INFOTECH journal
ISSN : 2460-1861 (Print), 2615-4250 (Online)
Vol. 11 No. 2 Desember 2025, pp. 234-241

PENERAPAN DATA MINING UNTUK MENENTUKAN KELAYAKAN KENDARAAN SEPEDA MOTOR BEKAS MENGGUNAKAN ALGORITMA C4.5

Ni Kadek Juliani[1], Ni Made Satvika Iswari[2], Nengah Widya Utami[3]
[1,2]Program Studi Informatika, Fakultas Tekonologi Informasi Dan Desain, Universitas Primakara
[3]Program Studi Sistem Informasi, Fakultas Teknologi Informasi Dana Desain, Universitas Primakara
email: juliani2002.id@gmail.com

ABSTRACT
Determining the feasibility of used motorcycles is one of the challenges for companies in selecting attributes that cover various factors, such as physical condition, maintenance history, and reasonable price. In this study, the researcher aims to analyze the existing problems and provide decision results by applying the C4.5 algorithm to determine the feasibility of used motorcycles based on relevant data. The C4.5 algorithm has the capability to build decision trees to automate and improve the accuracy of the feasibility determination process. In this research, attributes such as motorcycle model, year, engine, kilometers, fuel type, modifications, engine overhaul, oil type, transmission, engine type, and displacement are used as determining variables.
Furthermore, to avoid overfitting that may occur due to overly complex decision trees, the researcher also applies pruning techniques to the C4.5 algorithm. Pruning functions to trim insignificant branches of the tree so that the model becomes simpler. With pruning, it is expected that the resulting decision tree will be not only accurate but also efficient, enabling the feasibility determination process of used motorcycles to be conducted optimally. Therefore, after applying pruning techniques, the model achieved an accuracy of 72.41%, precision of 68.42%, recall of 86.67%, and F1-score of 76.47%.

Keywords: Motorcycle, Data Mining, Feasibility, C4.5 Algorithm, Decision Tree, Pruning.

ABSTRAK
Dalam menentukan suatu kelayakan kendaraan sepeda motor bekas menjadi salah satu tantangan bagi perusahaan untuk menentukan atribut yang mencakup berbagai faktor, seperti kondisi fisik, riwayat perawatan serta harga yang wajar. Dalam penelitian ini, Algoritma C4.5 memiliki kemampuan dalam membangun pohon keputusan untuk mengotomatisasi dan meningkatkan akurasi proses dalam penentuan kelayakan. Dalam penelitian ini, atribut seperti model motor, tahun, mesin, kilometer, bahan bakar, modifikasi, turun mesin, jenis oli, tranmisi, jenis mesin, kubikasi digunakan sebagai variable penentu. Selain itu, untuk menghindari overfitting yang dapat terjadi akibat pohon keputusan yang terlalu kompleks, peneliti juga menerapkan teknik prepruning pada algoritma C4.5. Prepruning berfungsi untuk memangkas cabang – cabang pohon yang kurang signifikan sehinggal model menjadi lebih sederhana. Dengan adanya prepruning, diharapkan pohon keputusan yang dihasilkan tidak hanya akurat tetapi efisien, sehingga proses penentuan kelayakan sepeda motor bekas dapat dilakukan secara optimal. Oleh karena itu, setelah menggunakan teknik prepruning mendapatkan akurasi 72,41%, precision 68,42%, recall 86.67% dan F1-score 76,47%.

Kata Kunci: Sepeda Motor, Data Mining, Kelayakan, Algoritma C4.5, Pohon Keputusan, Prepruning.

Riwayat Artikel :
Tanggal diterima : 25-06-2025
Tanggal revisi : 18-07-2025
Tanggal terbit : 05-08-2025

DOI :
https://doi.org/10.31949/infotech.v11i2.15338

INFOTECH journal by Informatika UNMA is licensed under CC BY-SA 4.0
Copyright © 2025 By Author

1. PENDAHULUAN

1.1. Latar Belakang

Dalam usaha otomotif, salah satunya masalah yang di hadapi khususnya dalam hal penjualan sepeda motor bekas adalah menentukan seberapa layak sepeda motor bekas untuk dibeli oleh pelanggan. Sepeda motor menjadi kebutuhan primer dan menggunakan sepeda motor mempermudah dan mempercepat setiap pekerjaan manusia yang membuat waktu, biaya, dan tenaga lebih efisien. Kualitas sepeda motor baru dan bekas tentunya berbeda, konsumen yang memilih sepeda motor bekas menghadapi beberapa masalah dalam menentukan apakah sepeda motor tersebut layak untuk dibeli atau tidak(Wibowo, 2013a). Faktor-faktor seperti kondisi mesin dan harga yang wajar menentukan kelayakan sepeda motor bekas. Ini penting untuk memastikan kepuasan pelanggan dan kepercayaan penjual (Madiono & Susanto, 2013)

Jaya Motor adalah penjualan sepeda motor belas yang berfokus pada transaksi jual beli motor second dari aneka merek. Pembelian motor bekas disebabkan oleh daya beli masyarakat yang mendorong ide bisnis jual beli motor bekas. Meskipun demikian, pelanggan sangat berhati-hati saat membeli sepeda motor bekas karena perlu mengetahui kondisi motor secara menyeluruh termasuk mesin dan jarak tempuh. Jaya Motor juga sering menemukan sepeda motor bekas yang tidak layak jual, beberapa faktor membentuk kelayakan itu, seperti kondisi mesin, Jarak tempuh dan sebagainya. Showroom akan dirugikan jika pemiliknya salah memprediksi kelayakan sepeda motor bekas, karena pemilik perlu memehami besaran biaya yang harus dikeluarkan untuk memperbaiki sepeda motor tersebut, (Hardiyan & Nadiyah, 2022a). Untuk menyelesaikan masalah ini, peneliti menggunakan metode data mining dengan algoritma pohon keputusan C4.5. Algoritma ini berperan dalam menghitung nilai entropi dan information gain untuk menentukan pemisah terbaik dalam membentuk pohon keputusan yang akan membantu dalam pengambilan keputusan kelayakan sepeda motor motor (Pratiwi et al., 2022). Namun, pohon keputusan yang dihasilkan dari algoritma C4.5 bisa menjadi berisiko overfitting ketika model terlalu detail mengikuti data pelatihan sehingga kurang mampu menyamaratakan data baru.

Untuk menyelesaikan permasalahan yang ada. Oleh karena itu, peneliti melakukan proses prepruning yang diterapkan untuk memangkas cabang – cabang pohon yang kurang signifikan atau berpotensi menyebabkan overfitting. Dalam hal ini, prepruning membantu dalam menyederhanakan model pohon keputusan sehingga menjadi lebih efektif dan akurat dalam memprediksi kelayakan motor bekas. Algoritma C4.5 yang dilengkapi dengan teknik prepruning diharapkan dapat menghasilkan model klasifikasi yang optimal untuk menentukan kelayakan sepeda motor bekas.

Ada beberapa beberapa algoritma yang digunakan dalam data mining, khususnya algoritma pohon keputusan C4.5 yang menggunakan entropi dan pendapatan infromasi sebagai pemisah dalam tree keputusan. Proses ini menghasilkan suatu simpul yang tidak dapat di pecah lagi. Node tersebut proses perhitungan numerik nilai normalisasi setiap atributnya yang diterapkan pada data untuk mencari nilai pada pohon keputusan C4.5. Dalam algoritma pohon keputusan, node yang memperoleh informasi dari sejumlah data yang sesuai untuk membantu dalam pengambilan keputusan (Pradana, 2018).

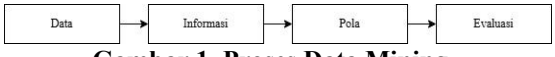
Adapun penelitian sebelumnya mengenai penelitian oleh (Egih Sugiatha, Anton Maulana Ibrahim, Ichwansyah Abdul Hadi) dengan judul "Implementasi Algoritma Klasifikasi C4.5 Untuk Memprediksi Kelayakan Pembelian Kendaraan" dengan hasil pemilihan kendaraan sepeda motor bekas menggunakan metode *decision tree* serta algoritma C4.5 dengan tingkat keakuratan menggunakan metode *confusion matrix* dengan *software* WEKA memiliki tingkat akurasi 92.4%, Precision 92.4% dan Recall 92.4%. Penelitian oleh (Moch Zacharia Azra, Intan Yuniar Purbasari, Basuki Rahmat) dengan judul "Penerapan Algoritma C4.5 Untuk Grading Kualitas Motor Bekas Di UD.Permata Motor" dengan hasil menangani kasus klasifikasi *multiclass* menggunakan algoritma C4.5 dengan hasil uji pengukuran akurasi, presisi dan *recall* yang mencapai 100% (Azra et al., 2021). Dalam penelitian terdahulu menjadi salah satu tumpuan yang baik bagi peneliti dalam melakukan penelitian sehingga dapat mengkaji penelitian yang dilakukan.

Berdasarkan penjelasan diatas, peneliti berniat untuk melaksanakan penelitian dengan judul "Penerapan Data Mining untuk Menentukan Kelayakan Kendaraan Sepeda Motor Bekas Menggunakan Algoritma C4.5" bertujuan untuk menggali lebih dalam informasi yang dapat mendukung peneliti untuk membuat keputusan akhir tentang masalah ini.

1.2. Tinjuan Pustaka

Dalam penelitian in terdapat beberapa landasan teori yang digunakan yaitu terkait Data Mining, Dataset, Tree Keputusan, Algoritma C4.5, Pemilihan Atribut, (Information Grain & Gain Ratio), Confusion Matrix, Kelayakan, Klasifikasi, Rapid Miner dan Google Collab.

Secara teknis, data mining dapat diartikan sebagai proses yang memanfaatkan teknik – teknik stastik, matematika dan kecerdasan buatan untuk mengesktrak dan mengidentifikasi informasi atau *knowledge* selanjutnya (Zai, 2022).



Gambar 1. Proses Data Mining

Pada proses gambar 1, proses pertama dilakukan adalah menyiapkan data mentah. Setelah

pemrosesan, pada tahap konversi, data dipilih dan diformat sesuai dengan teknologi atau metode yang digunakan. Informasi yang dihasilkan kemudian akan diproses menggunakan model untuk menandai pola tertentu. Dari keberhasilan mengidentifikasi pola-pola ini, Anda akan memperoleh pengetahuan yang berguna untuk evaluasi ketika mengambil keputusan kebijakan (Sulianta, 2023).

Dataset dalam konteks data mining dapat diartikan sebagai kumpulan data yang terdiri dari objek-objek yang mempunyai sifat atau variabel tertentu. Data dapat dalam berupa sebuah bahan untuk diskusi, pengambilan keputusan, perhitungan atau pengukuran dan data tidak selalu dalam bentuk kumpulan huruf dalam bentuk kata atau kalimat tapi data juga dapat dalam bentuk suara, gambar diam dan bergerak, baik dalam bentuk dua atau tiga dimensi (Sukriadi et al., 2023).

Tree Keputusan merupakan struktur yang berguna dalam memisahkan data ke beberapa kelompok lebih kecil menggunakan aturan keputusan yang telah ditentukan. Dengan kata lain, pohon keputusan terdiri dari tiga tahap penting. Pertama memilih atribut sebagai akarnya, kedua membuat cabang untuk setiap nilai yang mungkin. Ketiga membagi kasus ke dalam cabang yang memiliki kelas yang serupa (Setio et al. , 2020a).

Algoritma pohon keputusan merupakan salah satu dari metode yang digunakan untuk menganalisis data dengan C4.5 sebagai algoritma pohon keputusan yang banyak di pakai dalam data mining untuk membangun struktur pohon keputusan (Adriansa et al., 2022).

Pemilihan atribut adalah suatu metode dalam menentukan kriteria pemisahan yang optimal dalam memisahkan data dari atribut yang sudah diberikan label kelas menjadi kelas individu (Purbasari et al., 2013). Metrik pemilihan atribut juga sering dipakai sebagai aturan pemisahan karena mereka menentukan untuk memisahkan atribut suatu node tertentu (Setio et al., 2020).

1. *Information Gain*
   mengukur mengukur pengurangan ketidakpastian mengenai kelas target (label) yang dihasilkan ketika nilai atribut diketahui. *Information Gain* melibatkan beberapa tahapan yaitu:
   Perhitungan Entropi/Info(D) untuk ukuran ketidakpastian atau ketidakmurnian dalam suatu dataset. Entropi ini dapat dihitung menggunakan rumus:
   $$Entropy(D) = -\sum_{i=1}^{n} Pi * \log_2 (pi)$$
2. *Gain Ratio*
   mengatasi masalah ini dengan membagi *Information Gain* oleh *Split Information. Split Information* mengukur seberapa baik atribut data tanpa mempertimbangkan label Gain Ratio yang memberikan bobot tambahan dalam memastikan bahwa atribut dengan banyak nilai tidak mendominasi pemilihan atribut. Terdapat beberapa tahapan dalam perhitungan Gain Ratio (Eska, 2018).
   a. *Split Information* adalah proses dalam perhitungan nilai yang mempresentasikan potensi informasi yang dihasilkan dengan membagi dataset pelatihan D menjadi V berdasarkan atribut A. Perhitungan ini dapat di hitung menggunakan rumus:
   $$SplitInfo_A(D) = -\sum_{j=i}^{v} \frac{|Dj|}{|D|} x\ log_2 \left(\frac{|Dj|}{|D|}\right)$$
   Setiap bagian Dj diukur dengan mempertimbangkan jumlah tuple dalam bagian tersebut terhadap jumlah total tuple dalam D.
   b. Gain Ratio adalah metrik yang digunakan untuk pemilihan atribut dalam induksi pohon keputusan dan hasil perhitungan dari *Information Gain.* Perhitungan ini dapat dihitung menggunakan rumus:
   $$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)}$$

Atribut dengan gain ratio yang tinggi dipilih sebagai atribut pemisahan. Confusion matrix dapat diartikan sebagai suatu teknik diterapkan dalam analisis prediktif pada *machine learning.*

Confusion matrix pengukuran suatu performa untuk masalah klasifikasi yang dimana terdapat dua kelas atau lebih (Dinata et al., 2020). Dengan memvisualisasikan prediksi model dan membandingkannya dengan data aktual dan memberikan informasi mengenai klasifikasi benar dan salah.

Tabel 1. Informasi Klasifikasi

Klasifikasi		Kelas Prediksi	
		Positif	Negatif
Kelas Sebenarnya	Positif	TP	FN
	Negatif	FP	TN

Sesuai dengan parameter tersebut selanjutnya digunakan dalam rumus untuk menghitung metrik kinerja seperti akurasi (*accuracy*), presisi (*precision*), *recall*, dan hasil F1.

a. *Accuracy* adalah ukuran ketepatan yang dihitung dengan membagi jumlah data yang dipediksi positif dan yang sebenarnya positif, serta data yang diprediksi negatif dan benar – benar negatif dengan total jumlah data dalam dataset.
$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

b. *Precision* merupakan hasil akurasi dari jumlah daya yang bernilai positif dari keseluruhan data hasil diprediksi positif.
$$Precision = \frac{TP}{TP + FP}$$

c. *Recall* merupakan hasil akurasi dari jumlah yang bernilai positif dari keseluruhan data hasil yang benar positif.

$$Recall = \frac{TP}{TP + FN}$$

d. F1 merupakan hasil ukur nilai rata rata presisi dan *recall*.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Kelayakan merupakan penelitian mendalam yang dilakukan untuk menilai suatu proyek yang direncanakan dan dapat memberikan hasil sesuai dengan harapan. Dalam konteks ini penilaian layak berarti sejauh mana suatu suatu ide usaha untuk memberikan manfaat, baik dalam bentuk keuntungan finansial maupun manfaat sosial (Subagyo, 2007).

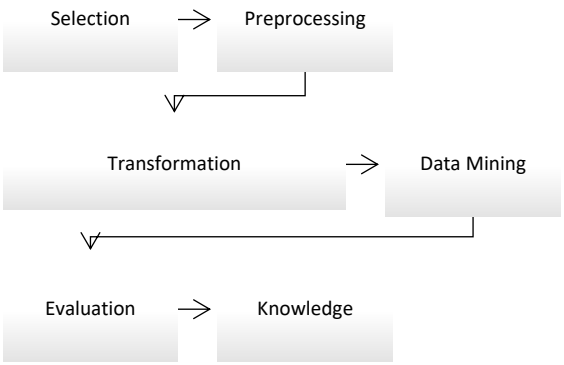
Klasifikasi merupakan salah satu metode dalam data mining yang digunakan untuk mengelompokkan sdata ke dalam kategori atau kelas tertentu. Fungsinya adalah untuk meramalkan kategori di mana suatu objek akan termasuk jika kelasnya belum diketahui, metode ini bisa disebut sebagai pembelajaran termonitor (Iswari et al., 2024).

Rapid Miner merupakan sebuah platform perangkat lunak yang digunakan untuk menganalisis dan mengolah data. Platform ini dilengkapi dengan berbagai alat dan fitur yang akan membantu dalam menjalankan proses eksplorasi data, pemodelan, dan visualisasi (Sinaga et al., 2021).

Google Colab atau disebut Google Colaboratory adalah sebuah lingkungan yang pengembangannya berbasis cloud untuk menulis dan menjalankan kode Phyton melalui browser web. Google colab dapat digunakan bersama sama oleh pengembang aplikasi, sehingga dapat mendukung kebutuhan kolaborasi antar anggota(Febby Wilyani et al., 2024).

1.3. Metodologi Penelitian

Metode yang diterapkan dalam penelitian ini adalah *Knowledge Discovery in Database* yang merupakan proses melibatkan pengumpulan dan pemanfaatan data historis untuk mengidentifikasi pola atau hubungan dalam dataset. KDD tidak hanya berfokus pada analisis data, tetapi juga mencakup transformasi data yang menjadi pengetahuan yang dapat ditarapkan. Proses KDD yang sistematis dan terstruktur memastikan pengetahuan yang dihasilkan adalah valid dan dapat diterapkan dalam konteks yang relavan (Mardi, 2017).



Gambar 2. Proses Knowledge Discovery In Database Https://Shorturl.At/Taudj

2. PEMBAHASAN

2.1. Data Selection

Pada tahap selection, permasalahan yang diidentifikasi adalah minimnya kemampuan dalam menganalisis standar kelayakan sepeda motor bekas, yang menyebabkan ketidakmampuan menentukan klasifikasi kelayakan secara langsung saat pemeriksaan manual yang bersifat subjektif. Penelitian ini bertujuan untuk menganalisis dan mengimplementasikan algoritma C4.5 dalam menentukan kelayakan sepeda motor bekas menggunakan Google Colab, melalui tahapan Data Selection, Preprocessing, Transformation, Data Mining, dan Evaluation. Pada tahap Data Selection, dilakukan pengumpulan dan analisis kualitas data, dimulai dengan menghubungkan Google Colab ke Google Drive agar dataset dapat diakses, dilanjutkan dengan import library *pandas* untuk pengolahan data, serta mencantumkan dataset ke Google Colab yang hasil tampilannya ditunjukkan pada Gambar 3.

```
==========================================================================================
 Informasi data 5 teratas
==========================================================================================
  ModelMotor  Tahun  Mesin  Kilo_Jaraktempuh BahanBakar Modifikasi TurunMesin  \
0      Honda   2022  Halus            354090   Pertamax     Ringan      Tidak
1      Honda   2014  Kasar            377020  Pertalite     Sedang        Iya
2      Honda   2010  Halus            325890  Pertalite     Sedang        Iya
3      Honda   2009  Kasar            557342  Pertalite     Sedang      Tidak
4      Honda   2012  Halus            478920  Pertalite      Berat        Iya

           JenisOli  Tranmisi JenisMesin  Kubikasi     Keputusan
0  Federal Ultratec  Otomatis      4-Tak       160  Layak dijual
1  Federal Ultratec  Otomatis      4-Tak       108   Tidak layak
2  Federal Ultratec  Otomatis      4-Tak       108   Tidak layak
3  Federal Ultratec  Otomatis      4-Tak       108   Tidak layak
4    Yamalube Matic  Otomatis      4-Tak       125   Tidak layak
-------------------------------------------------------------
```

Gambar 3. Tampilan Data

```
-------------------------------------------------------------
No.   Kolom                          Non-Null     Tipe Data
=============================================================
1     ModelMotor                     145          object
2     Tahun                          145          int64
3     Mesin                          145          object
4     Kilo_Jaraktempuh               145          int64
5     BahanBakar                     145          object
6     Modifikasi                     145          object
7     TurunMesin                     145          object
8     JenisOli                       145          object
9     Tranmisi                       145          object
10    JenisMesin                     145          object
11    Kubikasi                       145          int64
12    Keputusan                      145          object
=============================================================
```

Gambar 4. Tampilan Informasi Awal Dataset

Gambar 4. menunjukkan bahwa dataset terdiri dari 145 data dengan 12 atribut, yakni ModelMotor, Tahun, Mesin, Kilo_Jaraktempuh, BahanBakar, Modifikasi, TurunMesin, JenisOli, Tranmisi, JenisMesin, Kubikasi, dan Keputusan. Atribut Tahun, Kilo_Jaraktempuh, dan Kubikasi bertipe numerik, sementara sisanya bertipe kategorikal.

```
Saving Perhitungan_C45.csv to Perhitungan_C45 (2).csv
Kolom-kolom dataset: Index(['ModelMotor', 'Tahun', 'Mesin', 'Kilo_Jaraktempuh', 'BahanBakar', 'Modifikasi', 'TurunMesin', 'JenisOli', 'Tranmisi', 'JenisMesin', 'Kubikasi', 'Keputusan', 'Unnamed: 12', 'Unnamed: 13', 'Unnamed: 14', 'Unnamed: 15', 'Unnamed: 16', 'Unnamed: 17', 'Unnamed: 18', 'Unnamed: 19', 'Unnamed: 20', 'Unnamed: 21', 'Unnamed: 22', 'Unnamed: 23'], dtype='object')
```

Gambar 5. Tampilan File CSV Atau Excel

Gambar 5. menunjukkan hasil pembacaan file CSV ke dalam DataFrame, di mana seluruh atribut terdeteksi dengan benar. Namun, muncul kolom tambahan 'Unnamed: 12' yang dihasilkan secara otomatis oleh pandas akibat adanya kolom kosong dalam file CSV, menandakan bahwa data masih perlu dibersihkan pada tahap selanjutnya.

2.2. Preprocessing

```
==========================================
 Informasi Data Frame
==========================================
 Jumlah Data      : 145
 Jumlah Variabel : 12
```

Gambar 6. Informasi Dataframe

Melalui Gambar 6, dapat disimpulkan bahwa pada jumlah data “145” menunjukkan jumlah baris yang tersisa setelah proses pembersihan data dan jika data yang sebelumnya memiliki banyak baris, maka angka tersebut menandakan bahwa baris yang memiliki nilai kosong telah dihapus. Gambar ini menampilkan jumlah baris setelah proses pembersihan awal dan menunjukkan bahwa data duplikat atau baris kosong telah dihapus. Ini penting untuk mendemonstrasikan efektivitas tahap awal *preprocessing*.

```
------------------------------------------------------------
No.  Kolom                          Non-Null     Tipe Data
============================================================
1    ModelMotor                     145          object
2    Tahun                          145          int64
3    Mesin                          145          object
4    Kilo_Jaraktempuh               145          int64
5    BahanBakar                     145          object
6    Modifikasi                     145          object
7    TurunMesin                     145          object
8    JenisOli                       145          object
9    Tranmisi                       145          object
10   JenisMesin                     145          object
11   Kubikasi                       145          int64
12   Keputusan                      145          object
============================================================
```

Gambar 7. Konversi Tipe Data Kategorikal Dan Numerik

Melalui Gambar 7, dapat disimpulkan kolom yang bertipe int64 sudah dikonversi menjadi anga bulat (integer) kolom yang bertipe int64 tersebut ada pada atribut Tahun, Kilo_Jaraktempuh, dan Kubikasi. Gambar ini memperlihatkan hasil dari konversi tipe data yang esensial dalam *preprocessing*, seperti konversi dari kategorikal ke numerik dan pengisian nilai kosong. Ini menjadi dasar penting sebelum data diproses oleh algoritma C4.5.

	Model Motor	Mesin	Bahan Bakar	Modifikasi	Turun Mesin	Jenis Oli	Transmisi	Jenis Mesin	Keputusan
count	145	145	145	145	145	145	145	145	145
unique	2	2	2	3	2	2	2	1	2
top	Honda	Halus	Pertalite	Sedang	Iya	Yamalube Matic	Otomatis	4-Tak	Tidak layak
freq	108	75	106	74	77	90	128	145	85

Gambar 8. Statistik Kolom Kategorikal

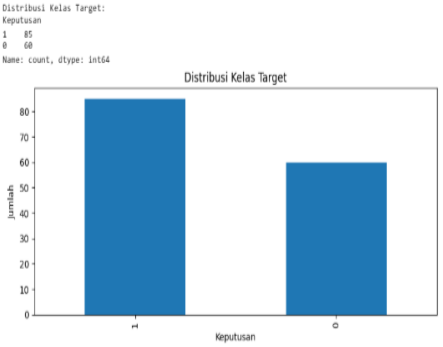
Melalui Gambar 8, dapat disimpulkan gambar diatas menampilkan statistik deskriptif untuk kolom kategorikal seperti ModelMotor, Mesin, BahanBakar, Modifikasi, TurunMesin, JenisOli, Tranmisi, JenisMesin dan Keputusan. Statistik deskriptif pada data kategorikal memberi gambaran umum tentang distribusi data, jumlah kategori unik, dan dominasi nilai tertentu seperti “Layak” atau “Tidak Layak”. Ini berguna dalam memahami karakteristik dataset sebelum klasifikasi.

2.3. Transformation (Transformasi Data)

```
0      0
1      1
2      1
3      1
4      1
      ..
140    1
141    1
142    1
143    0
144    0
Name: Keputusan, Length: 145, dtype: int64
   Keputusan
0          0
1          1
2          1
3          1
4          1
['Layak dijual' 'Tidak layak']
Output Label Encoding pada Kolom Keputusan
```

Gambar 9. Label Encoding

Gambar 9 menunjukkan hasil label encoding pada kolom 'Keputusan', di mana nilai kategorikal telah berhasil diubah menjadi angka, menandakan data siap untuk diproses lebih lanjut.



Gambar 10. Distribusi Kelas Target

Gambar 10 menunjukkan distribusi data pada kolom target setelah encoding, dengan kelas 1 sebanyak 85 dan kelas 0 sebanyak 60. Distribusi ini menunjukkan ketidakseimbangan data, yang dapat memengaruhi

performa model karena cenderung lebih fokus pada kelas yang jumlahnya lebih banyak.

```
Jumlah sampel training: 116
Jumlah sampel testing: 29
```

Gambar 11. Sampel Jumlah Training Dan Testing

Melalui gambar 11, dapat disimpulkan bahwa jumlah sampel training sebanyak 116 yang dimana data digunakan untuk training model decision tree dan jumlah sampel testing sebanyak 29 yang dimana data digunakan untuk testing performa model setelah ditraining.

2.4. Data Mining

```
Fitur dengan Gain Ratio tertinggi: Kilo_Jaraktempuh (0.1363)
```

Gambar 12. Fitur Dengan Gain Ratio Tertinggi

Melalui Gambar 12, dapat disimpulkan bahwa fitur yang memiliki Gain Ratio tertinggi adalah atribut Kilo_Jaraktempuh dengan nilai 0.1363. Ini menjadikannya atribut paling informatif untuk pemisahan awal dalam pohon keputusan.

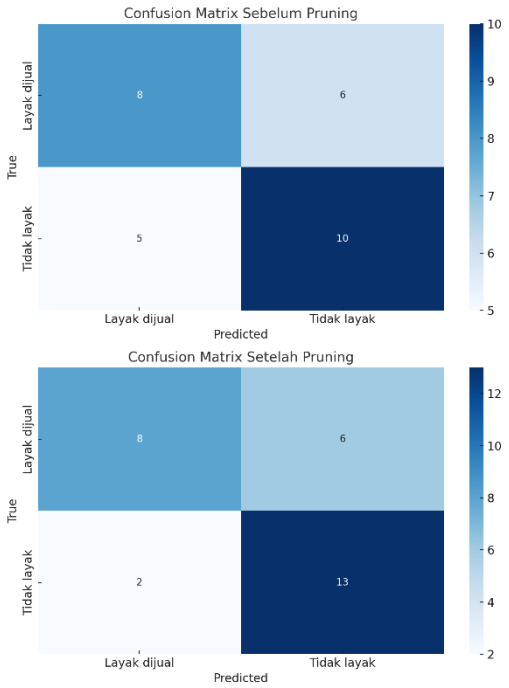
```
Model Setelah Pruning (CCP Alpha Optimal ):
Nilai Alpha Pruning : 0.0324
Akurasi pada Data Uji : 0.7241
```

Gambar 13. Model Setelah Pruning

Gambar 13 menunjukkan nilai alpha optimal sebesar 0.0324 yang menyederhanakan pohon keputusan dan meningkatkan akurasi dari 62% menjadi 72.41%. Pruning berhasil mengurangi overfitting dan meningkatkan generalisasi model.

Gambar 14. Perbandingan Akurasi Sebelum Dan Setelah Pruning

Melalui Gambar 14 diatas, dapat disimpulkan bahwa grafik dengan batang merah menunjukkan data sebelum pruning yang nilainya sekitar 0.62 (62%) dan grafik dengan batang hijau menunjukkan data setelah pruning yang nilainya sekitar 0.72 (72%).



Gambar 15. Confusion Matrix Sebelum Dan Setelah Pruning

Melalui Gambar 15 diatas, dapat disimpulkan bawah Accuracy meningkat setelah melakukan pruning, hal tersebut bertambahnya jumlah prediksi benar pada kategori “Tidak Layak” (False Positive) dari nilai sebelumnya 5 menjadi 2. Jumlah prediksi benar pada “Layak dijual” (True Positive) tetap sama dengan nilai 8 dan dengan prediksi salah pada “Layak dijual” (False Negative) dengan nilai 6. Model setelah pruning lebih optimal dalam mengidentifikasi data yang “Tidak layak”, sehingga mengurangi kesalahan False Positive.

```
Classification Report Sebelum Pruning
              precision    recall  f1-score   support

Layak dijual       0.62      0.57      0.59        14
 Tidak layak       0.62      0.67      0.65        15

    accuracy                           0.62        29
   macro avg       0.62      0.62      0.62        29
weighted avg       0.62      0.62      0.62        29

Classification Report Setelah Pruning
              precision    recall  f1-score   support

Layak dijual       0.80      0.57      0.67        14
 Tidak layak       0.68      0.87      0.76        15

    accuracy                           0.72        29
   macro avg       0.74      0.72      0.72        29
weighted avg       0.74      0.72      0.72        29
```

Gambar 16. Classification Report Sebelum Dan Setelah Pruning

Gambar 16 memperlihatkan peningkatan pada semua metrik evaluasi. Precision untuk “Layak dijual” meningkat dari 0.62 menjadi 0.80 dan Recall untuk “Tidak Layak” meningkat dari 0.67 menjadi 0.87. Ini membuktikan bahwa pruning berhasil meningkatkan performa klasifikasi.

2.5. Knowledge Representation

Tahap akhir dalam proses KDD adalah representasi pengetahuan, yaitu penyajian hasil mining dalam

bentuk yang mudah dipahami. Dalam studi ini, aturan kelayakan sepeda motor bekas direpresentasikan melalui diagram pohon keputusan dan sistem berbasis aturan (rule-based), guna memberikan penilaian yang objektif dan meminimalkan kesalahan.

Model hasil pruning disimpan dalam format .pkl menggunakan *joblib* untuk digunakan kembali tanpa pelatihan ulang (lihat Kode Program 4.37). Visualisasi pohon keputusan diekspor sebagai gambar .PNG dengan *export_graphviz* dan *pydotplus* (lihat Kode Program 4.38). Hasil akhir ditampilkan pada Gambar 17, yang menunjukkan keberhasilan penyimpanan model dan visualisasi pohon keputusan setelah proses pruning.

```
Model setelah pruning disimpan sebagai: decision_tree_after_pruning.pkl
Pohon keputusan setelah pruning disimpan sebagai: decision_tree_after_pruning.png
```

Gambar 17. Pemberitahuan Bahwa Model Dan Pohon Keputusan Disimpan

3. KESIMPULAN

Hasil penelitian dam implementasi sistem klasifikasi untuk menilai kelayakan sepeda motor bekas melalui algoritma C4.5 yang disesuaikan dengan melalui Knowledge in Database (KDD). Proses ini meliputi tahapan – tahapan KDD seperti seleksi data awal, pembersihan dan praproses data, transformasi data ke dalam format modeling, penerapan algoritma klasifikasi dan evaluasi performa model yang dihasilkan

Hasil evaluasi dari modeling dilakukannya teknik prepruning yang menetapkan bahwa model mampu mencapai akurasi 72,41% yang berarti dari seluruh data yang digunakan untuk diuji, lebih dari 7 dari sepuluh prediksi model adalah benar. Untuk nilai precision 68,42% yang mengindikasikan bahwa dari Sebagian besar sepeda motor yang diprediksi sebagai "layak" memang benar – benar layak. Sedangkan recall 86,67% menunjukkan bahwa kemampuan model untuk mendeteksi motor layak dengan baik. Nilai F1-score menunjukkan 76,47% yang keseimbangannya yang cukup baik antara precision dan recall. Oleh karena itu, algoritma C4.5 mampu mengklasifikasi kelayakan sepeda motor bekas dengan menggunakan teknik prepruning dan mendapatkan hasil yang cukup dan layak digunakan dalam proses pengambilan keputusan, terutama pada industri jual beli sepeda motor bekas.

Saran untuk pengembangan lebih lanjut, penulis ingin meningkatkan performa atau keandalan sistem seperti melakukan eskperimen dengan berbagai algoritma klasifikasi yang lain seperti Random Forest, Support Vector Machine (SVM). Dengan menggunakan perbandingan performa antar algoritma, peneliti dapat mengidentifikasi metode yang cocok digunakan pada dataset tertentu. Pendekatan berbasis pembelajaran (deep learning) guna untuk meningkatkan performa dan mengembangkannya menjadi aplikasi berbasis web sehingga mudah diakses oleh pengguna seperti pembeli motor.

Keterbatan penelitian ini mengenai ukuran dataset yang relatif kecil dan jumlah data yang terbatas memengaruhi kemampuan dalam model melakukan generalisasi terhadap data. Selain itu, dataset yang relatif kecil memungkinkan risiko overfitting, terutama pada model pohon keputusan yang kompleks. Oleh karena itu, hasil penelitian perlu diinterpretasikan dengan hati – hati dan disarankan untuk melakukan pengujian lebih lanjut menggunakan dataset yang lebih besar dan beragam agar model dapat memberikan hasil yang optimal.

PUSTAKA

Adriansa, M., Yulianti, L., & Elfianty, L. (2022). Analisis Kepuasan Pelanggan Menggunakan Algoritma C4. 5. *Jurnal Teknik Informatika UNIKA Santo Thomas*, 115–121.

Azra, M. Z., Purbasari, I. Y., & Rahmat, B. (2021). Penerapan Algoritma C4. 5 untuk Grading Kualitas Motor Bekas di UD. Permata Motor. *Jurnal Informatika Dan Sistem Informasi*, *2*(1), 93–98.

Dinata, R. K., Safwandi, S., Hasdyna, N., & Azizah, N. (2020). Analisis k-means clustering pada data sepeda motor. *INFORMAL: Informatics Journal*, *5*(1), 10–17.

Eska, J. (2018). *Penerapan data mining untuk prediksi penjualan wallpaper menggunakan algoritma C4. 5*.

Febby Wilyani, Qonaah Nuryan Arif, & Fitri Aslimar. (2024). Pengenalan Dasar Pemrograman Python Dengan Google Colaboratory. *Jurnal Pelayanan Dan Pengabdian Masyarakat Indonesia*, *3*(1), 08–14. https://doi.org/10.55606/jppmi.v3i1.1087

Madiono, L., & Susanto, A. (n.d.). Sistem Pendukung Keputusan Kelayakan Kredit Sepeda Motor Honda di Nusantara Sakti Cabang Sukun Semarang Dengan Metode Scoring System. *Techno. COM*, *12*(3), 168–174.

Mardi, Y. (2017). Data mining: Klasifikasi menggunakan algoritma c4. 5. *Jurnal Edik Informatika Penelitian Bidang Komputer Sains Dan Pendidikan Informatika*, *2*(2), 213–219.

Ni Made Satvika Iswari, Nunik Afriliana, Eddy Muntina Dharma, & Ni Putu Widya Yuniari. (2024). Enhancing Aspect-based Sentiment Analysis in Visitor Review using Semantic Similarity. *Journal of Applied Data Sciences*, *5*(2), 724–735.

Ni Wayan Oktha Pratiwi, Nengah Widya Utami, & I Gede Juliana Eka Putra. (2022). KLASIFIKASI PENENTUAN PENERIMA BANTUAN SOSIAL TUNAI (BST) MENGGUNAKAN ALGORITMA C4.5 DI DESA KERAMAS, GIANYAR BALI.

JINTEKS (Jurnal Informatika Teknologi Dan Sains), *4*(3), 101–107.

Pradana, T. (2018). SISTEM PENDUKUNG KEPUTUSAN KELAYAKAN DAN ANALISA RESIKOPEMBERIAN KREDIT SEPADA MOTOR DENGAN ALGORITMA C4. 5 PADA ADIRA FINANCE BANGIL. *SPIRIT*, *6*(1).

Purbasari, I. Y., Nugroho, B., & Implementasi, D. A. N. (2013). Benchmarking Algoritma Pemilihan Atribut Pada Klasifikasi Data Mining. *Snastia, 47â*, *54*.

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

Sinaga, T. H., Wanto, A., Gunawan, I., Sumarno, S., & Nasution, Z. M. (2021). Implementation of Data Mining Using C4. 5 Algorithm on Customer Satisfaction in Tirta Lihou PDAM. *Journal of Computer Networks, Architecture and High Performance Computing*, *3*(1), 9–20.

Subagyo, A. (2007). *Studi kelayakan*. Elex Media Komputindo.

Sukriadi, S., Ismail, I., & Andzar, A. M. (2023). Penerapan Text Mining Dalam Klasifikasi Judul Skripsi Yang Diusulkan Mahasiswa Menggunakan Metode Naïve Bayes. *Jurnal Ilmiah Sistem Informasi Dan Teknik Informatika (JISTI)*, *6*(2), 184–196.

Sulianta, F. (2023). *Basic Data Mining from A to Z*. Feri Sulianta.

Zai, C. (2022). Implementasi Data Mining Sebagai Pengolahan Data. *Jurnal Portal Data*, *2*(3).