**LAPORAN DATA MINING**

****

**Disusun Oleh:**

Isa Iant Maulana (A11.2022.14416)

A’laqim (A11.2022.14413)

Wahyu Aji Pamungkas (A11.2022.14595)

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS DIAN NUSWANTORO**

**2024/2025**

# DAFTAR ISI

[DAFTAR ISI 2](#_Toc156137252)

[BAB I PENDAHULUAN 3](#_Toc156137253)

[1.1. Latar Belakang 3](#_Toc156137254)

[BAB II TINJAUAN PUSTAKA 4](#_Toc156137255)

[2.1. Tinjauan Pustaka 4](#_Toc156137257)

[2.2. Data Mining 4](#_Toc156137260)

[2.3. Klasifikasi 5](#_Toc156137264)

[BAB III Rumusan Masalah, Tujuan Penelitian Dan Metode Analisis Data 6](#_Toc156137265)

[3.1. Rumusan Masalah 6](#_Toc156137267)

[3.2. Metode Analisis Data 6](#_Toc156137270)

[3.3. Tujuan Penelitian 6](#_Toc156137276)

[3.4. Manfaat Penelitian 6](#_Toc156137277)

[BAB IV PEMBAHASAN 7](#_Toc156137278)

[4.1. Data Penelitian 7](#_Toc156137280)

[4.2. Pembersihan Data 7](#_Toc156137281)

[4.3. Normalisasi Data 8](#_Toc156137282)

[4.4. Visualisasi Data 9](#_Toc156137283)

[4.5. Evaluasi Data 11](#_Toc156137284)

[4.6. Uji Performa Model 12](#_Toc156137285)

[4.7. Memvisualisasikan Decision Tree 13](#_Toc156137286)

[BAB V KESIMPULAN DAN SARAN 15](#_Toc156137287)

# BAB I PENDAHULUAN

## Latar Belakang

Data Mining adalah Proses Pengumpulan dan pengolahan data yang bertujuan untuk mengekstrak informasi penting pada data. Proses pengumpulan dan ekstraksi informasi tersebut dapat dilakukan menggunakan perangkat lunak dengan bantuan perhitungan statistik, matematika, ataupun teknologi AI (Artificial Intelligence).

Ada banyak teknik di dalam Data Mining Salah satu teknik dalam data mining adalah klasifikasi. Klasifikasi dilakukan untuk mengelompokkan sekumpulan data ke dalam kelas-kelas yang telah ditentukan terlebih dahulu berdasarkan kesamaan karakteristik yang dimiliki. Klasifikasi banyak digunakan diberbagai bidang, misal aplikasi klasifikasi di bidang pengenalan pola, diagnosis medis, deteksi kegagalan di industri dan lain sebagainya.

Klasifikasi adalah teknik dalam bidang machine learning dan data mining yang bertujuan untuk mengelompokkan objek atau data ke dalam kategori atau kelas yang telah ditentukan. Tujuan utama klasifikasi adalah mengembangkan model yang dapat memahami pola dan relasi dalam data, sehingga model tersebut dapat digunakan untuk mengklasifikasikan data baru yang belum pernah dilihat sebelumnya. Dengan demikian, klasifikasi memiliki potensi untuk memberikan wawasan yang berharga dan mendukung pengambilan keputusan yang lebih akurat.

Tantangan utama dalam klasifikasi adalah mengembangkan model yang dapat melakukan prediksi dengan tingkat akurasi yang tinggi, terutama saat dihadapkan pada data yang kompleks dan tidak terstruktur. Selain itu, penting untuk memperhatikan aspek etika dan privasi dalam penerapan teknik klasifikasi, mengingat sensitivitas informasi yang mungkin terlibat.

# BAB II TINJAUAN PUSTAKA



## Tinjauan Pustaka

Kami melakukan penelitian ini bertujuan untuk mencari label (juga dikenal sebagai kategori atau atribut target) atau mencari siswa/siswi dari SMAN 1 SETU yang masuk eligible. Cara yang digunakan dalam penelitian ini adalah dengan metode Klasifikasi.

Penelitian yang kami lakukan berfungsi mengembangkan model klasifikasi data mining untuk melihat nilai-nilai siswa/siswi SMAN 1 SETU manakah yang lebih dari target dan manakah yang kurang dari target. Hasil yang diperoleh dalam penelitian ini adalah ada sekitar 41% siswa dan siswi yang eligible, yang pada akhirnya dapat mempermudah para Guru untuk memilih siswa/siswi yang lulus menjadi eligible.

Penelitian yang sekarang direncanakan untuk mengaplikasikan model data mining pada proses estimasi pemilihan siswa/siswi yang nilainya lebih dari target yang ditentukan oleh sekolah dan. Proses ini bermanfaat bagi pihak sekolah untuk mengumpulkan data semua nilai siswanya yang lebih dari target.



## Data Mining

Data yang digunakan dalam kegiatan ini adalah data kesiswaan yaitu data nilai akhir para siswa di SMAN 1 Setu. Data yang berkaitan dengan nilai para siswa yang diambil dari semester 1 - 5 yang dirata-rata dan menghasilkan nilai akhir. Setiap atribut dalam dataset ini mencakup nilai - nilai untuk setiap siswa beserta label yang menunjukkan apakah siswa tersebut dianggap “eligible” atau “tidak eligible”.

Setiap atribut dataset ini terdiri dari berbagai mata pelajaran seperti Agama, PPKN, Bahasa Indonesia, Sejarah, Seni, PJOK, TIK, Geografi, Sosio, Ekonomi, Kimia, Bahasa Jepang, PKWU, Fisika, Biologi. Setiap siswa diberi atribut “Eligible” yang dimana itu merupakan label untuk menentukan klasifikasi dalam Data Mining. Dengan menggunakan dataset ini, analisis dapat dilakukan untuk menentukan faktor-faktor yang mempengaruhi label para siswa , membantu pengambilan keputusan atau kebijakan tertentu.



## Klasifikasi

Klasifikasi merupakan sebuah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep dan kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang kelasnya tidak diketahui.

Pada dasarnya ada dua langkah proses dalam klasifikasi data, yaitu membangun model berdasarkan sekumpulan data yang telah ditetapkan sebagai training set dan menggunakan model untuk klasifikasi. Banyak algoritma yang telah dikembangkan untuk melakukan proses klasifikasi, antara lain Decision Tree (Pohon Keputusan), Klasifikasi Bayesian, Neural Network dan lain-lain

# BAB III Rumusan Masalah, Tujuan Penelitian Dan Metode Analisis Data



## Rumusan Masalah

Bagaimana melakukan prediksi apakah siswa tersebut termasuk kedalam kategori Eligible atau tidak dengan menggunakan teknik klasifikasi.



## Metode Analisis Data

Data dianalisis menggunakan salah satu algoritma klasifikasi yaitu algoritma Decision Tree. Dalam algoritma ini setiap alternatif kemungkinan yang akan terjadi ditelaah lebih lanjut.



## Tujuan Penelitian

Mencari perumusan proses penyaringan data dengan akurat dan terpercaya menggunakan teknik klasifikasi berdasarkan label yang berada disetiap data siswa/siswi.

## Manfaat Penelitian

Hasil penelitian ini akan memberikan kontribusi pemanfaatan dataset yang ada dibidang Pendidikan sebagai dasar untuk pengelompokan nilai siswa yang lebih dari target dan yang kurang dari target menggunakan teknik klasifikasi pada kolom label pada dataset. Pada bidang pendidikan ini bisa sangat memudahkan para guru yang ada disekolah yang siswanya sangat banyak dengan penelitian ini dapat mempersingkat waktu untuk pengelompokan siswanya.

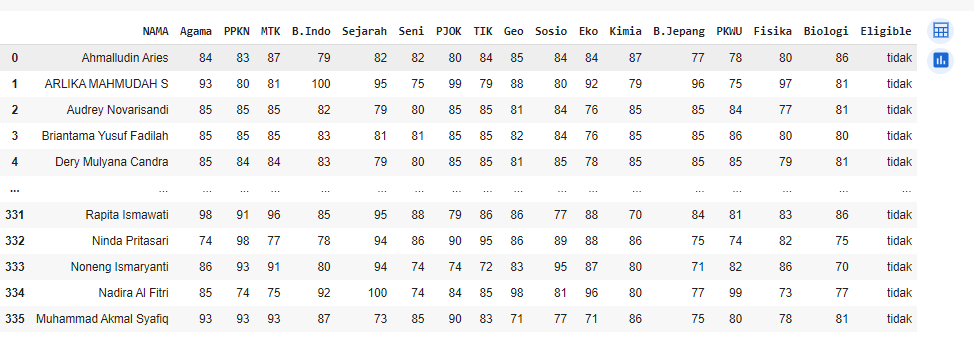
# BAB IV PEMBAHASAN



## Data Penelitian

Data penelitian yang kami ambil adalah data private.Data penelitian ini diambil dari nilai rata - rata atau akhir dalam kegiatan pembelajaran yang dilakukan oleh siswa - siswi SMAN 1 Setu. Label dikelompokkan menjadi 2 yaitu :

1. Label yang berisi “ya” merupakan siswa yang termasuk kedalam kategori eligible.
2. Label yang berisi “tidak” merupakan siswa yang tidak termasuk kedalam kategori eligible/



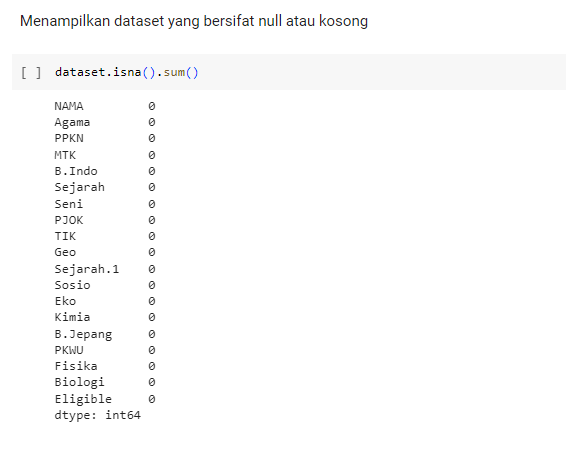
Tabel 4.1 Data Penelitian

Terlihat bahwa beberapa nama orang memiliki berbagai atribut dan Label sendiri - sendiri. Jumlah siswa dari dataset yang kami dapatkan adalah 336 siswa dan siswi yang sudah dilabeli sesuai dengan labelnya.

## Pembersihan Data

Sebelum dilakukan proses data mining, perlu dilakukan proses pembersihan data yang guna untuk mengidentifikasi, menanggulangi, dan mengatasi ketidakakuratan, ketidaklengkapan , dan ketidakteraturan dalam dataset.

Tujuannya adalah untuk memastikan bahwa data yang akan digunakan siap untuk diproses dan digunakan untuk menguji model Machine Learning. Pembersihan data sangat penting karena kualitas data yang buruk akan menghasilkan model yang tidak akurat.



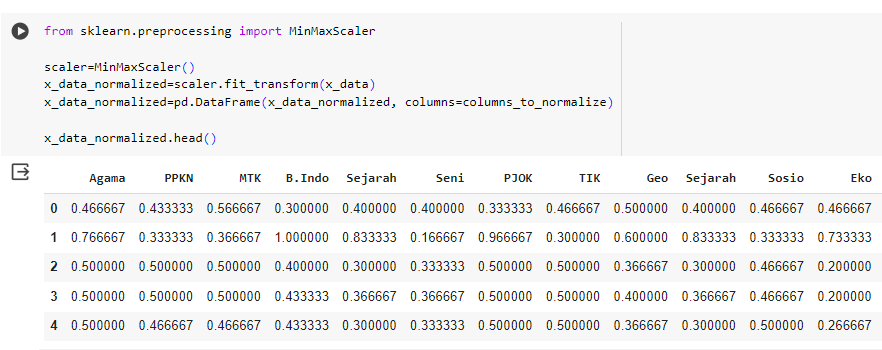
Gambar 4.2 Pembersihan Data

Terlihat digambar bahwa tidak ada dataset yang bersifat null atau kosong. Jika ada dataset yang kosong atau null maka kita harus membersihkan datanya atau menghapus datanya karena itu akan berpengaruh kepada akurasi untuk pemodelan.

## Normalisasi Data

Normalisasi Data adalah sebuah proses transformasi data pada suatu skala tertentu hingga data tersebut memiliki rentang atau skala yang seragam. Tujuan melakukan normalisasi adalah untuk membawa nilai - nilai dari berbagai fitur atau variabel dalam dataset ke dalam rentang yang konsisten , sehingga setiap fitur memiliki dampak yang seimbang pada analisis atau model machine learning.

Di Dalam normalisasi data, kami menggunakan salah satu metode yang umum digunakan yaitu Min - Max Scaling, yang merubah setiap nilai dalam fitur menjadi persentase relatif terhadap rentang nilai di fitur tersebut.

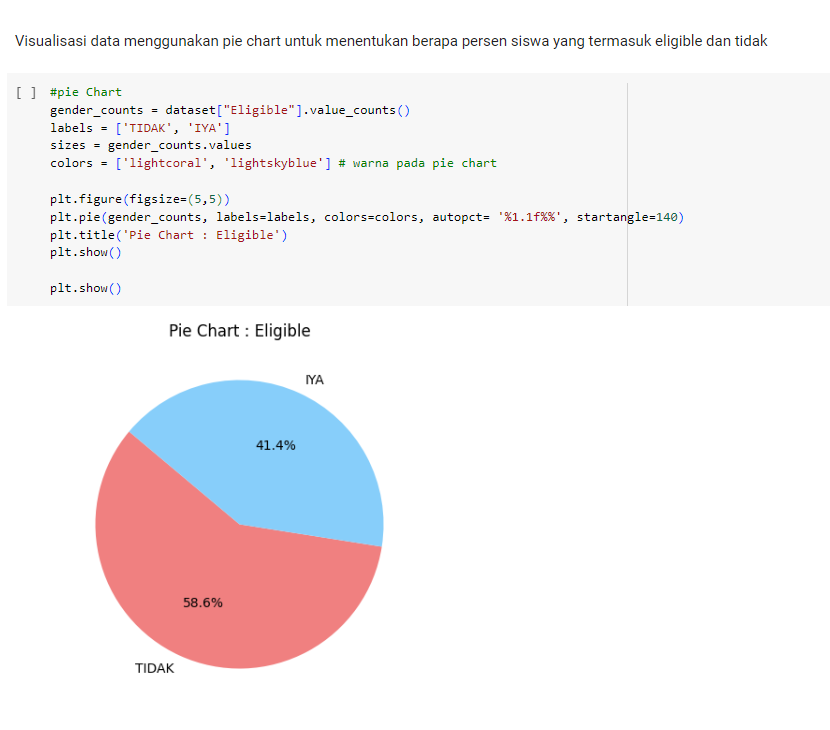


Gambar 4.4 Normalisasi Data

Terlihat bahwa kami mengubah nilai yang semula 0 - 100 menjadi 0 - 1.Ini guna untuk membantu algoritma Machine Learning yang rentang terhadap perbedaan skala antar fitur, seperti K- Means atau regresi linier.

## Visualisasi Data

Visualisasi data merupakan representasi informasi secara grafis atau diagramatik untuk mempermudah pemahaman pola, tren, dan hubungan dalam suatu dataset. Disini kami menggunakan salah satu model visualisasi data yang umum digunakan yaitu pie chart, dan count plot.



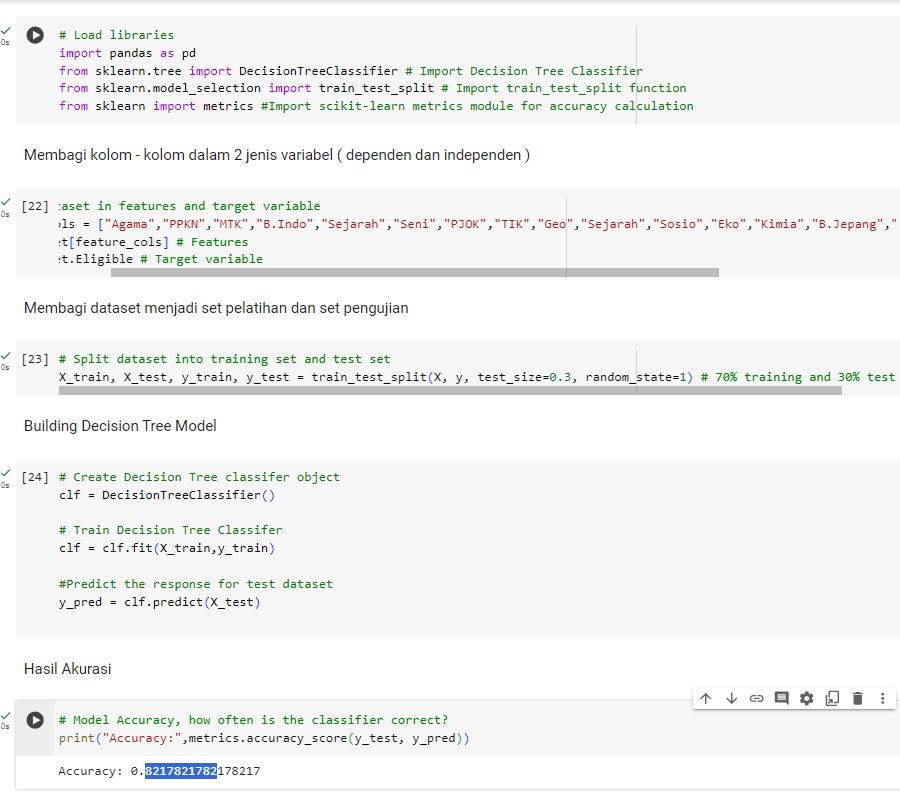
Gambar 4.3 Visualisasi Data

Kami menggunakan Pie Chart untuk visualisasi agar terlihat lebih mudah dipahami karena bentuk lingkaran yang intuitif, cocok untuk menggambarkan sebagian kecil dari data kategori yang signifikan, terutama ada satu atau dua data yang mendominasi.Terlihat di gambar bahwa data menunjukkan bahwa 41.4% menunjukkan label “iya” sedangkan sisanya yaitu 58.6% menunjukkan label “tidak”.

## Evaluasi Data

Evaluasi data adalah proses penilaian dan analisis terhadap kualitas, keterpercayaan, dan kinerja suatu model atau sistem berbasis data. Tujuannya adalah untuk mengukur sejauh mana model tersebut dapat memberikan hasil akurat, relevan, dan sesuai dengan kebutuhan atau tujuan yang diinginkan.

Mengukur akurasi adalah salah satu metrik evaluasi yang paling sederhana dan umum digunakan dalam Machine Learning. Ini mengukur sejauh mana model memberikan prediksi yang benar dibandingkan dengan total prediksi yang dilakukan.



Gambar 4.5 Evaluasi Data

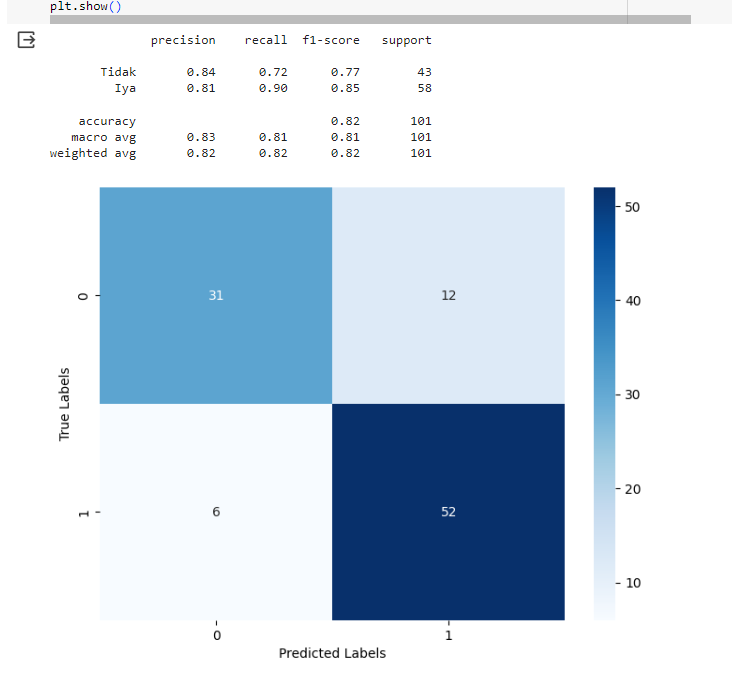
Kami menggunakan Algoritma Decision Tree untuk mengevaluasi model sehingga akurasi dapat dihitung dengan membandingkan nilai set pengujian aktual dan nilai prediksi.Sehingga diperoleh tingkat klasifikasi sebesar 82.17821782178217 %, yang dianggap akurasi yang baik.

## Uji Performa Model

Sete;ah melakukan modelling menggunakan Decision Tree , kami mencoba melakukan pengujian performa model . Uji performa model melibatkan evaluasi sejauh mana model dapat bekerja dengan baik pada data yang tidak digunakan selama pelatihan, dengan tujuan mengukur kinerja dan kemampuan generalisasi model. Tujuannuya untuk Menilai seberapa akurat, konsisten, dan dapat diandalkan model saat memprediksi atau mengklasifikasikan data baru. Untuk pengujian performa model, kami menggunakan Confusion Matrix. Confusion Matrix adalah sebuah visualisasi berbentuk tabel yang digunakan untuk mengevaluasi performa dari model klasifikasi yang kita pakai untuk mengolah sebuah data.

Confusion Matrix pada dasarnya seperti tabel perbandingan atau class pada dataset antara data aktual dan data prediksi. Confusion Matrix ini memiliki 4 nilai yaitu :

1. True Positive (TP)
2. True Negative (NP)
3. False Positive (FP)
4. False Negative (FN)

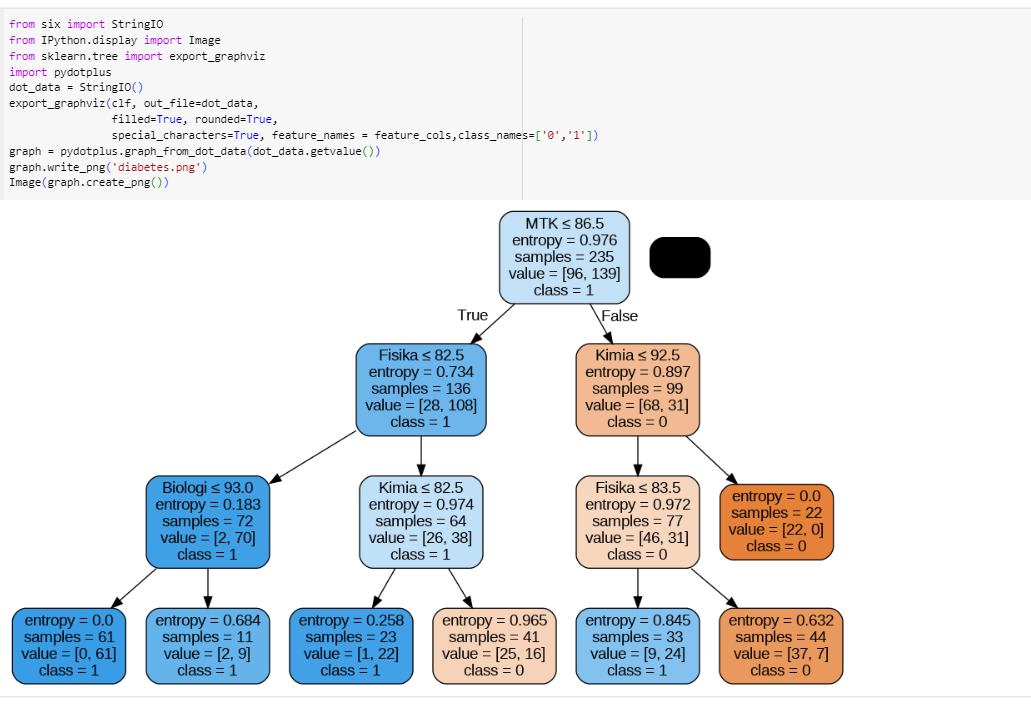


Gambar 4.6 Confusion Matrix

Terlihat di gambar bahwa siswa yang memiliki prediksi True Positive (TP) atau diprediksi positif dan dia memang positif sebanyak 52 siswa, True Negative (TN) atau diprediksi negatif tetapi dia positif sebanyak 31 siswa, False Positive (FP) atau diprediksi positif tetapi dia negatif sebanyak 6 siswa, dan False Negative (NP) atau diprediksi negatif dan dia negatif sebanyak 12 siswa.

## Memvisualisasikan Decision Tree

Visualisasi Decision Tree adalah representasi grafis dari struktur Decision Tree yang dihasilkan oleh model Machine Learning. Visualisasi Model ini juga memudahkan kita memahami tentang aturan dan logika keputusan yang dihasilkan oleh model Decision Tree. Bukan hanya itu visualisasi ini juga membantu kita dalam mengidentifikasi dan memvalidasi model atau pola tertentu yang guna untuk pembentukan model.



Gambar 4.7 Memvisualisasikan Decision Tree\

Dalam bagan Decision Tree, setiap node internal memiliki aturan keputusan yang membagi data. Entropy adalah ukuran ketidakpastian atau kekacauan dalam suatu himpunan data, samples mengacu pada jumlah observasi atau instansi dalam suatu simpul atau cabang dalam Decision Tree, value adalah distribusi kelas atau nilai target pada suatu simpul atau cabang, dan class adalah nilai target yang dihasilkan pada suatu simpul atau cabang tertentu dalam Decision Tree.

# BAB V KESIMPULAN DAN SARAN

Penelitian ini memberi hasil bahwa menggunkan konsep klasifikasi data mining diterapkan dalam aktifitas pengelompokan data dalam bidang pendidikan mampu memberikan kemampuan memprediksi yang akurat dan dapat dipercaya. Dalam penelitian ini menggunakan Confusion Matrix dengan menggunakan itu kita bisa mempermudah dalam melihat hasil pengelompokan data yang sudah kita buat dengan menggunakan klasifikasi tadi. Perumusan praktis yang dapat disumbangkan bagi kegiatan dalam bidang pengolahan data yang terdapat label pada datasetnya. Pada akhirnya perumusan ini aladalah untuk mempermudah dalam pengelompokan data yang akurat dan dapat dipercaya dan dalam waktu yang cepat.