Prediksi Kelulusan Siswa Dengan Pendekatan Algortma C5.0 Pada SMAN 2 Cikarang Selatan

Prediction of Student Graduation Using the C5.0 Algorithm Approach at SMAN 2 South Cikarang

Putri Nabila Amir*[1], Muhamad Fatchan [2], Edora[3]
[1,2,3] *Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa*
e-mail: [1]***nblputri2304@gmail.com, [2]fatchan@pelitabangsa.ac.id, [3]edora@pelitabangsa.ac.id

Abstrak

Penelitian yang berjudul "Prediksi Kelulusan Siswa Dengan Pendekatan Algortma C5.0 Pada SMAN 2 Cikarang Selatan". Penelitian ini bertujuan untuk menerapkan metode Decision Tree C5.0 dalam memprediksi kelulusan siswa Sekolah Menengah Atas (SMA). Data kelulusan siswa dari SMA Negeri 2 Cikarang Selatan digunakan untuk membangun model prediksi kelulusan siswa. Metode Decision Tree C5.0 berhasil menghasilkan model prediksi dengan tingkat akurasi 100%. Model ini dapat mengidentifikasi siswa yang berisiko tinggi untuk tidak lulus, sehingga tindakan perbaikan yang tepat dapat diambil. Model Decision Tree C5.0 juga memberikan interpretasi aturan keputusan yang dapat digunakan oleh sekolah dan tenaga pendidik. Penelitian ini memberikan sumbangan penting bagi dunia pendidikan dengan meningkatkan efektivitas pengambilan keputusan di bidang pendidikan. Penerapan metode Decision Tree C5.0 dalam memprediksi kelulusan siswa membantu sekolah untuk mengidentifikasi siswa yang membutuhkan perhatian khusus. Dengan tingkat akurasi 100%, model ini dapat membantu meningkatkan kualitas pendidikan dan memastikan kesuksesan siswa dalam menyelesaikan pendidikan menengah atas.

Kata Kunci : *Decision Tree C5.0, Prediksi Kelulusan Siswa, Pendidikan, Akurasi*

Abstract

The research entitled "Prediction of Student Graduation with C5.0 Algorithm Approach at SMAN 2 Cikarang Selatan". This study aims to apply the Decision Tree C5.0 method to predict the graduation of high school students (SMA). Student graduation data from SMA Negeri 2 Cikarang Selatan is used to build the student graduation prediction model. The Decision Tree C5.0 method successfully produces a prediction model with 100% accuracy. This model can identify students at high risk of not graduating, enabling appropriate intervention measures to be taken. The Decision Tree C5.0 model also provides interpretation of decision rules that can be used by schools and educators. This research provides a significant contribution to the field of education by enhancing the effectiveness of decision-making in education. The application of the Decision Tree C5.0 method in predicting student graduation helps schools to identify students who need special attention. With 100% accuracy, this model can assist in improving the quality of education and ensuring the success of students in completing their high school education.

Keywords *: Decision Tree C5.0, Student Graduation Prediction, Education, Accuracy.*

PENDAHULUAN

Pendidikan adalah jembatan untuk manusia agar dapat mengembangkan potensi diri melalui proses pembelajaran yang di dapat. Tertuang di dalam UUD 1945 pasal 31 Ayat 1 yang menyebutkan bahwa: "setiap warga negara berhak mendapatkan pendidikan". Jadi, sudah jelas bahwa setiap orang berhak atas pendidikan. Pendidikan diharapkan dapat menghasilkan generasi penerus bangsa yang cerdas dan berkualitas, yang mampu memanfaatkan kemajuan

Informasi Artikel:
Submitted: Maret 2023, **Accepted:** Mei 2023, **Published:** Mei 2023
ISSN: 2685-4902 (media online), Website: http://jurnal.umus.ac.id/index.php/intech

saat ini dengan sebaik mungkin. Selain itu, generasi berikutnya akan memiliki rasa nasionalisme yang kuat. Pendidikan merupkan kunci dari kemajuan[1].

Sekolah Menengah Atas (SMA) adalah jenjang pendidikan menengah yang ditempuh setelah Sekolah Menengah Pertama (SMP) dan sebelum perguruan tinggi. Sekolah Menengah Atas (SMA) bertujuan untuk memberikan pendidikan yang lebih lanjut dan mendalam kepada siswa dalam berbagai bidang, seperti ilmu pengetahuan, sosial, dan humaniora, serta mempersiapkan siswa untuk melanjutkan pendidikan ke jenjang yang lebih tinggi atau memasuki dunia kerja. Sekolah Menengah Atas (SMA) juga dapat diartikan sebagai lembaga pendidikan formal yang menyediakan program pendidikan untuk siswa pada jenjang pendidikan menengah atas. Program pendidikan SMA biasanya mencakup mata pelajaran umum seperti matematika, bahasa Inggris, sains, dan sejarah, serta mata pelajaran pilihan seperti seni, musik, dan olahraga. Selain itu, SMA juga dapat menawarkan program keagamaan atau kejuruan, tergantung pada kebijakan dan kebutuhan masing-masing sekolah[2].

Pendidikan yang berkualitas pada Sekolah Menengah Atas (SMA) memiliki manfaat yang signifikan bagi masyarakat, terutama bagi generasi penerus bangsa yaitu siswa. Salah satu ukuran keberhasilan pendidikan adalah tingkat kelulusan siswa. Namun, tingkat kelulusan tidak selalu mencerminkan kualitas pendidikan yang baik karena bisa dipengaruhi oleh faktor-faktor eksternal seperti kondisi ekonomi dan lingkungan. Oleh karena itu, memprediksi kelulusan siswa menjadi penting dalam membantu pihak sekolah dan orang tua dalam pengambilan keputusan terkait pendidikan.

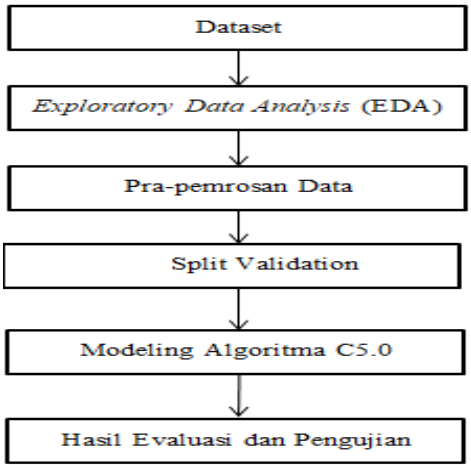
METODE PENELITIAN

Metode yang digunakan dalam penelitian ini adalah Decision Tree C5.0. Data yang digunakan berasal dari sumber data sekunder. Informasi tentang kelulusan siswa diperoleh dari SMA Negeri 2 Cikarang Selatan yang berlokasi di Jalan Utama Jl. Raya Perum Bumi Cikarang Makmur, Sukadami, Cikarang Selatan, Kabupaten Bekasi, Jawa Barat 17530. Data yang diperoleh merupakan data sekunder dalam format csv yang terdiri dari 138 kolom. Beberapa di antaranya adalah Rata Rata Semester, Nilai Ujian Sekolah (US), Nilai Sikap (NS), Jurusan, dan Status Kelulusan. Berikut adalah tabel atribut field yang dipilih dan 10 contoh data dari total 138 isi dataset :

Tabel 1. Dataset

Rata_Rata_Semester	US	NS	Jurusan	Status_Kelulusan
83	87	85	MIPA	Lulus
80	83	80	MIPA	Lulus
72	74	75	IPS	Tidak Lulus
88	91	88	MIPA	Lulus
75	75	70	MIPA	Tidak Lulus
80	84	81	MIPA	Lulus
84	88	85	MIPA	Lulus
80	83	83	MIPA	Lulus
86	89	88	IPS	Lulus
90	79	78	IPS	Lulus

Tahapan langkah-langkah peneliti untuk menjalankan tahapan atau proses penelitian :



Gambar 1 Proses Penelitian

2. 1 Algoritma Decision Tree C5.0

Decision Tree adalah algoritma pembelajaran mesin untuk klasifikasi dan regresi. Model berbentuk pohon terdiri dari simpul (fitur), cabang (keputusan), dan daun (hasil). Rekursif membagi data berdasarkan fitur yang dipilih, menciptakan pohon akurat dan efisien. Pembagian ditentukan dengan mengevaluasi kenaikan informasi atau indeks Gini, mencari atribut terbaik untuk subset yang murni [3].

Algoritma C5.0 adalah metode pembentukan pohon keputusan untuk klasifikasi data dengan menggunakan entrophy dan information gain. Atribut dengan information gain tertinggi menjadi simpul akar pohon keputusan. Penting untuk memilih fitur yang memberikan informasi terbanyak saat mengelompokkan objek ke dalam kelas. Simpul akar ditentukan oleh atribut dengan gain tertinggi dan entrophy terendah [4].

Untuk menentukan nilai entropi dan gain. Rumus untuk menghitung nilai entropi dapat ditemukan dalam persamaan(1) [5]:

$$\text{Entropy}(S) = -P(+)\log 2P(+) - P(-)\text{-}\log 2P(-) \qquad \text{persamaan(1)}$$

Dimana:

S : Sampel yang digunakan untuk pembelajaran (data).
P (-) : jumlah solusi negatif atau kuantitas data negatif untuk kriteria yang relevan, berdasarkan data sampel.
P (+) : jumlah solusi positif atau kuantitas data positif untuk kriteria yang relevan, berdasarkan data sampel.
Entropi (S) : 0 ketika setiap contoh dari S berada dalam kelompok yang sama.
Entropi (S) : 1 ketika jumlah sampel positif dan negatif dalam S sama.

Nilai entropi (S) berada di antara 0 dan 1 jika jumlah sampel positif dan negatif dalam S tidak sama. Selanjutnya, untuk mencari nilai gain, digunakan persamaan (2).

$$\text{Gain}(S,A) = \textit{Entropy}(S) - \sum_{i=1}^{n} - \frac{|Si|}{|S|} * \textit{Entropy} \qquad \text{persamaan(2)}$$

Keterangan :

S : Himpunan Kasus
A : Atribut
n : Jumlah Partisi Atribut A
| Si | : Jumlah Kasus Pada Partisi Ke – i

| S | : Jumlah Kasus Dalam S

Karena Si merupakan himpunan kasus dalam kelas ke-i, dan A adalah komponen yang digunakan, sedangkan n adalah banyaknya kelas pada komponen A, maka |Si| adalah jumlah kasus dalam kategori ke-i, dan |S| adalah jumlah kasus dalam keseluruhan S.

Rumus dasar untuk menghitung rasio keuntungan adalah Persamaan (3).:

$$\text{Gain Rasio} = \frac{\text{Gain}(S,A)}{\sum_{i=1}^{n} Entrophy\,(Si)} \quad \text{persamaan(3)}$$

Gain (S, A), disnini mengacu pada nilai gain dari variabel, yaitu jumlah nilai entropi dalam variabel yang dinyatakan sebagai $\sum_{i=1}^{n} Entropy(Si)$.

2. 2 Data Mining

Data mining adalah serangkaian proses yang berguna untuk menggali dan mencari nilai-nilai berupa informasi dan hubungan kompleks yang tersembunyi dalam suatu basis data. Dengan melakukan analisis pola informasi pada data, kita dapat memanipulasi data tersebut menjadi informasi baru yang lebih bermanfaat. Data mining juga membantu dalam mengidentifikasi dan mengekstraksi pola-pola berharga atau menarik dari data yang ada dalam basis data. Penggunaan data mining dapat membantu dalam mengelola data yang besar dan memfasilitasi penyimpanan data transaksi serta pengolahan data gudang data (data warehousing) untuk mendapatkan informasi yang relevan bagi pengguna[6].

2. 3 Tahapan Data Mining

Bagian dari proses *knowledge discovery from data* (KDD) adalah data mining. Berikut adalah tahapan proses KDD[7]:

1. Data selection
   Pada tahap pertama, peneliti menggunakan database kelulusan siswa dari SMA Negeri 2 Cikarang Selatan sebagai input, dan proses ini menghasilkan database kelulusan siswa yang dipilih pada tahun 2020.
2. Prepocessing
   Tahap kedua terdiri dari pembersihan dan preprocessing data untuk memastikan kualitas data yang baik. Ini termasuk menghilangkan data yang tidak relevan, mengatasi nilai yang hilang, dan mengatasi data yang tidak seimbang.
3. Transformation:
   Pada tahap ini, data diubah untuk menyesuaikan dengan jenis atau pola informasi yang dicari. Ini dilakukan berdasarkan data yang telah dibuat pada tahap sebelumnya dan disesuaikan dengan kebutuhan peneliti untuk mendapatkan hasil analisis yang lebih akurat dan sesuai dengan tujuan penelitian.
4. Data mining
   Pada tahap ketiga, algoritma data mining, metode Decission Tree C5.0, digunakan untuk melakukan proses prediksi kelulusan siswa. Proses ini mencakup pencarian pola atau informasi pada data yang telah dipilih menggunakan metode atau teknik tertentu. Output dari proses ini dikumpulkan melalui beberapa kali proses pelatihan untuk menghasilkan pola informasi yang dianggap memahami oleh peneliti.
5. Interpretation/evaluation

Pada tahap terakhir, hasil data mining harus diinterpretasikan sehingga orang lain dapat memahaminya. Ini adalah proses menerjemahkan pola data atau informasi yang telah dikumpulkan ke dalam bentuk yang lebih mudah dipahami untuk semua pihak yang berkepentingan, termasuk sekolah dan siswa.

2. 4 Exploratory Data Analysis

Exploratory Data Analysis (EDA) adalah metode visual dan deskriptif dalam analisis data untuk mempelajari karakteristik, pola, dan anomali. Tujuan utamanya adalah mencari pola data, sejalan dengan data mining. Di era big data, eksplorasi pola data jadi lebih sulit karena volume data yang besar. EDA berguna untuk meningkatkan pemahaman analisis data lewat visualisasi atau reduksi dimensi. Ini juga membantu mengoptimalkan pengetahuan tentang data, mengidentifikasi variabel penting, menemukan anomali, outlier, dan menguji asumsi awal. Melalui data mining, semua ini dapat diterapkan untuk meningkatkan analisis data dan optimalisasi hasil klasifikasi [8].

2. 5 Split Validation

Validasi Pemisahan (*Split Validation*) adalah teknik yang digunakan dalam pembelajaran mesin untuk mengevaluasi performa suatu model. Metode ini melibatkan pembagian data yang tersedia menjadi dua set, yaitu set pelatihan dan set validasi. Set pelatihan digunakan untuk melatih model, sementara set validasi digunakan untuk mengukur kinerja model. Keunggulan dari validasi pemisahan adalah kemampuannya untuk melakukan evaluasi cepat terhadap performa model tanpa memerlukan data tambahan. Namun, metode ini dapat sensitif terhadap pembagian data, dan dalam beberapa kasus, validasi silang menjadi pilihan yang lebih baik[9].

2.6 Counfusion Matrix

Matriks kebingungan(*Counfusion Matrix)* adalah tabel untuk menggambarkan kinerja model klasifikasi. Tabel ini menunjukkan jumlah prediksi benar positif, benar negatif, positif salah, dan negatif salah. Ini digunakan untuk menghitung metrik evaluasi seperti akurasi, presisi, recall, dan skor F1. Matriks kebingungan banyak digunakan dalam pembelajaran mesin dan tugas klasifikasi untuk mengevaluasi model. Memberikan pemecahan rinci tentang prediksi model dan membantu memahami kesalahan yang dibuat oleh model.

Metrik-metrik evaluasi tersebut memberikan wawasan tentang kemampuan model untuk mengklasifikasikan instance dengan benar dan mengidentifikasi prediksi benar positif, benar negatif, positif salah, dan negatif salah. Contohnya, akurasi dihitung dengan membagi jumlah prediksi benar positif dan benar negatif dengan total instance. Presisi dihitung sebagai rasio prediksi benar positif dibagi dengan total prediksi positif, termasuk prediksi benar positif dan positif salah. Recall dihitung sebagai rasio prediksi benar positif dibagi dengan total instance yang seharusnya diprediksi positif, termasuk prediksi benar positif dan negatif salah. Skor F1 adalah kombinasi presisi dan recall, memberikan ukuran seimbang kinerja model.

Matriks kebingungan penting untuk memahami kelebihan dan kelemahan model klasifikasi serta membimbing perbaikan atau penyesuaian lebih lanjut pada model [10].

2. 7 Bahasa Pemograman Python

Python adalah bahasa pemrograman yang populer dan sering digunakan dalam pengembangan machine learning, NLP (Natural Language Processing), dan neural network. Python memiliki beberapa keunggulan diantaranya sebagai bahasa pemrograman, terutama dalam konteks pengembangan machine learning dan data science, keunggulan bahasa pemograman python diantaranya populer dalam pengembangan machine learning, mudah dipelajari maupun digunakan, dapat diintegrasikan dengan baik, Dukungan pustaka yang luas, lintas platform, dan sumber terbuka.

Python juga banyak digunakan dalam bidang data science. Python memiliki banyak pustaka (library) yang mendukung penelitian, seperti Numpy, Pandas, Sklearn, Matplotlib, dan lainnya. Berikut ini penjelasan pustaka (library) seperti Numpy, Pandas, Sklearn, Matplotlib[11]:

1. Numpy adalah pustaka yang digunakan untuk melakukan operasi matematika dan manipulasi array multidimensi. Pustaka ini menyediakan fungsi-fungsi yang efisien untuk melakukan perhitungan numerik dalam Python
2. Pandas adalah pustaka yang digunakan untuk melakukan manipulasi dan analisis data. Pustaka ini menyediakan struktur data yang fleksibel dan efisien, seperti DataFrame, yang memungkinkan pengguna untuk melakukan operasi seperti filtering, grouping, dan joining data.
3. Sklearn (Scikit-Learn) adalah pustaka yang digunakan untuk machine learning. Pustaka ini menyediakan berbagai algoritma dan fungsi yang dapat digunakan untuk melakukan pemodelan, evaluasi, dan prediksi dalam machine learning.
4. Matplotlib adalah pustaka yang digunakan untuk visualisasi data. Pustaka ini menyediakan fungsi-fungsi untuk membuat berbagai jenis plot, seperti scatter plot, line plot, dan histogram, sehingga memudahkan pengguna untuk memvisualisasikan data dan hasil analisis.

Dalam penelitian ini, pustaka-pustaka tersebut digunakan untuk melakukan pra-pemrosesan data, proses machine learning dengan Decision Tree, dan visualisasi hasil.

HASIL DAN PEMBAHASAN

3.1 Import Libray dan Dataset

Langkah pertama dalam langkah ini, akan memanfaatkan library-library yang sering digunakan dalam analisis dan pembelajaran mesin, seperti pandas, numpy, dan sklearn. Selain itu, dataset yang relevan juga harus diimpor agar dapat digunakan dalam proses pengolahan dan pembelajaran. Berikut ini adalah kode untuk langkah ini

```
# Import libray yang dibutuhkan
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import precision_score, recall_score, f1_score, accuracy_score
from sklearn.metrics import roc_curve, roc_auc_score
from sklearn import preprocessing
from sklearn import tree


# Memuat dataset
df = pd.read_csv('Dataset.csv')
df.index = df.index + 1
# Menampilkan informasi dataset
df.head(5) # Menampilkan beberapa baris pertama dari dataset
```

Gambar 2. *Import Libray* dan Dataset

kita telah mengimport library-library yang diperlukan, seperti pandas, numpy, seaborn, train_test_split dari sklearn.model_selection, *DecisionTreeClassifier* dari sklearn.tree, precision_score dari sklearn.metrics, recall_score dari sklearn.metrics, f1 score dari sklearn.metrics dan seterusnya seperti pada gambar 2. Selanjutnya, menggunakan fungsi pd.read_csv untuk memuat dataset dari file CSV. Setelah memuat dataset, kita dapat menampilkan beberapa baris pertama dari dataset menggunakan dataset.head() dan melihat dimensi dataset menggunakan dataset.shape.

3.2 Pre-processing Data

Setelah mengimpor dataset, langkah ke dua adalah melakukan pra-pemrosesan data guna mempersiapkan dataset sebelum diaplikasikan pada algoritma Decision Tree C5.0. Pada tahap ini, penulis melakukan pemeriksaan dan penghapusan data yang potensial mengandung noise atau nilai yang tidak valid. Berikut kode program dan hasil untuk langkah ini:

```
# Fungsi reusable pribadi saya untuk mendeteksi data yang hilang
def missing_value_describe(data):
# Periksa nilai yang hilang dalam data
    missing_value_stats = (data.isnull().sum() / len(data)*100)
    missing_value_col_count = sum(missing_value_stats > 0)
    missing_value_stats = missing_value_stats.sort_values(ascending=False)[:missing_value_col_count]
    print("Jumlah kolom dengan nilai yang hilang:", missing_value_col_count)
    if missing_value_col_count != 0:
        # Mencetak nama kolom dengan persentase nilai yang hilang
        print("\nPersentase hilang (menurun):")
        print(missing_value_stats)
    else:
        print("Tidak ada data yang hilang!!!")
missing_value_describe(df)
```

```
Jumlah kolom dengan nilai yang hilang: 0
Tidak ada data yang hilang!!!
```

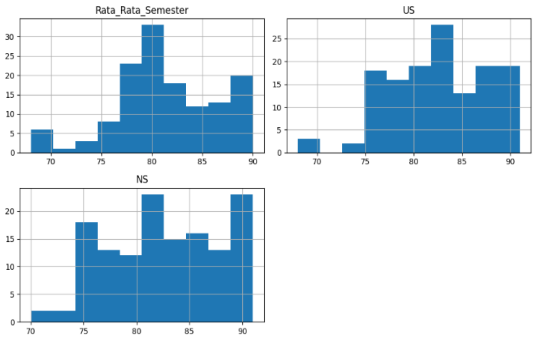
Gambar 3. Pre-processing Data

Data ini siap untuk *Exploratory Data Analysis*, seperti yang ditunjukkan pada Gambar 4.

3.3 Exploratory Data Analysis

Langkah ketiga adalah melakukan *Exploratory Data Analysis* (EDA). Langkah ini membantu menemukan pola, hubungan, dan karakteristik penting dalam data sebelum memulai proses pemodelan. Visualisasi akan dilakukan dengan beberapa cara, seperti grafik untuk setiap variabel numerik, matriks korelasi, dan grafik untuk melihat bagaimana dua variabel numerik dan variabel kategorikal berinteraksi satu sama lain.

a) Variabel Numerik,

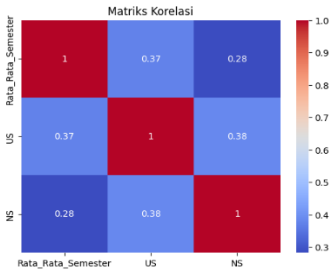


Gambar 4. Histogram untuk setiap variabel numerik

Gambar 5 menunjukkan histogram dengan distribusi miring ke kiri, yang berarti bahwa nilai-nilai data akan muncul dengan frekuensi yang lebih tinggi pada nilai-nilai yang lebih tinggi. Histogram dengan distribusi miring ke kiri akan memiliki puncak yang lebih rendah dan ekor yang lebih panjang di sebelah kiri.

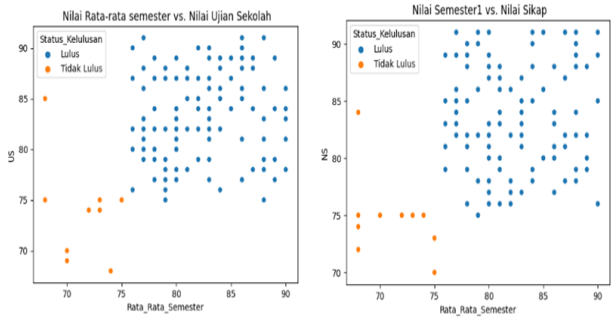
b) Matriks Korelasi

Karena banyak korelasi di bawah 0,6 menunjukkan tidak adanya korelasi antara dua variabel dalam dataset, matriks korelasi ini menunjukkan bahwa hubungan antara variabel dalam dataset tidak cukup baik.



Gambar 5. Matriks Korelasi

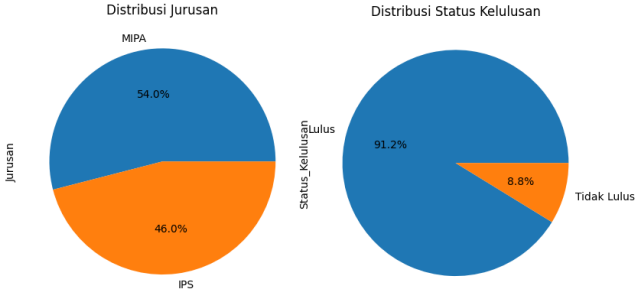
c) Dua Variabel Numerik



Gambar 6. Dua Variabel Numerik

Pada Gambar 7 dapat disimpulankan bahwa siswa yang dinyatakan tidak lulus mempunyai niali kisaran 68 – 75 dan siswa yang dinyatakan lulus mempunyai nilai kisaran 76 – 91.

d) Variabel Kategorikal



Gambar 7. Pie Chart variabel kategorikal

Pada Gambar 8untuk disribusi jurusan terdapat 54 % kelas MIPA dan 46% kelas IPS. Untuk distribusi status kelulusan siswa yang lulus sebesar 91,2 % dan siswa yang tidak lulus 8,8%.

3.4 Transformasi Data

Setelah proses *Exploratory Data Analysis* (EDA) selesai, langkah ke empat adalah melakukan transformasi data. Transformasi data adalah proses mengubah atau memanipulasi data agar sesuai dengan tujuan analisis yang akan dilakukan. Berikut adalah penjelasan dalam transformasi data:

Tabel 2. Proses Data Transformation

No	Atribut	Subset	Nilai
1	Rata Rata Semester	< 60	1
		60-75	2
		75-85	3
		85-100	4
2	US	< 60	1
		60-75	2
		70-85	3
		85-100	4
3	NS	< 60	1
		60-75	2
		70-85	3
		85-100	4
4	Jurusan	MIPA	1
		IPS	0

```
# Daftar fitur yang perlu ditransformasi
fitur = ['Rata_Rata_Semester','US','NS']

# Kriteria transformasi
batas = [0, 60, 75, 85, 100]
labels = [1, 2, 3, 4]

# Melakukan transformasi data

label_encoder = preprocessing.LabelEncoder()
X['Jurusan']= label_encoder.fit_transform(X['Jurusan'])

for f in fitur:
    X[f] = pd.cut(X[f], bins=batas, labels=labels, include_lowest=True)

# Menampilkan dataset setelah transformasi
print(X.head())
```

Gambar 8 Kode Program Transformasi Data

Tabel 3. Hasil Data Transformation

Nama Siswa	Rata Rata Semester	US	NS	Jurusan	Status Kelulusan
Ahmad Syahputra	3	4	3	1	Lulus
Siti Fatimah	3	3	3	1	Lulus
Budi Santoso	2	2	2	0	Tidak Lulus
Rina Novianti	4	4	4	1	Lulus
Riko Sebastian	2	2	2	1	Tidak Lulus

Faisal Rahman	3	3	3	1	Lulus
Rizki Pratama	3	4	3	1	Lulus
...	...	...	...	...	...
Maya Sari	3	3	3	1	Lulus
Adi Nugroho	4	4	4	0	Lulus

3.5 Validasi Pemisahan (Split Validation)

Setalah transformasi, langkah ke lima melakukan validasi pemisahan(*split validation*) dimana bagi data menjadi data pelatihan dan data pengujian. Data pelatihan seperti X_train dan y_train digunakan untuk melatih model, sedangkan data pengujian seperti X_test dan y_test digunakan untuk menguji kinerja model yang telah dilatih. Pada penelitian kali ini data dibagi dalam perbandingan 80:20. Berikut kode program *split validation*:

```
# Membagi data menjadi data latih dan data uji
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=52)
```

Gambar 9. Validasi Pemisahan (*Split Validation*)

3.6 Pembentukan Model Decision Tree C5.0

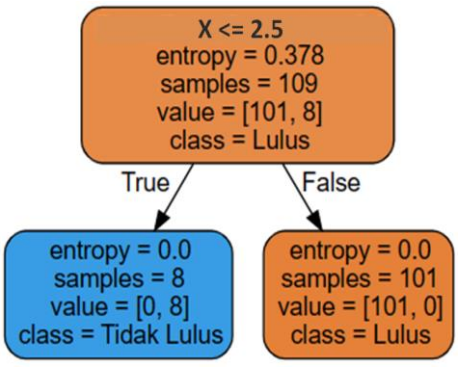
Setelah validasi pemisahan (*split validation*), langkah terakhir adalah membangun model menggunakan algoritma Decision Tree C5.0. Dalam langkah ini, penulis akan melaksanakan visualisasi hasil prediksi dan kode program untuk melatih model Decision Tree C5.0 menggunakan data pelatihan yang telah dipisahkan sebelumnya. Setelah melakukan ini, model akan dipelajari dan disesuaikan dengan pola data pelatihan yang ada, sehingga penulis dapat menghasilkan aturan keputusan yang dapat digunakan untuk memprediksi kelulusan siswa. Dengan algoritma Decision Tree C5.0, diharapkan model yang dibangun mampu memberikan prediksi kelulusan yang akurat berdasarkan faktor-faktor yang relevan.

```
# Membuat model Decision Tree C5.0
C5_0 = DecisionTreeClassifier(criterion="entropy",min_samples_split=75 , max_leaf_nodes=5)

# Melatih model dengan data latih s
C5_0.fit(X_train, y_train)

# Memprediksi data uji
y_pred = C5_0.predict(X_test)
```

Gambar 10. Kode Program Decision Tree C5.0



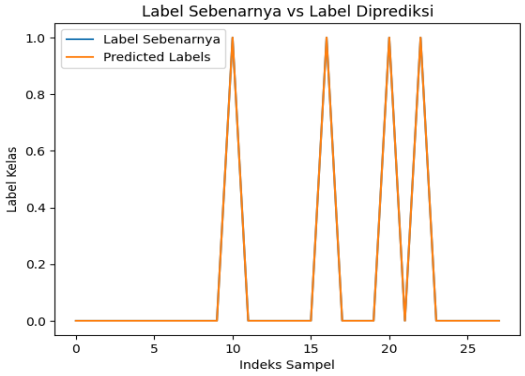
Gambar 11. Decission Tree.

Sehingga diperoleh suatu pohon keputusan dengan aturan sebagai berikut:

1. IF “Rata_Rata_Semester” <= 2.50 and “US” <= 2.50 and “NS” <= 2.50 THEN Tidak Lulus.
2. IF “Rata_Rata_Semester” > 2.50 and “US” > 2.50 and “NS” > 2.50 THEN Lulus.

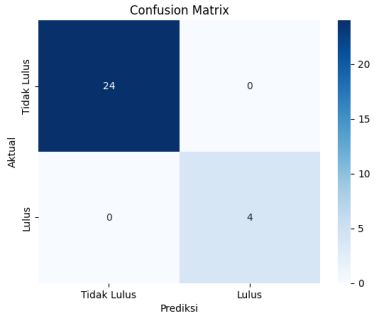
```
|--- feature   <= 2.50
|   |--- class: Tidak Lulus
|--- feature   >  2.50
|   |--- class: Lulus
```

Gambar 12. Deskripsi Decission Tree



Gambar 13. Visualisasi Hasil Prediksi Algoritma Decission Tree C5.0

Berdasarkan visualisasi pada Gambar 13, dapat ditarik kesimpulan bahwa model menghasilkan kinerja yang sangat baik karena mampu secara efektif mengklasifikasikan data ke dalam kelas 1(Tidak Lulus) dan kelas 0(Lulus). Selanjutnya, validasi dan pengukuran keakuratan hasil yang dicapai oleh model dilakukan dengan melihat confussion matrix, seperti yang ditunjukkan pada gambar 14.



Gambar 14. Confusion matrix

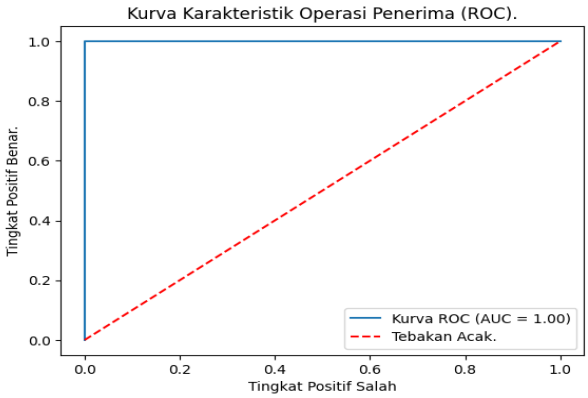
Pada gambar 14 terdapat 24 prediksi yang benar untuk kelas prediksi lulus (*True* lulus) dan 4 prediksi yang benar untuk kelas prediksi tidak lulus (*True* tidak lulus). Tidak ada kasus yang salah dalam prediksi sebagai kelas tidak lulus (*False* tidak lulus) maupun kelas lulus (*False* lulus).

Selanjutnya akan melihat hasil *Accuracy*, *Precision*, *Recall* dan *F1 Score*. Hasil analisis yang dilakukan menggunakan pembelajaran mesin (machine learning) menggunakan bahasa pemograman Python dengan pengukuran Decission tree c5.0 ditunjukkan dalam tingkat:

Tabel 4. Accuracy, Precision, Recall dan F1 Score

Accuracy	*Precission*	*Recall*	*F1 Score*
100%	100%	100%	100%

Untuk melakukan evaluasi secara komprehensif, metrik *ROC* dan *AUC* digunakan:



Gambar 15. Visualisasi ROC dan AUC

Berdasarkan visualisasi pada Gambar 14 hasil evaluasi penelitian ini masuk ke dalam tingkat diagnosa Excellent classification karena menunjukkan nilai AUC = 1.0. Nilai AUC memiliki kinerja yang sangat baik dalam membedakan antara kelas positif dan negatif.

KESIMPULAN

Kesimpulan dari penerapan metode Decision Tree C5.0 di SMA Negeri 2 Cikarang Selatan adalah bahwa model prediksi yang dihasilkan efektif dan akurat dalam mengklasifikasikan siswa sebagai lulus atau tidak lulus, dengan tingkat akurasi mencapai 100%. Penggunaan atribut rata-rata semester, nilai ujian sekolah (US), dan nilai sikap (NS) dalam model tersebut berhasil menghasilkan prediksi yang akurat. Evaluasi menggunakan metrik ROC dan AUC menunjukkan nilai 1.0, mengindikasikan bahwa model ini sangat mampu dalam membedakan siswa yang lulus dan tidak lulus, serta performanya dalam mengklasifikasikan siswa ke dalam kategori yang benar sangat baik. Secara keseluruhan, penerapan metode Decision Tree C5.0 dalam prediksi kelulusan siswa di SMA Negeri 2 Cikarang Selatan membawa hasil yang memuaskan dan dapat diandalkan. Tingkat akurasi yang tinggi dan kemampuan model dalam memprediksi kelulusan siswa menunjukkan bahwa pendekatan ini adalah solusi efektif untuk menghadapi masalah prediksi kelulusan di sekolah tersebut. Dengan hasil evaluasi yang sangat baik, model ini dapat menjadi alat yang berguna bagi pihak sekolah dalam mengidentifikasi siswa yang berisiko tinggi untuk tidak lulus dan memberikan intervensi lebih lanjut agar dapat meningkatkan peluang kelulusan siswa.

DAFTAR PUSTAKA

[1] S. F. N. Fitri, “Problematika Kualitas Pendidikan di Indonesia,” *J. Pendidik. Tambusai*, vol. 5, no. 1, pp. 1617–1620, 2021.

[2] J. Educatio, “Keberagamaan dan Pola Belajar Siswa Berprestasi Akademik di Sekolah Menengah Atas,” vol. 9, no. 2, pp. 831–840, 2023, doi: 10.31949/educatio.v9i2.5016.

[3] V. Rahmayanti, Y. Azhar, and A. E. Pramudita, “Penerapan algoritma C5.0 pada analisis faktor-faktor pengaruh kelulusan tepat waktu mahasiswa Teknik Informatika UMM,” *J. Repos.*, vol. 1, no. 2, p. 131, 2020, doi: 10.22219/repositor.v1i2.545.

[4] K. Smk, M. A. Arif, N. U. Al, and M. Bekasi, "1300-Article Text-2682-1-10-20220906," *J. Teknol. Pelita Bangsa*, vol. 11, no. 1, 2020.

[5] A. Charis Fauzan, "Penerapan Algoritma Decision Tree C5.0 Untuk Memprediksi Tingkat Kematian Pasien Penyakit Gagal Jantung Application of The C5.0 Decission Tree Algorithm to Predict Patient Mortality Rate Heart Failure Disease," *J. Ilm. Intech Inf. Technol. J. UMUS*, vol. 4, no. 02, pp. 216–222, 2022.

[6] D. P. Utomo and B. Purba, "Penerapan Datamining pada Data Gempa Bumi Terhadap Potensi Tsunami di Indonesia," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, no. September, p. 846, 2019, doi: 10.30645/senaris.v1i0.91.

[7] W. Cholil, A. R. Dalimunthi, and L. Atika, "Model Data Mining Dalam Mengidentifikasi Pola Laju Pertumbuhan Antar Sektor Ekonomi di Provinsi Sumatera Selatan dan Bangka Belitung," *Teknika*, vol. 8, no. 2, pp. 103–109, 2019, doi: 10.34148/teknika.v8i2.181.

[8] E. D. Wahyuni, A. A. Arifiyanti, and M. Kustyani, "Exploratory Data Analysis dalam Konteks Klasifikasi Data Mining," *Pros. Nas. Rekayasa Teknol. Ind. dan Inf. XIV Tahun 2019*, vol. 2019, no. November, pp. 263–269, 2019.

[9] E. Retnoningsih and R. Pramudita, "Mengenal Machine Learning Dengan Teknik Supervised Dan Unsupervised Learning Menggunakan Python," *Bina Insa. Ict J.*, vol. 7, no. 2, p. 156, 2020, doi: 10.51211/biict.v7i2.1422.

[10] I. Düntsch and G. Gediga, "Confusion Matrices and Rough Set Data Analysis," *J. Phys. Conf. Ser.*, vol. 1229, no. 1, 2019, doi: 10.1088/1742-6596/1229/1/012055.

[11] S. P. Barus, "Penerapan Model Decision Tree pada Machine Learning untuk Memprediksi Calon Potensial Mahasiswa Baru," *J. Ikraith Inform.*, vol. 6, no. 2, pp. 59–62, 2022.