  object
 6   Kuliah sambil bekerja         301 non-null    object
 7   Mengikuti Sertifikasi         301 non-null    object
dtypes: datetime64[ns](1), object(7)
```

Gambar 12. Dataset sesudah diseleksi

4.2. *Data Cleaning*

Setelah melakukan seleksi data, kemudian data akan diproses ke tahap selanjutnya yaitu *data cleaning*. Pada tahap ini, peneliti akan mencari data yang duplikat, *missing values* atau data yang kosong (nilai *null*) dan data yang tidak lengkap menggunakan Phyton. Dalam Phyton data cleaning memanfaatkan library, yaitu Pandas dan NumPy untuk mempermudah sekaligus mempercepat pembersihan data. Pada tahap pertama Namun, dikarenakan data tersebut berasal dari kuisioner yang mana bersifat pernyataan tertutup, maka pada tahap ini tidak berikan perlakuan apapun terhadap data karena tidak terdapat *missing values* atau data yang kosong. Hal tersebut dapat dilihat pada Gambar 13.

```
# Data Cleaning
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df_raw = pd.read_excel("data_terbaru_lulus.xlsx")
df_duplikat = df_raw.duplicated(subset=['NPM'])
count_duplikat = df_duplikat.sum()
print('Duplikat NPM', df_duplikat)
print('jumlah nilai yang duplikat', count_duplikat)
spec_duplikat = df_raw[df_duplikat]
print(spec_duplikat)
df_cleaned = df_raw.drop_duplicates(subset=['NPM'])
df = df_cleaned.drop(columns=['NPM', 'Nama Lengkap', 'Program Studi', 'Angkatan Tahun'])
df.info()
```

Gambar 13. Kode untuk cek missing values dan cek duplikat

```
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Gender                        270 non-null    object
 1   Asal Sekolah                  270 non-null    object
 2   Jalur Masuk Universitas       269 non-null    object
 3   Mengikuti Organisasi Kuliah   269 non-null    object
 4   Kuliah sambil bekerja         269 non-null    object
 5   Mengikuti Sertifikasi         269 non-null    object
 6   Kategori Lulus Tepat Waktu    269 non-null    object
dtypes: object(7)
memory usage: 17.3+ KB
```

Gambar 14. Hasil dari cek missing values

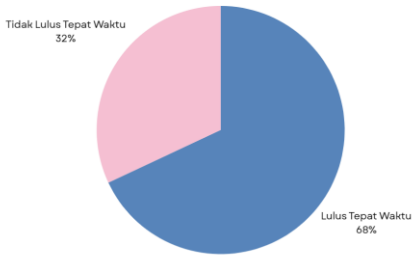
Seperti yang dijelaskan pada Gambar 14 tidak terdapat nilai *null* dalam dataset ini. Maka dilanjut ke tahap kedua, yaitu peneliti akan mencari data duplikat dari 299 data tersebut. Data yang duplikat dapat mengganggu analisis dan menyebabkan interpretasi. Menghapus data yang duplikat membantu

menghindari redudansi. Berikut Gambar 15 merupakan kode dan hasil program membersihkan duplikat. Atribut yang menjadi patokan untuk mencari data yang duplikat adalah menggunakan atribut 'NPM' karena NPM merupakan ID bagi mahasiswa dan tidak ada yang memiliki NPM yang sama. Data yang awalnya ada 299 data menjadi 269 data karena ada 30 data yang duplikat.

```
Duplikat NPM 0      False
1      False
2      False
3      False
4      False
       ...
288     True
289     True
290    False
291     True
292     True
Length: 293, dtype: bool
jumlah nilai yang duplikat 30
```

Gambar 15. Kode dan hasil cek data duplikat

Setelah melalui *data cleaning*, dimana ada 26 data yang terdapat duplikasi, maka dapat dilihat pada Gambar 16 diagram perbandingan kelulusan mahasiswa tepat waktu. Mahasiswa yang lulus tepat waktu 183 dan mahasiswa yang tidak lulus tepat waktu terdapat 86 mahasiswa. Dataset inilah yang kedepannya akan menjadi dataset yang akan diolah untuk tahap tahap selanjutnya sebanyak 274 data.



Gambar 16. Perbandingan kelulusan mahasiswa

4.3. *Data Transformation*

Data yang sudah dibersihkan selanjutnya disamakan bentuknya yang berguna untuk proses *data mining*. Berdasarkan dataset, atribut yang dilakukan untuk *transformation data* yaitu "kategori_lulus_tepat_waktu", atribut tersebut akan diubah menjadi "Ya" dan "Tidak" dengan ketentuan berdasarkan angkatan tahun. Angkatan 2017 mempunyai ketentuan "Kategori Lulus Tepat Waktu" jika tanggal lulus tidak melebihi dari Agustus 2021. Sedangkan untuk angkatan 2018 mempunyai ketentuan "Kategori Lulus Tepat Waktu" jika tanggal lulus lebih dari Agustus 2022 dan angkatan 2019 mempunyai ketentuan "Kategori Lulus Tepat Waktu" jika tanggal lulus lebih dari Agustus 2023.

Selanjutnya adalah proses *label encoding* yang berperan untuk mengubah variable kategori menjadi variable numerik. Label encoding diimplementasikan dengan mengimpor kelas "LabelEncoder" dari modul "sklearn.preprocessing" yang dapat dilihat pada Gambar 17.

```python
# Data Transformation
from sklearn.preprocessing import LabelEncoder

label_encoders = {}

for column in df.columns:
    if df[column].dtype == 'object':
        label_encoders[column] = LabelEncoder()
        df[column] = label_encoders[column].fit_transform(df[column])
        print(f"Label encoding for column {column}:")
        for index, class_label in enumerate(label_encoders[column].classes_):
            print(f"{class_label}: {index}")
        print()

print(df.head())
```

Gambar 17. Kode untuk label encoding

```
   Gender  Asal Sekolah  Jalur Masuk Universitas  Mengikuti Organisasi Kuliah  \
0       0             0                        2                            1
1       0             1                        3                            1
2       0             1                        3                            1
3       0             0                        3                            1
4       1             0                        3                            1

   Kuliah sambil bekerja  Mengikuti Sertifikasi  Kategori Lulus Tepat Waktu
0                      1                      1                           0
1                      0                      1                           1
2                      1                      1                           1
3                      1                      1                           0
4                      1                      1                           1
```

Gambar 18. Hasil dari label encoding

Dapat dilihat pada Gambar 18 atribut yang dilakukan untuk *transformation data* yaitu "kategori_lulus_tepat_waktu", atribut tersebut akan diubah menjadi "Ya" dan "Tidak" dengan ketentuan berdasarkan angkatan tahun. Angkatan 2017 mempunyai ketentuan "Kategori Lulus Tepat Waktu" jika tanggal lulus tidak melebihi dari Agustus 2021. Sedangkan untuk angkatan 2018 mempunyai ketentuan "Kategori Lulus Tepat Waktu" jika tanggal lulus lebih dari Agustus 2022 dan angkatan 2019 mempunyai ketentuan "Kategori Lulus Tepat Waktu" jika tanggal lulus lebih dari Agustus 2023. Berikut hasil *transformation data* pada atribut "Kategori Lulus Tepat Waktu". Atribut "Gender" ditransformasi dari yang laki laki dirubah menjadi 0 dan perempuan menjadi 1. Atribut "Asal Sekolah" dari yang SMA diubah menjadi 0 dan SMK diubah menjadi 1. Atribut "Jalur Masuk" dari yang Mandiri diubah menjadi 0, Prestasi diubah menjadi 1, SBMPTN dirubah menjadi 2 dan SNMPTN diubah menjadi 3. Atribut "Mengikuti Organisasi" dari yang Tidak diubah menjadi 0 dan Ya diubah menjadi 1. Atribut "Kuliah Sambil Bekerja" dari yang Tidak diubah menjadi 0 dan Ya diubah menjadi 1. Atribut "Mengikuti Sertifikasi" dari yang Tidak diubah menjadi 0 dan Ya diubah menjadi 1. Dan yang terakhir, atribut lulus tepat waktu dari yang Tidak diubah menjadi 0 dan Ya diubah menjadi 1.

4.4. *Data Mining*

Dalam proses *data mining*, permodelan dataset dibagi dan diuji menggunakan teknik *train test split*, yang mana data akan dibagi menjadi 3 bagian, pertama data *training* sebesar 70% dan test sebesar 30%, kedua data *training* sebesar 80% dan test sebesar 20%, data *training* sebesar 90% dan test sebesar 10%. Selanjutnya, dilakukan proses permodelan menggunakan algoritma *Decision Tree* (C4.5). Algoritma ini dipilih karena dapat menangani klasifikasi pada dataset lulus tepat waktu. Pada

langkah ini, model diinisialisasi, dilatih menggunakan data *training*, dan diprediksi menggunakan data *testing* untuk menilai kinerja dan generalisasi model.

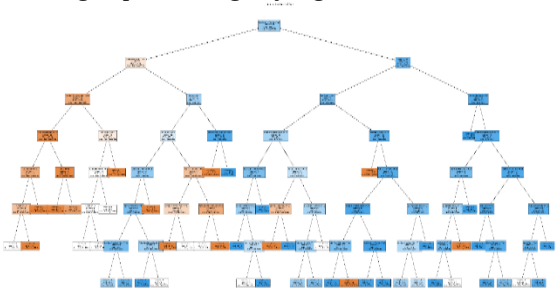
Penerapan langkah-langkah permodelan dilakukan menggunakan 269 data setelah melalui tahap-tahap sebelumnya. Permodelan dilanjutkan dengan menggunakan algoritma *Decision Tree* (C4.5) dengan pembagian rasio 70:30, yang mana 70% (188 data) merupakan data training dan 30% (81 data) merupakan data *testing*, untuk pembagian rasio 80:20, yang mana 80% (215 data) merupakan data training dan 20% (54 data) merupakan data *testing*, untuk pembagian rasio 90:10, yang mana 90% (242 data) merupakan data training dan 10% (27 data) merupakan data *testing*. Perbandingan tersebut dipilih untuk membandingkan hasil akurasi yang tertinggi. karena menghasilkan nilai akurasi yang lebih tinggi daripada perbandingan yang lainnya di penelitian ini. Setelah semua data sudah dibagi, selanjutnya dilakukan proses klasifikasi dengan membentuk sebuah *decision tree* sebagai *output*nya.

Decision tree (C4.5) akan dibangun dengan terlebih dahulu menghitung nilai *entropy* dan *gain*. Berdasarkan hasil perhitungan *entropy* dan *gain*, nilai *gain* tertinggi terdapat pada atribut Mengikuti Sertifikasi. Hasil dari perhitungan nilai *entrpy* dan *gain* terhadap data latih lebih jelasnya dapat dilihat pada tabel 2 berikut.

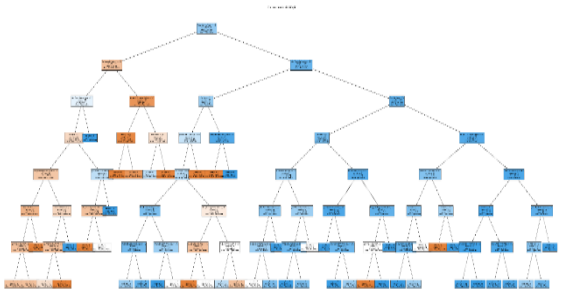
Tabel 2. Hasil perhitungan *entropy* dan *gain*

	Ket	**Jumlah Kasus**	**Ya**	**Tidak**	***Entropy***	***Gain***
Total		269	183	86	0.9041	
Gender						0.0273
	Laki-laki	167	102	65	0.9284	
	Perempuan	102	81	21	0.7619	
Sekolah						0.0698
	SMA	147	81	66	0.9984	
	SMK	122	102	20	0.5272	
Jalur masuk univ						0.1565
	SBM	105	86	19	0.5392	0.5392
	SNM	66	58	8	0.4972	0.4972
	Mandiri	97	39	58	0.9724	0.9724
	Prestasi	1	0	1	0	0
Mengikuti organisasi saat kuliah						0.0630
	Ya	186	146	40	0.6875	
	Tidak	83	37	46	0.5175	
Kerja sambil kuliah						0.0382
	Ya	97	52	45	0.9996	
	Tidak	172	131	41	0.6059	
Mengikuti sertifikasi						0.1689
	Ya	194	160	34	0.9996	
	Tidak	75	23	52	0.8412	

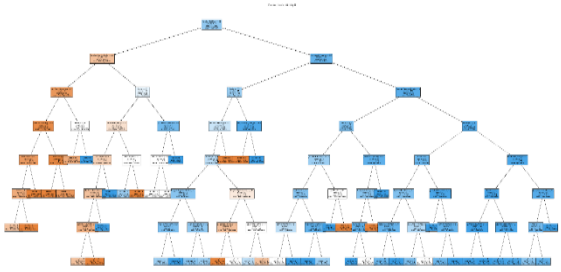
Setelah mengetahui hasil *gain* dari setiap atribut, maka atribut dengan *gain* tertinggi akan dijadikan *root node*. *Root node* pada penelitian ini adalah atribut "Mengikuti Sertifikasi". Terdapat 3 hasil dari *decision tree* dengan perbandingan yang berbeda.



Gambar 19. Hasil *decision tree* perbandingan 70:30

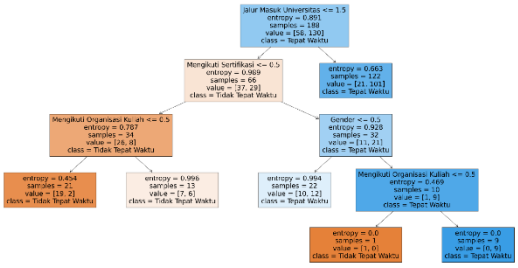


Gambar 20. Hasil *decision tree* perbandingan 80:20

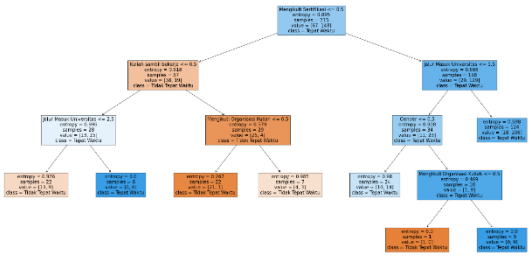


Gambar 21. Hasil *decision tree* perbandingan 90:10

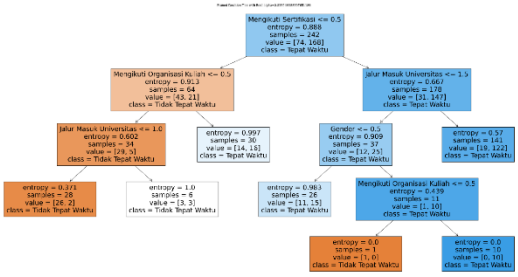
Setelah mendapatkan hasil dari *decision tree* setiap perbandingan, selanjutnya adalah dengan melakukan *pruning*. *Pruning* adalah langkah penting dalam pengembangan model decision tree yang bertujuan untuk mengurangi kompleksitas model dan meningkatkan akurasi prediksi. Berikut hasil *pruning* dari setiap perbandingan.



Gambar 22. Hasil *pruning* untuk perbandingan 70:30

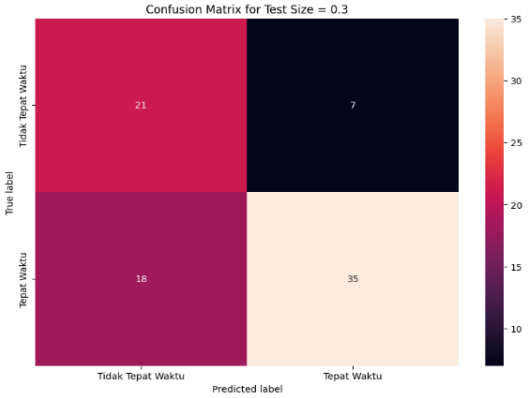


Gambar 23. Hasil *pruning* untuk perbandingan 80:20



Gambar 24. Hasil *pruning* untuk perbandingan 90:10

4.5. *Pattern Evaluation*



Gambar 25. *Confusion matrix* pada *split* data 70:30

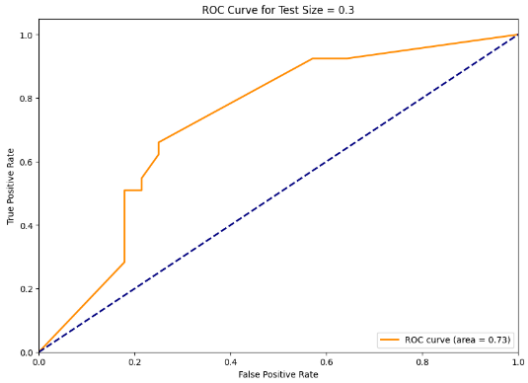
Setelah proses permodelan dengn *data mining*, selanjutnya akan melakukan evaluasi dengan menggunakan *confusion matrix* dan *classification report*. Terdapat 3 *confusion matrix* untuk setiap *split data*. Pertama, *confusion matrix* pada *split data* 70:30 menunjukan, 21 *True Negatives* (TN), 35 *True Positives* (TP), 7 *False Positives* (FP), 18 *False Negatives* (FN), dengan total semua yaitu 81.

Dengan adanya *confusion matrix* evaluasi lain juga dapat diketahui dengan *classification report* yang mana ini akan menunjukan nilai *accurasi*, *precision*, *recall*, *f1-score*, dan *support*. Untuk melihat lebih detailnya ada pada tabel 3.

Tabel 3. *Classification report*

	Precision	**Recall**	**F1-score**	**Support**
0	0.54	0.75	0.63	28
1	0.83	0.66	0.74	53
Accuracy			0.69	81
Macro avg	0.69	0.71	0.68	81
Weighted avg	0.73	0.69	0.70	81

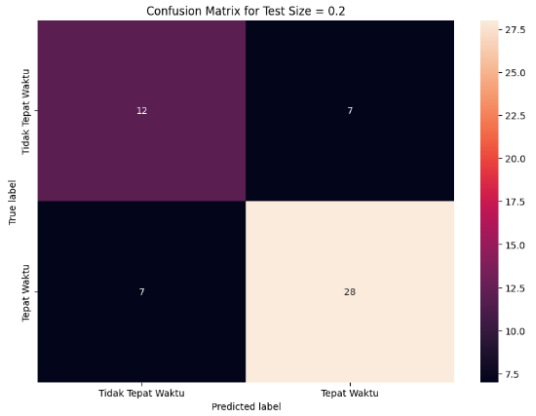
Dengan menganalisis *classification report*, model ini memiliki akurasi sekitar 69%. Setelah itu ada juga evaluasi menggunakan ROC, metrik ini akan membandingkan nilai *False Positive Rate* (FPR) dan *True Positive Rate* (TPR) yang juga didapatkan dari *confusion matrix* sebelumnya. Dengan mengetahui nilai dari kedua parameter tersebut, perbandingan kedua parameter tersebut dapat dijadikan grafik. Detail grafik tersebut dapat dilihat pada Gambar 26.



Gambar 26. ROC metrik test 0.3

Berdasarkan gambar 26 di atas, dihasilkan bahwa nilai yang didapatkan dijadikan sebagai bahan evaluasi yaitu nilai AUC (*Area Under The Curve*), nilai ini didapat dengan menghitung area atau luas di bawah kurva. AUC diperlukan karena kurva yang mendekati nilai 1 menendakan model mempunyai akurasi yang baik atau artinya nilai AUC yang semakin besar maka akan semakin bagus. Dari grafik di atas diketahui nilai AUC sebesar 0.73.

Kedua, *confusion matrix* pada *split data* 80:20 menunjukan, 12 *True Negatives* (TN), 28 *True Positives* (TP), 7 *False Positives* (FP), 7 *False Negatives* (FN), dengan total semua yaitu 54.



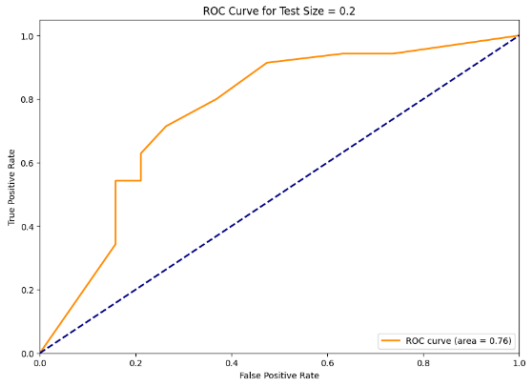
Gambar 27. *Confusion matrix* pada *split* data 80:20

Dengan adanya *confusion matrix* evaluasi lain juga dapat diketahui dengan *classification report* yang mana ini akan menunjukan nilai *accurasi*, *precision*, *recall*, *f1-score*, dan *support*. Untuk melihat lebih detailnya ada pada tabel 4.

Tabel 4. *Classification report*

	Precision	**Recall**	**F1-score**	**Support**
0	0.63	0.63	0.63	19
1	0.80	0.80	0.80	35
Accuracy			0.74	54
Macro avg	0.72	0.72	0.72	54
Weighted avg	0.74	0.74	0.74	54

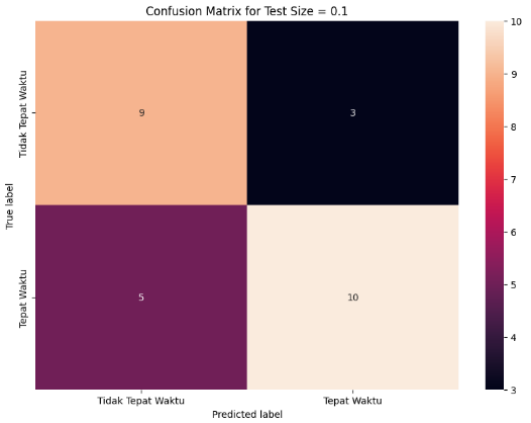
Dengan menganalisis *classification report*, model ini memiliki akurasi sekitar 74%. Setelah itu ada juga evaluasi menggunakan ROC, metrik ini akan membandingkan nilai *False Positive Rate* (FPR) dan *True Positive Rate* (TPR) yang juga didapatkan dari *confusion matrix* sebelumnya. Dengan mengetahui nilai dari kedua parameter tersebut, perbandingan kedua parameter tersebut dapat dijadikan grafik. Detail grafik tersebut dapat dilihat pada gambar 28.



Gambar 28. ROC metrik test 0.2

Berdasarkan dari gambar 28 di atas, dihasilkan bahwa nilai yang didapatkan dijadikan sebagai bahan evaluasi yaitu nilai AUC (*Area Under The Curve*), nilai ini didapat dengan menghitung area atau luas di bawah kurva. AUC diperlukan karena kurva yang mendekati nilai 1 menendakan model mempunyai akurasi yang baik atau artinya nilai AUC yang semakin besar maka akan semakin bagus. Dari grafik di atas diketahui nilai AUC sebesar 0.76.

Ketiga, *confusion matrix* pada *split data* 90:10 menunjukan, 9 *True Negatives* (TN), 10 *True Positives* (TP), 3 *False Positives* (FP), 5 *False Negatives* (FN), dengan total semua yaitu 27.



Gambar 29. *Confusion matrix* pada *split* data 90:10

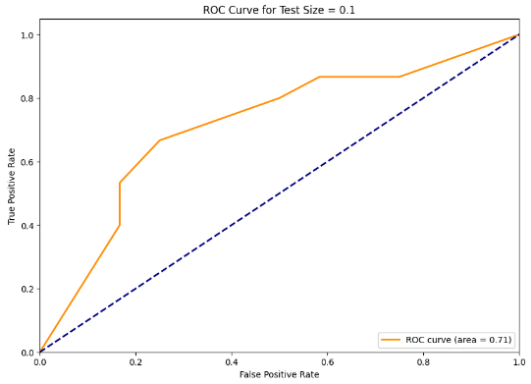
Dengan adanya *confusion matrix* evaluasi lain juga dapat diketahui dengan *classification report* yang mana ini akan menunjukan nilai *accurasi*, *precision*, *recall*, *f1-score*, dan *support*. Untuk melihat lebih detailnya ada pada tabel 5.

Tabel 5. *Classification report*

	Precision	**Recall**	**F1-score**	**Support**
0	0.64	0.75	0.69	12
1	0.77	0.67	0.71	15
Accuracy			0.70	27
Macro avg	0.71	0.71	0.70	27
Weighted avg	0.71	0.70	0.70	27

Dengan menganalisis *classification report*, model ini memiliki akurasi sekitar 70%. Setelah itu ada juga evaluasi menggunakan ROC, metrik ini akan membandingkan nilai *False Positive Rate* (FPR) dan *True Positive Rate* (TPR) yang juga didapatkan dari *confusion matrix* sebelumnya. Dengan mengetahui nilai dari kedua parameter tersebut, perbandingan kedua parameter tersebut dapat dijadikan grafik. Detail grafik tersebut dapat dilihat pada gambar 30.

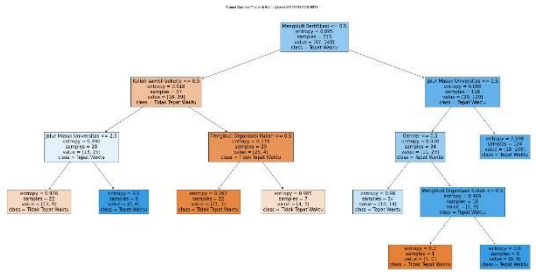
Berdasarkan dari gambar 30 di atas, dihasilkan bahwa nilai yang didapatkan dijadikan sebagai bahan evaluasi yaitu nilai AUC (*Area Under The Curve*), nilai ini didapat dengan menghitung area atau luas di bawah kurva. AUC diperlukan karena kurva yang mendekati nilai 1 menendakan model mempunyai akurasi yang baik atau artinya nilai AUC yang semakin besar maka akan semakin bagus. Dari grafik di atas diketahui nilai AUC sebesar 0.71.



Gambar 30. ROC metrik test 0.1

4.6. *Knowledge Presentation*

Knowledge Presentation yang diperoleh dari perhitungan algoritma *decision tree* (C4.5) dapat dipresentasikan sesuai dengan namanya yaitu *decision tree*. Hasil *decision tree* yang diambil ialah dengan nilai akurasi yang paling tinggi yaitu sebesar 74% yang mana merupakan perbandingan 80:20. Hasil *pruning decision tree* dengan akurasi tertinggi dapat dilihat pada gambar 31 berikut.



Gambar 31. Hasil *pruning decision tree* pada perbandingan 80:20

Berdasarkan hasil *pruning decision tree* pada perbandingan 80:20 menunjukkan bahwa terdapat 5 kedalaman *decision tree*. Hasil model klasifikasi tersebut bahwa atribut "Mengikuti Sertifikasi" menjadi root node hal ini terjadi karena mempunyai nilai gain tertinggi, atribut ini juga menjadi pengaruh tertinggi terhadap ketentuan kategori lulus tepat waktu.

5. KESIMPULAN DAN SARAN

Kesimpulan yang dapat diambil dari penelitian yang sudah dilakukan yaitu pemodelan dataset menggunakan algoritma *decision tree* (C4.5) dengan pembagian rasio 80:20 menghasilkan nilai akurasi tertinggi, yaitu 74%. Hasil pruning *decision tree* menunjukkan bahwa atribut "Mengikuti Sertifikasi" menjadi *root node* dan memiliki nilai gain tertinggi. Atribut ini juga memiliki pengaruh tertinggi terhadap kategori lulus tepat waktu.

Saran yang dapat dilakukan untuk penelitian selanjutnya yaitu meningkatkan jumlah data training dengan menggunakan data dari sumber lain, menggunakan teknik cross-validation untuk memastikan kinerja model yang lebih stabil. Dan terakhir menggunakan algoritma lain seperti Random Forest atau Support Vector Machine untuk membandingkan hasil.

DAFTAR PUSTAKA

[1] Y. T. Samuel, B. Jonathan, and J. Naibaho, "Memprediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Decision Tree J48 Di Universitas Advent Indonesia," *TeIKa*, vol. 9, no. 1, pp. 43–52, 2019.

[2] D. A. Wulandari, B. N. Sari, and T. N. Padilah, "Prediction of Student Graduation Accuracy Using C45 Algorithm (Case Study: Fasilkom Unsika): Prediksi Ketepatan Kelulusan Mahasiswa Menggunakan Algoritma C45 (Studi Kasus: Fasilkom Unsika)," *Systematics*, vol. 4, no. 1, pp. 281–372, 2022.

[3] M. Nur, "Data Mining Untuk Memprediksi Kelulusan Mahasiswa Jurusan Teknik Informatika UIN Syarif Hidayatullah Jakarta Menggunakan Metode Klasifikasi C4. 5." Fakultas Sains dan Teknologi Universitas Islam Negeri Syarif Hidayatullah ….

[4] A. A. Fauzi and T. Pahlevi, "Analisis Hubungan Keaktifan Berorganisasi Terhadap Hasil Prestasi Akademik Mahasiswa Fakultas Ekonomi Universitas Negeri Surabaya," *J. Pendidik. Adm. Perkantoran*, vol. 8, no. 3, pp. 449–457, 2020.

[5] A. Srirahayu and L. S. Pribadie, "Review Paper Data Mining Klasifikasi Data Mining," *J. Ilm. Inform. Glob.*, vol. 14, no. 1, 2023.

[6] B. I. Nugroho, Z. Ma'arif, and Z. Arif, "Tinjauan Pustaka Sistematis: Penerapan Data Mining Metode Klasifikasi Untuk Menganalisa Penyalahgunaan Sosial Media: Array," *J. Sist. Inf. dan Teknol. Perad.*, vol. 3, no. 2, pp. 46–51, 2022.

[7] H. P. Rajagukguk, "Pendekatan Data Mining untuk Memilih Produk Terlaris Menggunakan Algoritma Naive Bayes." Program Studi Teknik Informatika, 2023.

[8] B. Gustiandi, "Langkah Awal menguasai Bahasa pemrograman python," 2023.

[9] C. Cahyadi, "Pengaruh Kualitas Produk Dan Harga Terhadap Keputusan Pembelian Baja Ringan Di Pt Arthanindo Cemerlang," *EMaBI Ekon. dan Manaj. Bisnis*, vol. 1, no. 1, pp. 60–73, 2022.

[10] A. F. A. Rahman, "PREDIKSI KELULUSAN MAHASISWA MENGGUNAKAN ALGORITMA C4. 5 (STUDI KASUS DI UNIVERSITAS PERADABAN): Array," *Indones. J. Informatics Res.*, vol. 1, no. 2, pp. 70–77, 2020.