Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 6 Iss. 1 January 2026, pp: 354-362
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Decision Tree Classification Using Equal Width and Logarithmic Binning for Sustainable Tourism Data

Klasifikasi Data Pariwisata Berkelanjutan Menggunakan *Decision Tree* dengan *Equal Width* dan *Logaritma Binning*

Fahri Alviansyah[1*], Mukti Adi Azhari[2], Attar Raihan Nazhif[3], Suharto[4], Denny Saryanto[5], Kusrini[6]

[1,2,3,4,5,6]Program Studi PJJ Magister Informatika, Fakultas Ilmu Komputer,
Universitas Amikom Yogyakarta, Indonesia

E-Mail: [1]f.alviansyah@students.amikom.ac.id, [2]muktigeb@gmail.com, [3]a.raihan@students.amikom.ac.id, [4]suharto@students.amikom.ac.id, [6]kusrini@amikom.ac.id

Received Nov 13th 2025; Revised Dec 09th 2025; Accepted Jan 11th 2026; Available Online Jan 31th 2026
Corresponding Author: Fahri Alviansyah

Abstract

Tourism recommendation data management requires accurate predictive modeling to classify recommendation targets based on tourist behavior. However, numerical variables such as Sustainable Score often exhibit skewed data distributions, which can degrade the performance of machine learning algorithms, particularly Decision Tree models. This study investigates the effectiveness of two discretization techniques applied during the preprocessing stage, namely Equal Width Binning (EWB) and Logarithmic Binning (LB), in improving classification performance. The proposed methodology includes handling missing values and extracting temporal features from travel date data. The dataset is then processed under two discretization scenarios Equal Width Binning and Logarithmic Binning before being modeled using the Decision Tree classification algorithm. Model performance is evaluated using Accuracy, Precision, Recall, and F1-Score metrics. Experimental results demonstrate that the Decision Tree model with Equal Width Binning achieves 82% accuracy, whereas Logarithmic Binning yields 90% accuracy. Moreover, Logarithmic Binning reduces tree depth and mitigates overfitting, resulting in a more robust and generalizable model for tourism recommendation prediction..

Keywords: Data Preprocessing, Decision Tree, Equal Width Binning, Logarithmic Binning

Abstrak

Pengelolaan data rekomendasi pariwisata memerlukan pemodelan prediktif yang akurat untuk mengklasifikasikan target rekomendasi berdasarkan perilaku wisatawan. Namun, variabel numerik seperti Rekomendasi Score seringkali memiliki distribusi data yang tidak merata (skewed), yang dapat memengaruhi performa algoritma pembelajaran mesin seperti Decision Tree. Penelitian ini bertujuan untuk membandingkan efektivitas dua teknik preprocessing dengan diskritisasi, yaitu *Equal Width Binning* (EWB) dan *Logarithmic Binning* (LB), dalam meningkatkan kinerja model klasifikasi. Metodologi penelitian ini mencakup beberapa tahapan preprocessing data, antara lain handling missing value, serta ekstraksi fitur temporal dari data tanggal perjalanan. Data kemudian diproses menggunakan dua skenario binning yang berbeda sebelum dilatih menggunakan algoritma Decision Tree. Hasil penelitian dievaluasi menggunakan metrik Akurasi, Presisi, Recall, dan F1-Score. Hasil perbandingan menunjukkan bahwa Equal Width Binning nilai akurasi sebesar 82 %, dan *Logarithmic* Binning memberikan nilai akurasi sebesar 90%. Diskritisasi melalui logaritma bining mampu mengurangi kedalaman pohon (tree depth) dan mencegah *overfitting*, sehingga menghasilkan model yang lebih tangguh dalam memprediksi target rekomendasi pariwisata.

Kata Kunci: *Decision Tree, Equal Width Binning, Logarithmic Binning,* Pra-Proses Data

1. PENDAHULUAN

Pariwisata jalan raya (*road tourism*) hal penting industri pariwisata global yang memberikan kontribusi luas terhadap pertumbuhan ekonomi regional dan pertukaran budaya. Sebagai modal perjalanan yang menawarkan fleksibilitas, penggunaan kendaraan pribadi, bus, atau sepeda motor memfasilitasi

konektivitas hingga ke daerah terpencil [1]. Namun, seiring meningkatnya kepedulian terhadap kelestarian lingkungan, terdapat urgensi untuk menyelaraskan pengembangan pariwisata dengan praktik ramah lingkungan guna meminimalkan jejak karbon dan kemacetan [1]. Tantangan utamanya adalah konsumsi sumber daya yang berlebihan dan polusi udara yang secara langsung berdampak pada perubahan iklim di destinasi wisata [2].

Pengembangan pariwisata berkelanjutan membutuhkan perencanaan rute cerdas yang mampu menyeimbangkan permintaan wisatawan dengan kapasitas infrastruktur. Sayangnya, metode tradisional dalam perencanaan rute dan peramalan sering kali gagal menangkap sifat dinamis dari preferensi pengunjung yang dipengaruhi oleh perilaku masa lalu, kondisi lalu lintas, dan faktor lingkungan yang kompleks. Selain itu, data yang dikumpulkan dari berbagai sumber sering kali memiliki masalah kualitas, seperti nilai yang hilang, derau statistik (*noise*), dan pencilan (*outliers*) yang dapat membiaskan hasil prediksi [1].

Penelitian ini bertujuan untuk mengatasi batasan tersebut dengan mengusulkan penggunaan model *Decision Tree* untuk memprediksi tren masa depan dalam pariwisata jalan raya yang berkelanjutan. *Decision Tree* dipilih karena kemampuannya dalam membagi dataset menjadi kelompok kelas yang terukur dan menyediakan logika pengambilan keputusan yang transparan dalam bentuk aturan "IF-THEN" [3]. Untuk memastikan keandalan model, penelitian ini menerapkan serangkaian tahap prapemrosesan data yang meliputi diskritisasi, handling missing values melalui teknik imputasi guna menjaga integritas analisis, serta encoding untuk mengubah data kategorikal menjadi format numerik yang dapat diproses algoritma [1].

Sebagai inovasi pada tahap diskritisasi, penelitian ini mengintegrasikan dua teknik binning: *Equal Width Binning* dan *Logarithmic Binning*. *Equal Width Binning* diterapkan untuk mendiskritisasi variabel kontinu menjadi interval kategori yang seragam guna meningkatkan akurasi algoritma klasifikasi [3]. Sementara itu, *Logarithmic Binning* digunakan secara khusus untuk menangani distribusi data yang memiliki *noise* tinggi [4]. Penggunaan binning logaritmik sangat krusial untuk mengungkap tren tersembunyi yang tidak terlihat pada data mentah dan memberikan estimasi yang lebih akurat terhadap fenomena berskala besar, seperti lonjakan emisi karbon atau kepadatan kendaraan yang ekstrem [5]. Melalui integrasi teknik ini, penelitian diharapkan dapat menghasilkan wawasan strategis yang lebih robust bagi pengambil kebijakan dalam mengoptimalkan efisiensi jaringan jalan demi pariwisata yang bertanggung jawab secara lingkungan.

Penelitian sebelumnya menunjukkan bahwa tahapan diskritisasi atau binning merupakan langkah prapemrosesan data yang sangat penting untuk meningkatkan hasil klasifikasi pada algoritma *Decision Tree* secara signifikan [3], [6]. Banyak metode data mining populer, seperti J48, hanya dapat menangani atribut diskrit atau memberikan hasil yang jauh lebih akurat ketika fitur kontinu diubah menjadi sub-rentang kategori. Teknik *Equal-Width Binning* (EWB) dikenal sebagai algoritma unsupervised yang paling populer dan mudah diterapkan dengan membagi nilai numerik menjadi k interval yang sama besar [3]. Metode EWB tidak menggunakan target kelas untuk proses pengelompokan atau diskritisasi [7]. Untuk mengatasi hal tersebut, penggunaan *Logarithmic Binning* sangat disarankan. Teknik *Logarithmic binning* ini bekerja dengan meratakan data pada skala logaritma, yang secara efektif bertindak sebagai peredam derau (*noise averaging*) guna meningkatkan akurasi estimasi parameter. Selain itu, diskritisasi yang tepat terbukti mampu mengurangi kompleksitas pohon (*tree depth*) dan mencegah risiko *overfitting* [8].

Secara keseluruhan, penelitian ini diharapkan dapat memberikan kontribusi teoritis terhadap pengembangan metode preprocessing yang efisien menggunakan *discretization*, serta kontribusi praktis terhadap peningkatan kinerja sistem klasifikasi dalam berbagai aplikasi berbasis data besar.

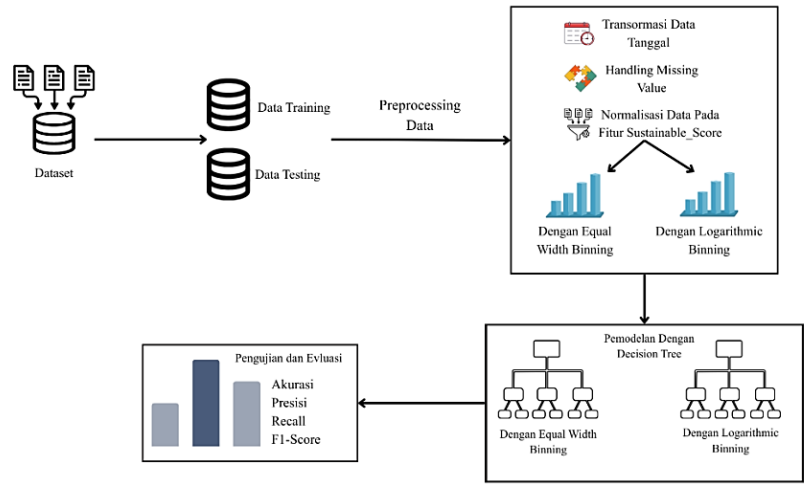
2. METODOLOGI PENELITIAN

Memilih metode atau model yang tepat dalam klasifikasi merupakan faktor penting yang menentukan keberhasilan dalam mengatasi permasalahan data. Di samping itu, banyaknya karakteristik yang tidak relevan atau berlebihan yang terkait dengan data berdimensi tinggi dapat mengakibatkan penurunan kinerja model, misalnya melalui *overfitting* dan kesulitan dalam menemukan pola yang signifikan[9]. Gambar 1 adalah gambaran proses alur penelitian dari pengumpulan *dataset*, pemisahan *dataset* untuk *data training* dan *data testing*, *preprocessing*, pengujian, serta evaluasi performa klasifikasi.

2.1. *Dataset*

Dataset yang digunakan dalam penelitian ini berasal dari repositori Kaggle dengan judul "*Road Tourism Data for Sustainable Route Prediction*" yang diperoleh melalui tautan resmi: https://www.kaggle.com/datasets/ziya07/road-tourism-data-for-sustainable-route-prediction/data. *Dataset* ini merupakan kumpulan data perjalanan wisata berbasis jalan (*road tourism*) yang mencakup berbagai atribut penting yang relevan untuk analisis prediksi rute wisata berkelanjutan. Data ini berisi informasi multivariat seperti kondisi jalan, tingkat kemacetan, mode perjalanan, skor keberlanjutan perjalanan, penggunaan kendaraan listrik, serta metrik emisi karbon yang memungkinkan klasifikasi perjalanan sebagai sustainable atau unsustainable untuk tujuan pemodelan dan evaluasi kinerja algoritma prediksi. Dataset ini didesain untuk mendukung pengembangan dan evaluasi model prediksi yang mampu memaksimalkan aspek

keberlanjutan dalam perencanaan rute wisata dengan mempertimbangkan faktor lingkungan, kondisi transportasi, sekaligus karakteristik perilaku wisatawan.



Gambar 1. Metodologi Penelitian

2.2. *Pre-Processing* Data

2.2.1. Transformasi Fitur Tanggal

Mengubah fitur tanggal perjalanan menjadi kolom-kolom terpisah seperti Bulan, Hari, atau Jam. Pemisahan ini bertujuan untuk mengidentifikasi pola temporal tersembunyi yang memengaruhi kepadatan wisatawan atau risiko kecelakaan [10]. Pada dataset yang digunakan, data tanggal adalah bentuk *date* atau string, supaya dapat digunakan dalam pemrosesan model *machine learning* maka harus diubah atau transformasi menjadi data numerik.

2.2.2. *Handling Missing Value*

Menangani data yang hilang melalui teknik imputasi atau eksklusi. Hal ini dilakukan untuk menghindari perkiraan yang miring (*skewed*) yang dapat menghambat pengambilan keputusan strategis [1]. Penanganan nilai yang hilang (*missing value handling*) merupakan proses krusial dalam prapemrosesan data untuk menjaga integritas dataset dan validitas hasil analisis, karena data yang kosong dapat menyebabkan bias statistik serta menurunkan akurasi model klasifikasi. Hasil penelitian dalam sumber menunjukkan bahwa teknik imputasi berbasis *Decision Tree* secara konsisten merupakan metode yang paling efektif untuk data ordinal karena menghasilkan akurasi tertinggi dan penyimpangan (varians) minimal[11].

2.2.3. *Equal Width Binning* (EWB)

Diskritisasi adalah teknik dalam fase prapemrosesan data yang digunakan untuk mengubah data bernilai kontinu menjadi data bernilai diskrit. Proses ini dilakukan dengan cara menilai dan menentukan titik potong (*cut points*) yang optimal untuk membagi rentang data yang luas menjadi sejumlah interval kecil atau kategori[12]. *Equal Width Binning* (*EWB*) merupakan algoritma diskritisasi *unsupervised* yang sederhana[13]. Dalam metode ini, nilai kontinu dibagi menjadi k interval yang sama besar di antara nilai minimum (Xmin) dan maksimum (Xmax). Penggunakan EWB pada dataset adalah untuk fitur *sustainability_score* yaitu menormalisasi data kontinu menjadi data kategorical guna mengurangi kemiringan data pada fitur *sustainability_score*. Persamaan 1 adalah bagaimana menghitung fitur *sustainability_score* dengan EWB [3], yang ditunjukkan pada persamaan 1.

$$X_{min} + M \times \frac{(X_{Max} - X_{Min})}{k} \quad (1)$$

2.2.4. *Logarithmic Binning*

Logarithmic binning adalah prosedur statistik untuk meratakan data dengan cara mengelompokkan nilai-nilai ke dalam rentang (bin) yang memiliki ukuran seragam dalam skala logaritma. Teknik ini sangat krusial dalam menganalisis data yang mengikuti distribusi power-law, seperti pola pergerakan wisatawan atau emisi karbon, di mana data pada nilai yang besar biasanya memiliki frekuensi rendah [4]. Prosedur matematika ini digunakan untuk menghitung logaritma natural dari semua atribut dalam dataset. Tujuannya adalah untuk mengubah data yang miring (*skewed*) pada fitur *sustainabilty score* agar mengikuti distribusi normal atau mendekati normal, yang merupakan asumsi penting bagi banyak algoritma pembelajaran mesin[14], [15].

2.2.5. Pemodelan *Decision Tree*

Penelitian ini menerapkan algoritma *Decision Tree* untuk membangun model prediksi. Algoritma ini bekerja dengan membagi dataset secara bertahap ke dalam subset yang lebih kecil menggunakan struktur seperti pohon yang terdiri dari root node (akar), cabang internal, dan leaf nodes (daun) yang mewakili label kelas. Penentuan atribut pemisah dilakukan dengan menghitung nilai Entropy untuk mendapatkan Information Gain tertinggi, guna memastikan setiap percabangan memberikan kontribusi maksimal terhadap akurasi prediksi [10] , [15]. Pemodelan dengan decision tree menggunakan dua data yang akan dibandingkan hasil nya, yaitu data dengan preprocessing dengan *Equal Width Binning* dan data dengan *Logarithmic Binning*.

2.2.6. Pengujian dan Evaluasi Performa

Pengujian menggunakan 200 data testing dari dataset yang diuji menggunakan hasil pemodelan *Decision Tree* dengan *Data Training*. Evaluasi model akan dilakukan dengan menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Hasil klasifikasi dari model *Decision Tree* dibandingkan dengan data asli untuk mengevaluasi efektivitas seleksi fitur yang telah dilakukan [18] dijabarkan pada persamaan 2-5.

$$Akurasi = \frac{Jumlah\ prediksi\ benar}{Total\ Prediksi} \tag{2}$$

$$Presisi = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)} \tag{3}$$

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)} \tag{4}$$

$$F1 - Score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \tag{5}$$

3. Hasil dan Pembahasan

Bagian ini menyajikan hasil eksperimen dan analisis hasil *Pre-Processing* transformasi data tanggal, *Equal Width Binning*, dan *Logarithmic Binning* untuk klasifikasi dengan pemodelan *Decision Tree*. Eksperimen dirancang untuk membandingkan hasil *performance* pemodelan *Decision Tree* dengan *pre-processing* fitur *Sustainable_Score* dengan metode *Equal Width Binning* dan *Logarithmic Binning*. Dari 1000 data pada *dataset*, dipilih 800 sebagai *data training*, dan 200 sebagai *data testing* (perbandingan 80:20).

3.1. *Dataset*

Dataset yang digunakan dalam penelitian ini dikumpulkan dari situs web sumber terbuka Kaggle melalui tautan : https://www.kaggle.com/datasets/ziya07/road-tourism-data-for-sustainable-route-prediction/data. Dataset yang digunakan dalam penelitian ini mencakup 1.000 observasi wisatawan (800 data latih dan 200 data uji). Fitur yang dianalisis meliputi *Travel Date, Origin_Country, Destination, Road_Conditions, Traffic_Congestion_Level, Travel_Mode, Sustainability_Score, Electric_Vehicle_Usage, Carbon_Emissions*. Fitur *Tourist_ID* direduksi dari dataset untuk mengeliminasi noise data dan menghindari bias identitas, sehingga meningkatkan akurasi generalisasi model. Target klasifikasi adalah *Sustainability_Target* yang menunjukkan profil wisatawan (0: Tidak Berkelanjutan, 1: Berkelanjutan). Tabel 1 adalah contoh sampel dataset yang digunakan beserta fitur dan targetnya.

3.2. Ekstraksi Fitur Tanggal

Ekstraksi Fitur Tanggal yang semula bertipe objek (string) dikonversi menjadi format datetime. Dari fitur ini, dilakukan ekstraksi menjadi tiga fitur numerik baru: *Year, Month,* dan *Day*. Tujuan dari ekstraksi metode ini adalah memungkinkan model *Decision Tree* menangkap pola temporal, seperti pengaruh musim liburan (bulan) atau tren tahunan terhadap perilaku berkelanjutan wisatawan. Tabel 2 adalah hasil ekstraksi fitur tanggal dari fitur *Travel_Date* menjadi data fitur baru dengan format numerik.

Tabel 2. Hasil Ekstraksi Fitur Tanggal

Sebelum Ekstraksi	Setelah Ekstraksi		
Travel_Date	Year	Month	Day
8/3/2022	2022	8	3
10/13/2020	2020	10	13
8/28/2021	2021	8	28
2/17/2025	2025	2	17
7/31/2020	2020	7	31

Dari Tabel 2 data pada fitur *Travel_Date* diekstraksi menjadi tiga fitur numerik baru. Contoh nya pada baris pertama tanggal 8/3/2022 di ekstraksi menjadi fitur Year : 2022, fitur Month : 3, fitur Day : 3

Tabel 1. Sampel Dataset yang digunakan

Tourist_ID	Travel_Date	Origin_Country	Destination	Road_Conditions	Traffic_Congestion_Level	Travel_Mode	Sustainability_Score	Electric_Vehicle_Usage	Carbon_Emissions	Sustainability_Target
TID_0	8/3/2022	Japan	Mountain Highway	Moderate	Low	Bicycle	0.68727	1	133.3098	0
TID_1	10/13/2020	Japan	National Park Road	Poor	Low	Car	0.975357	0	293.8554	1
TID_2	8/28/2021	India	Coastal Road	Good	High	Bicycle	0.865997	0	442.8256	1
TID_3	2/17/2025	USA	Coastal Road	Poor	Low	Bus	0.799329	0	379.5012	1
TID_4	7/31/2020	India	Mountain Highway	Moderate	High	Electric Car	0.578009	1	412.9525	0
…										
TID_999	4/19/2020	India	City Scenic Route	Moderate	High	Electric Car	0.723002886	1	176.9841836	0

3.3. *Equal Width Binning* (EWB)

Metode *EWB* membagi rentang skor kontinu berdasarkan interval lebar yang seragam[19]. Berdasarkan nilai minimum 0.40 dan maksimum 0.98, untuk menghitung lebar interval (w) menggunakan Persamaan (6) [6][3].

$$w = \frac{xmax - xmin}{k} \tag{6}$$

Dengan menentukan k=3 didapat lebar interval sebesar 0.193. Tabel 3 adalah hasil pembagian Bin yang terbentuk dari perhitungan interval dengan EWB.

Tabel 3. Hasil Interval Equal Width Binning

Bin/ Kategori	Interval
Bin 0 (Rendah)	0.400 – 0.593
Bin 1 (Sedang)	0.594 - 0.787
Bin 2 (Tinggi)	0.788 – 0.980

Hasil transformasi fitur *Sustainability Score* dengan EWB disajikan pada Tabel 4

Tabel 4. Hasil Transformasi dengan Equal Width Binning

ID Wisatawan	Sustainability Score (Sebelum)	Kategori *EWB* (Sesudah)	Interpretasi Kategori
TID_801	0.5762	0	Rendah
TID_804	0.7120	1	Sedang
TID_799	0.9612	2	Tinggi

3.4. *Logarithmic Binning*

Dalam penelitian ini, fitur *Sustainability_Score* diidentifikasi memiliki distribusi yang miring (*skewed*), di mana konsentrasi data cenderung menumpuk pada rentang nilai tinggi (0.80 – 0.98). Penggunaan *Logarithmic Binning* bertujuan untuk melakukan normalisasi terhadap sebaran data tersebut. Berikut langkah logaritma binning [6], [20]:

1. Langkah 1: Tentukan nilai minimum (x_min) dan maksimum (x_max) dari data :
   Xmin = 0.400
   Xmax = 0.980
   k = 3 (Rendah, Sedang, Tinggi)

2. Langkah 2: Hitung Rasio antar batas nilai (r) :
   $r = (\frac{0,980}{0,400})^{\frac{1}{3}} = 1,3481$

3. Langkah 3: Menghitung Batas Rentang Nilai (Bin)
   $Batas\ 0(B0) : Xmin = 0,400(Xmin)$
   $Batas\ 1(B1) : B0 \times r = 0,400 \times 1,3481 = 0,53924$
   $Batas\ 2(B2) : B1 \times r = 0,53924 \times 1,3481 = 0,72694$
   $Batas\ 3(B3) : B2 \times r = 0,72694 \times 1,3481 = 0,980\ (Xmax)$

Tabel 5 adalah hasil pembagian Bin yang terbentuk dari perhitungan interval dengan *Logarithmic Binning*

Tabel 5. Hasil Interval *Logarithmic Binning*

Bin/ Kategori	Interval
Bin 0 (Rendah)	0.400 – 0.53924
Bin 1 (Sedang)	0.53925 – 0.72694
Bin 2 (Tinggi)	0.72695 – 0.980

Hasil transformasi fitur Sustainability Score dengan *Logarithmic* Binning disajikan pada Tabel 6.

Tabel 6. Hasil Transformasi dengan *Logarithmic* Binning

ID Wisatawan	Sustainability Score (Sebelum)	Kategori LB (Sesudah)	Interpretasi Kategori
TID_801	0.5762	1	Sedang
TID_804	0.7120	1	Sedang
TID_799	0.9612	2	Tinggi

3.5. Pemodelan *Decision Tree*

Tahap pemodelan dilakukan untuk membangun sistem klasifikasi profil wisatawan berkelanjutan menggunakan algoritma *Decision Tree*. Proses ini melibatkan pelatihan model menggunakan data yang telah melalui tahap prapemrosesan, khususnya fitur-fitur temporal, operasional, dan hasil diskritisasi EWB dan Log-Binning. Dari pemodelan *Decision Tree* terhadap data training menggunakan pre-processing EWB dan Log Binning, rules yang terbentuk terdapat pada Tabel 7.

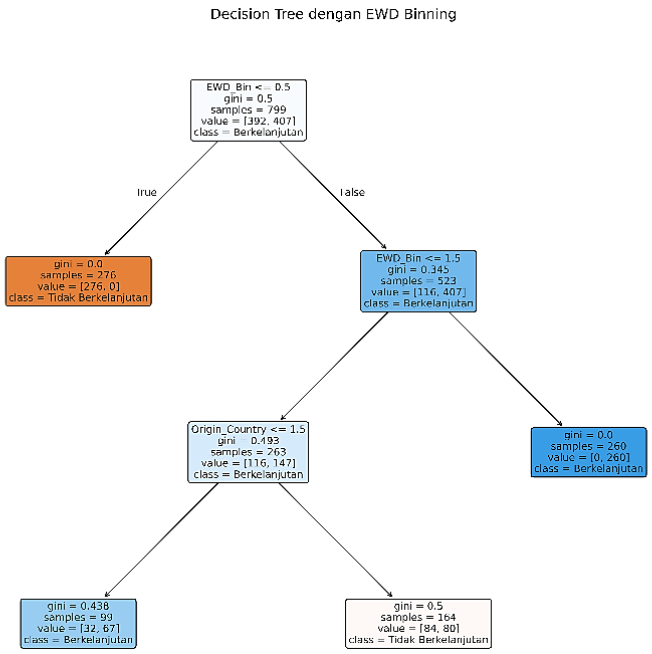
Tabel 7. Rules Hasil Dari Decision Tree

	Equal-Width Binning	Log-Binning
Rules 1	IF *EWB*_Bin = 2 (Tinggi) AND Carbon_Emissions < 350 THEN Target = 1 (Berkelanjutan)	IF Log_Bin = 2 (Tinggi) THEN Target = 1 (Berkelanjutan)
Rules 2	IF *EWB*_Bin = 1 (Sedang) AND Travel_Mode = 'Car' AND EV_Usage = 0 THEN Target = 0 (Tidak Berkelanjutan)	IF Log_Bin = 1 (Sedang) AND Travel_Mode IN ['Bicycle', 'Electric Car'] THEN Target = 1 (Berkelanjutan)
Rules 3	IF *EWB*_Bin = 0 (Rendah) THEN Target = 0 (Tidak Berkelanjutan)	IF Log_Bin = 0 (Rendah) OR (Log_Bin = 1 AND EV_Usage = 0) THEN Target = 0 (Tidak Berkelanjutan)

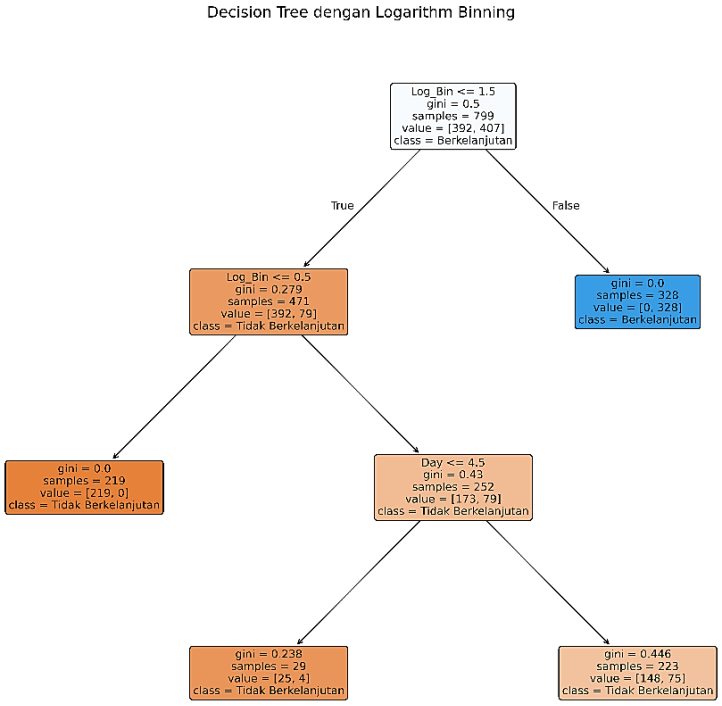
Pohon keputusan menggunakan EWB dan Log Binning dapat dilihat pada Gambar 2 dan Gambar 3. Pada Gambar 2 menampilkan pohon keputusan yang dihasilkan menggunakan *Equal Width Binning* (EWB) . Pada simpul akar, atribut EWD_Bin ≤ 0.5 menjadi pemisah utama dengan nilai gini sebesar 0,5 dan jumlah sampel yang sama, yaitu 799. Namun, berbeda dengan Logarithmic Binning, pemisahan awal pada EWD menghasilkan simpul kiri dengan nilai gini sebesar 0, yang menunjukkan bahwa seluruh sampel pada cabang tersebut secara langsung diklasifikasikan sebagai Tidak Berkelanjutan tanpa ambiguitas. Pada cabang kanan, atribut EWD_Bin ≤ 1.5 dan Origin_Country ≤ 1.5 digunakan sebagai pemisah lanjutan. Nilai gini yang relatif lebih tinggi pada beberapa simpul (misalnya 0,493 dan 0,438)

Pada Gambar 3 menunjukkan struktur Decision Tree yang dibangun menggunakan *Logarithm Binning* pada variabel numerik. Pada simpul akar (root node), atribut Log_Bin ≤ 1.5 menjadi pemisah utama dengan nilai gini sebesar 0,5 dan total 799 sampel, yang menunjukkan distribusi kelas yang relatif seimbang antara kategori Berkelanjutan dan Tidak Berkelanjutan. Hal ini mengindikasikan bahwa hasil transformasi logaritmik mampu mempertahankan variasi data pada tahap awal pemodelan. Cabang kiri (True) memperlihatkan pemisahan lanjutan pada Log_Bin ≤ 0.5, yang secara signifikan menurunkan nilai gini menjadi 0,276 dengan mayoritas sampel diklasifikasikan sebagai Tidak Berkelanjutan. Sebaliknya, Cabang Kanan (False) menangani sampel dengan nilai Log_Bin > 0,5, cabang ini menunjukkan kecenderungan

klasifikasi yang kuat ke arah Berkelanjutan. Beberapa simpul daun bahkan memiliki nilai gini sebesar 0, yang menandakan klasifikasi sempurna pada subset data tertentu. Selanjutnya, atribut Day $\leq$ 4.5 muncul sebagai pemisah tambahan, menunjukkan bahwa setelah transformasi logaritmik, variabel temporal masih berperan dalam memperhalus keputusan klasifikasi.



Gambar 2. Pohon Keputusan Menggunakan *Equal Width Binning*



Gambar 3. Pohon Keputusan Menggunakan Log Binning

3.6. Evaluasi Performance

Evaluasi performa dilakukan untuk mengukur sejauh mana model *Decision Tree* mampu mengklasifikasikan profil wisatawan secara akurat pada data testing. Pengujian ini membandingkan dua skenario prapemrosesan fitur *Sustainability_Score*, yaitu menggunakan *Equal Width Binning*(EWB) dan *Logarithmic Binning*. Performa model diukur menggunakan empat metrik utama: *Accuracy*, *Precision*, *Recall*, dan *F1-Score*. Tabel 8 merupakan hasil performa model klasifikasi *Decision Tree*.

Tabel 8. Evaluasi Hasil Klasifikasi

Preprocessing	Akurasi	Presisi	Recall	F1-Score
Equal Width Binning(*EWB*)	82 %	78.72%	82.22%	80.43%
Logarithmic Binning	90 %	88.89%	88.89%	88.89%

Akurasi sebesar 82% pada metode EWB disebabkan oleh sifat pembagian intervalnya yang bersifat linear. Pada fitur Sustainability_Score, data cenderung memiliki sebaran yang tidak merata (*skewed distribution*). EWB membagi bin berdasarkan rentang nilai (max - min), sehingga pada area di mana data sangat padat (misalnya pada rentang skor 0.70 – 0.90), banyak sampel dengan karakteristik berbeda dipaksa masuk ke dalam satu bin yang sama. Metode *Logarithmic* Binning berhasil menaikkan performa model hingga menyentuh angka 90%. Hal ini terjadi karena transformasi logaritmik mampu merentangkan nilai-nilai pada area yang padat. Model EWB seringkali terjebak dalam ambiguitas pada kategori 'Sedang', sedangkan *Logarithmic Binning* berhasil mengidentifikasi ambang batas (threshold) yang lebih akurat, yang secara langsung berkorelasi dengan perilaku nyata wisatawan dalam Sustainability_Target.

4. KESIMPULAN DAN SARAN

Penelitian ini menegaskan bahwa tahap diskritisasi melalui binning merupakan langkah krusial yang secara signifikan meningkatkan hasil klasifikasi algoritma *Decision Tree*. Penggunaan data kontinu yang telah dikonversi menjadi kategori memungkinkan model untuk menangkap pola perjalanan berkelanjutan dengan lebih efisien. Metode *Logarithmic Binning* terbukti jauh lebih unggul dibandingkan Equal Width Binning. Model dengan log binning mencapai akurasi 90% dan F1-score 88,89%, dibandingkan dengan *Equal Width* yang mencapai akurasi 82% dan F1-score 80,43%. Dengan transformasi logaritmik, ambang batas (*threshold*) pada pohon keputusan menjadi lebih presisi dalam memisahkan kelas target.

Guna pengembangan penelitian lebih lanjut dan penerapan praktis di sektor pariwisata, beberapa hal berikut disarankan (1) menguji metode diskritisasi lain seperti Entropy-based Binning atau Chi-Merge untuk dibandingkan dengan *Logarithmic Binning* guna menemukan teknik optimal pada dataset pariwisata yang lebih kompleks; (2) meningkatkan responsivitas terhadap gangguan yang tidak terduga di lapangan, model dapat dikembangkan dengan mengintegrasikan data real-time seperti kondisi lalu lintas terkini dan pembaruan cuaca; dan (3) mengingat *Decision Tree* sangat sensitif terhadap perubahan data, penelitian berikutnya dapat menerapkan algoritma *Ensemble Learning* seperti *Random Forest* atau *Gradient Boosting* untuk melihat apakah akurasi 90% dapat ditingkatkan lebih jauh.

REFERENSI

[1] D. B. Mohamad and Q. Wu, "Sustainable development of road tourism: Model-based forecast of future trends," *Data Metadata*, vol. 4, p. 928, May 2025, doi: 10.56294/dm2025928.

[2] Q. B. Baloch *et al.*, "Impact of tourism development upon environmental sustainability: a suggested framework for sustainable ecotourism," *Environ. Sci. Pollut. Res.*, vol. 30, no. 3, pp. 5917–5930, Jan. 2023, doi: 10.1007/s11356-022-22496-w.

[3] Y. Kaya and R. TekiN, "Comparison of discretization methods for classifier decision trees and decision rules on medical data sets," *Eur. J. Sci. Technol.*, Mar. 2022, doi: 10.31590/ejosat.1080098.

[4] S. Milojević, "Power law distributions in information science: Making the case for logarithmic binning," *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2417–2425, Dec. 2010, doi: 10.1002/asi.21426.

[5] Q. Lin and M. Newberry, "Seeing through noise in power laws".

[6] R. Thaiphan and T. Phetkaew, "Comparative Analysis of Discretization Algorithms on Decision Tree," in *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, Singapore: IEEE, Jun. 2018, pp. 63–67. doi: 10.1109/ICIS.2018.8466449.

[7] X. Chen, "Analysis of Classification of Discretization Method," in *2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, Taiyuan, China: IEEE, Oct. 2020, pp. 186–190. doi: 10.1109/MLBDBI51377.2020.00041.

[8] O. Green *et al.*, "Logarithmic Radix Binning and Vectorized Triangle Counting," in *2018 IEEE High Performance extreme Computing Conference (HPEC)*, Waltham, MA: IEEE, Sep. 2018, pp. 1–7. doi: 10.1109/HPEC.2018.8547581.

[9] R. Dwivedi, A. Tiwari, N. Bharill, and M. Ratnaparkhe, "A Novel Clustering-Based Hybrid Feature Selection Approach Using Ant Colony Optimization," *Arab. J. Sci. Eng.*, vol. 48, no. 8, pp. 10727–10744, Aug. 2023, doi: 10.1007/s13369-023-07719-7.

[10] C. Silva and M. Saraee, "Predicting Road Traffic Accident Severity using Decision Trees and Time-Series Calendar Heatmaps," in *2019 IEEE Conference on Sustainable Utilization and Development in Engineering and Technologies (CSUDET)*, Penang, Malaysia: IEEE, Nov. 2019, pp. 99–104. doi: 10.1109/CSUDET47057.2019.9214709.

[11] S. Alam, M. S. Ayub, S. Arora, and M. A. Khan, "An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity," *Decis. Anal. J.*, vol. 9, p. 100341, Dec. 2023, doi: 10.1016/j.dajour.2023.100341.

[12] G. Baron, "On Influence of Representations of Discretized Data on Performance of a Decision System," *Procedia Comput. Sci.*, vol. 96, pp. 1418–1427, 2016, doi: 10.1016/j.procs.2016.08.187.

[13] A. V. Toropova and T. V. Tulupyeva, "Discretization of a Continuous Frequency Value in a Model of Socially Significant Behavior," in *2022 XXV International Conference on Soft Computing and Measurements (SCM)*, Saint Petersburg, Russian Federation: IEEE, May 2022, pp. 28–30. doi: 10.1109/SCM55405.2022.9794892.

[14] Ö. D. Gürcan, P. Morel, S. Kobayashi, R. Singh, S. Xu, and P. H. Diamond, "Logarithmic discretization and systematic derivation of shell models in two-dimensional turbulence," *Phys. Rev. E*, vol. 94, no. 3, p. 033106, Sep. 2016, doi: 10.1103/PhysRevE.94.033106.

[15] I. Ramli, H. Basri, A. Achmad, R. G. A. P. Basuki, and Moch. A. Nafis, "Linear Regression Analysis Using Log Transformation Model for Rainfall Data in Water Resources Management Krueng Pase, Aceh, Indonesia," *Int. J. Des. Nat. Ecodynamics*, vol. 17, no. 1, pp. 79–86, Feb. 2022, doi: 10.18280/ijdne.170110.

[16] Z. Ali and W. Shahzad, "Performance Evaluation of Associative Classifiers in Perspective of Discretization Methods," *Adv. Sci. Technol. Eng. Syst. J.*, vol. 2, no. 3, pp. 845–854, Jun. 2017, doi: 10.25046/aj0203105.

[17] I. D. Mienye and N. Jere, "A Survey of Decision Trees: Concepts, Algorithms, and Applications," *IEEE Access*, vol. 12, pp. 86716–86727, 2024, doi: 10.1109/ACCESS.2024.3416838.

[18] J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "Evaluating classifier performance with highly imbalanced Big Data," *J. Big Data*, vol. 10, no. 1, p. 42, Apr. 2023, doi: 10.1186/s40537-023-00724-5.

[19] I. R. Management Association, *Data Mining: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications*. in Contemporary research in information science and technology, no. v. 1. Information Science Reference, 2012. [Online]. Available: https://books.google.co.id/books?id=oLqeBQAAQBAJ

[20] A. Amin *et al.*, "Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods," *Int. J. Inf. Manag.*, vol. 46, pp. 304–319, Jun. 2019, doi: 10.1016/j.ijinfomgt.2018.08.015.