Analisis Preferensi Pengunjung Wisata Kabupaten Pekalongan Menggunakan Algoritma Decision Tree

Visitor Preference Analysis of Tourism in Pekalongan Regency Using the Decision Tree Algorithm

Muhamad Rizaludin*[1], Husni Hidayat[2], Mujibul Hakim[3]
Program Studi Teknologi Informasi, ITSNU Pekalongan, Jl.Karangdowo No.9 Kedungwuni Kab. Pekalongan, 0285 7831614
E-mail : rizal.lonly@gmail.com[1], husni.hidayat1@gmail.com[2], mujibul.hakim@gmail.com[3]*
**Corresponding author*
Received 28 November 2025; Revised 25 December 2025; Accepted 6 January 2026

Abstrak - Pariwisata merupakan sektor strategis yang memiliki potensi besar dalam meningkatkan pertumbuhan ekonomi daerah, termasuk di kabupaten Pekalongan. Dalam upaya meningkatkan daya saing dan efektivitas promosi wisata, dibutuhkan pemahaman yang mendalam terhadap preferensi pengunjung. Penelitian ini bertujuan untuk menganalisis preferensi pengunjung wisata berdasarkan atribut demografis seperti jenis kelamin, usia, tingkat pendidikan, dan pekerjaan, serta kategori wisata yang diminati. Algoritma Decision Tree digunakan sebagai metode klasifikasi untuk memodelkan hubungan antara data demografis dan ketegori wisata yang dipilih. Tahapan penelitian meliputi pengumpulan data, preprocessing, pemisahan data menjadi training dan testing, pembangunan model Decision Tree, evaluasi kinerja model, hingga interpretasi hasil pohon keputusan. Data dikumpulkan dari 500 responden lokal melalui kuesioner online dan diproses secara sistematis meliputi imputasi nilai hilang, penghapusan duplikasi, validasi kategori, encoding numerik serta oversampling untuk menangani ketidakseimbangan kelas. Model klasifikasi dibangun menggunakan algoritma Decision Tree, dengan pembagian data latih dan uji. Evaluasi terhadap data uji menunjukkan akurasi sebesar 76,92%, dan model mampu menginterpretasikan faktor-faktor demografis yang paling efektif dalam memprediksi kategori wisata sesuai preferensi pengunjung.

Kata Kunci - Decision Tree, Preferensi Pengunjung Wisata, Klasifikasi Demografis, Rekomendasi Wisata

Abstract - *Tourism is strategic sector with substantial to elevate regional economic growth, particularly in Pekalongan regency. To enhance competitiveness and the effectiveness of tourism promotion, it is essential to deeply understand visitor preferences. This study aims to analyze tourism preferences based on demographic attributes such as gender, age, education level, and occupation, as well as preferred tourism categories. The Decision Tree (C4.5) algorithm is used as a classification method to model the relationship between demographic data and chosen tourism categories. The research steps include data collection, preprocessing, splitting into training and testing sets, model building using Decision Tree, model evaluation, and interpretation of the resulting Decision Tree. Data were collected from 500 local respondents via an online questionnaire and were processed systematically, including imputation of missing values, removal of duplicates, category validation, numeric encoding, and oversampling to deal with class imbalance. The classification model was constructed using the Decision Tree, with data split into training and test sets. Evaluation on the test set revealed an accuracy of 76,92%, indicating that the model effectively captures complex patterns and reduces overfitting through informed attribute splitting and pruning techniques typical of Random Forest. The model also identifies which demographic factors are most influential tourism category preferences.*

Keywords - *Decision Tree, Tourist Visitor Preferences, Demographic Classification, Tourism Recommendations*

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15171

1. PENDAHULUAN

Kabupaten Pekalongan merupakan salah satu wilayah di provinsi Jawa Tengah yang memiliki kekayaan potensi wisata yang sangat beragam. Keanekaragaman ini mencakup berbagai bentuk wisata seperti wisata alam, budaya, sejarah, religi, kuliner, edukasi, hingga wisata belanja. Dengan keanekaragaman tersebut, kabupaten Pekalongan berpeluang besar untuk mengembangkan sektor pariwisata sebagai salah satu pilar perekonomian daerah [1]. Keindahan panorama alam seperti pegunungan, air terjun (Curug Telaga Lumbu, Curug Kali Banteng, Curug Lawe, Curug Muncar, Curug Bajing), dan pantai (Sunter Beach Depok, Wonokerto Beach, Curug Siwatang, Bukit Pawuluhan), serta kekayaan budaya seperti batik khas Pekalongan yang telah mendunia, di mana Pekalongan menjadi anggota UNESCO *Creative Cities Network* sejak desember 2014 (bidang kerajinan dan seni rakyat), menjadi daya tarik tersendiri bagi wisatawan lokal maupun mancanegara [2].

Namun, terlepas dari potensi yang besar tersebut, masih banyak tantangan yang harus dihadapi, salah satunya adalah pemahaman yang terbatas mengenai preferensi pengunjung terhadap jenis wisata yang ditawarkan. Preferensi wisatawan terhadap suatu destinasi sangat dipengaruhi oleh berbagai faktor demografis seperti usia, jenis kelamin, tingkat pendidikan, dan jenis pekerjaan. Pemahaman terhadap preferensi ini sangat penting dalam perencanaan dan pengembangan destinasi wisata. Tanpa informasi yang memadai mengenai kebutuhan dan keinginan pengunjung, strategi promosi dan pengembangan wisata cenderung tidak tepat sasaran [3]. Pendekatan berbasis data menjadi sangat relevan di tengah peningkatan kunjungan wisatawan, seperti kunjungan wisata di Pekalongan pada semester pertama tahun 2025 mencapai peningkatan sebesar 35,38%, dengan pantai Pasir Kencana tetap menjadi primadona utama. Oleh karena itu, diperlukan pendekatan berbasis data yang mampu mengungkapkan pola dan kecenderungan preferensi wisatawan secara sistematis dan objektif [4].

Penggunaan algoritma Decision Tree dalam konteks pariwisata memiliki berbagai keunggulan. Pertama, *Decision Tree* mampu menangani atribut data yang bersifat kategorikal maupun numerik. Kedua, algoritma ini menghasilkan *output* dalam bentuk yang mudah dimengerti oleh pengguna awam sekalipun, karena struktur pohon keputusan menyerupai proses pengambilan keputusan secara logis. Ketiga, *Decision Tree* dapat menunjukkan atribut-atribut mana yang paling berpengaruh dalam menentukan preferensi wisatawan. Dengan keunggulan-keunggulan tersebut, *Decision Tree* sangat cocok diterapkan dalam analisis preferensi pengunjung wisata yang memiliki berbagai atribut demografis [5].

Implementasi *Decision Tree* dalam penelitian ini akan menggunakan atribut demografis sebagai variabel *input*, seperti jenis kelamin, usia, pendidikan, dan pekerjaan. Selanjutnya, preferensi terhadap jenis wisata seperti wisata alam, sejarah, religi, petualangan, seni dan budaya, kuliner, dan lainnya akan menjadi variabel target atau *output*. Model yang diharapkan mampu mengklasifikasikan atau memprediksi jenis wisata yang disukai oleh pengunjung dengan karakteristik tertentu. Selain itu, penelitian ini juga akan mengevaluasi kinerja model berdasarkan metrik-metrik seperti akurasi, *precision*, *recall*, dan *f1-score* untuk memastikan bahwa model yang dibangun memiliki tingkat keandalan yang baik [6].

Penelitian ini memiliki beberapa aspek kebaruan dibandingkan dengan penelitian-penelitian terdahulu dalam domain pariwisata, khususnya yang menggunakan algoritma *Decision Tree* untuk analisis preferensi wisata:

1.1. Analisis preferensi berbasis demografi lengkap: Penelitian sebelumnya menggunakan *Decision Tree* untuk rekomendasi wisata di Lombok Timur, namun fokusnya pada sistem pakar tanpa analisis mendalam terhadap atribut demografis seperti jenis kelamin, usia, pendidikan, dan pekerjaan secara bersama-sama sebagai variabel prediktor utama preferensi wisatawan [7].

1.2. Jumlah responden yang lebih besar dan representatif: Studi lain pemanfaatan *Decision Tree* untuk prediksi kunjungan wisata di Pati menggunakan data sekunder besar, tetapi tidak

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15171

melibatkan responden langsung dalam jumlah besar yang mewakili populasi lokal. Penelitian ini menggunakan 500 responden survei online, memberikan basis data yang lebih kuat untuk analisis preferensi aktual masyarakat lokal [8].

1.3. Preprocessing data menyeluruh: Pada studi hanya fokus pada penerapan algoritma tanpa detail preprocessing yang komprehensif. Penelitian ini mencakup imputasi nilai hilang, pemeriksaan duplikat, validasi kategori, transformasi numerik, *one-hot encoding*, dan oversampling untuk menangani ketidakseimbangan kelas, sehingga model lebih akurat dan representatif [8].

1.4. Penggunaan oversampling dalam model klasifikasi wisata: Teknik oversampling tidak disebut secara eksplisit pada penelitian pariwisata *Decision Tree* terdahulu. Penerapan oversampling di penelitian ini membantu menyamakan distribusi kelas sehingga model lebih adil terhadap kategori wisata minoritas [9].

1.5. Evaluasi performa yang lebih komprehensif: Studi lain membandingkan akurasi antara algoritma *Decision Tree* dan SVM, penelitian ini memberikan analisis performa yang lebih focus pada interpretasi pohon keputusan serta dampaknya terhadap kebijakan pariwisata lokal [10].

1.6. Fokus pada konteks kebupaten tertentu: Penelitian lain untuk sistem rekomendasi diterapkan di kota Yogyakarta. Penelitian ini memfokuskan pada konteks lokal Kabupaten Pekalongan, memberikan *insight* yang lebih kontekstual untuk pengembangan pariwisata daerah [9].

1.7. Menghubungkan hasil algoritma dengan strategi promosi dan kebijakan pariwisata: Penelitian ini tidak hanya membangun model, tetapi juga mengevaluasi implikasi hasilnya terhadap strategi pengembangan pariwisata, sedangkan studi terdahulu sering berhenti pada aspek teknis model saja [8].

1.8. Integrasi responden lokal dan calon wisatawan luar daerah dalam survei: Pada studi yang hanya mengandalkan data sekunder atau data ulasan online, penelitian ini melibatkan responden lokal dan calon wisatawan eksternal secara aktif melalui survei online yang terdistribusi luas [9].

1.9. Tingkat akurasi yang kompetitif untuk konteks pariwisata regional: Hasil akurasi 76,92% menunjukkan model *Decision Tree* yang lebih stabil dalam konteks preferensi berbagai kategori wisata, mengungguli beberapa studi klasifikasi sebelumnya yang akurasinya kurang optimal tanpa preprocessing matang [8].

1.10. Interpretasi faktor demografis yang paling berpengaruh: Walaupun algoritma Decision Tree sering digunakan, penelitian ini menekankan analisis atribut demografis strategis yang paling berpengaruh terhadap preferensi wisatawan, namun pada aspek ini tidak banyak dianalisis secara mendalam pada penelitian sebelumnya [11].

1.11. Penerapan data mining yang mencakup aspek pemberdayaan masyarakat: Pendekatan pengumpulan data yang langsung melibatkan masyarakat lokal turut memberikan kontribusi sosial dalam pemetaan preferensi, tidak hanya sekedar analisis teknis [9].

1.12. Relevansi terhadap kebijakan pariwisata daerah: Kebaruan lain ada pada kemampuan model untuk memberikan rekomendasi yang mendukung perencanaan promosi wisata di tingkat pemerintahan daerah, bukan hanya rekomendasi berbasis sistem IT saja [8].

1.13. Pendekatan survei online adaptif terhadap tren digital: Metode pengumpulan data secara online dengan distribusi strategis di media sosial memperkaya data dibandingkan penelitian yang hanya mengandalkan survei tradisional atau data sekunder [9].

1.14. Kontribusi terhadap literatur *Decision Tree* di konteks pariwisata Indonesia: Meskipun ada penelitian Decision Tree di pariwisata pada level kota atau obyek wisata tertentu, penelitian ini memperluas wawasan analitik dengan fokus pada klasifikasi preferensi wisata secara demografis di tingkat kabupaten [7].

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15171

Dengan demikian, dapat disimpulkan bahwa pemanfaatan algoritma *Decision Tree* untuk menganalisis preferensi pengunjung wisata di kabupaten Pekalongan merupakan pendekatan yang relevan dan potensial. Melalui penelitian ini, diharapkan dapat diperoleh model klasifikasi yang akurat dan informatif, yang tidak hanya bermanfaat dalam meningkatkan efektivitas strategi promosi wisata, tetapi juga mendukung pengembangan destinasi wisata yang lebih sesuai dengan kebutuhan dan minat masyarakat. Hasil penelitian ini juga dapat menjadi kontribusi ilmiah dalam bidang data mining terapan dan pengembangan sistem rekomendasi berbasis kecerdasan buatan di sektor pariwisata [12].

2. METODE PENELITIAN

Penelitian ini dilakukan melalui serangkaian tahapan sistematis untuk menganalisis preferensi pengunjung wisata di kabupaten Pekalongan menggunakan algoritma *Decision Tree*. Tujuan *Decision Tree* adalah membagi dataset menjadi subset yang semakin homogen terhadap target kelas. Pemecahan dilakukan berdasarkan atribut-atribut input yang memiliki kemampuan terbaik dalam memisahkan kelas target. Berikut rumus *Decision Tree*:

- *Entropy* adalah ukuran ketidakpastian (*impurity*) dalam dataset. Semakin rendah nilai *entropy*, semakin homogen subset data tersebut terhadap satu kelas. *Entropy* dihitung dengan rumus:

  $$Entropy\ (S) = -\sum_{i=1}^{k} pi \log_2(pi)$$

  Dimana:
  S = keseluruhan dataset pada node tertentu
  k = jumlah kelas target
  pi = proporsi observasi pada kelas ke-i dalam dataset S
- *Information Gain* (IG) mengukur pengurangan *entropy* setelah dataset dibagi berdasarkan atribut A. atribut yang menghasilkan IG paling besar adalah atribut terbaik untuk split pertama dan selanjutnya. Rumusnya:

  $$Gain(S, A) = Entropy(S) - \sum_{v \epsilon Values(A)} \frac{|S_v|}{|S|} . Entropy(S_v)$$

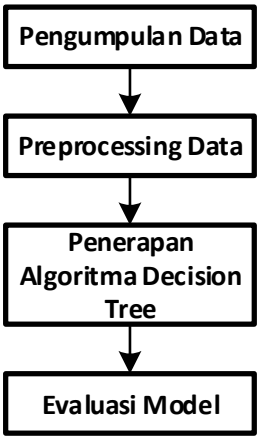
  Dimana:
  Values (A) = set nilai unik atribut A
  S_v = subset data di mana atribut A bernilai v
  $|S_v|/|S|$ = proporsi subset terhadap keseluruhan data.
- *Gain Ratio* (Pada C4.5) dapat memperbaiki bias IG terhadap atribut dengan banyak nilai melalui *Gain Ratio*. Rumusnya:

  $$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)}$$

  Dimana *SplitInfo* adalah ukuran entropi berdasarkan fragmentasi atribut itu sendiri:

  $$SplitInfo(S, A) = -\sum_{v \epsilon Values(A)} \frac{|S_v|}{|S|} \log_2 \left(\frac{|S_v|}{|S|}\right)$$

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15171



Gambar 1. Tahapan Penelitian

2.1. Pengumpulan Data

Data dikumpulkan dari berbagai sumber yang relevan dan terpercaya, mencakup survei langsung kepada pengunjung wisata, wawancara mendalam dalam pelaku industri pariwisata, serta data resmi yang diperoleh dari Dinas Pariwisata kabupaten Pekalongan. Proses pengumpulan data dilakukan secara sistematis untuk memastikan kelengkapan dan akurasi, meliputi pengumpulan informasi demografis pengunjung seperti jenis kelamin, usia tingkat pendidikan, dan pekerjaan. Pendekatan yang komprehensif ini diharapkan dapat memberikan gambaran yang lebih utuh mengenai preferensi wisatawan, sehingga hasil analisis nantinya dapat digunakan secara efektif dalam perencanaan dan pengembangan destinasi wisata di kabupaten Pekalongan.

2.2. *Preprocessing* Data

Setelah seluruh data terkumpul dari berbagai sumber, tahap berikutnya adalah melakukan *preprocessing* atau pra-pemrosesan data untuk memastikan bahwa data yang digunakan memiliki kualitas tinggi, konsisten, dan siap dianalisis. Proses ini dimulai dengan Langkah data *cleaning* atau pembersihan data, yaitu mengidentifikasi serta menghapus data yang duplikat, memperbaiki kesalahan pengetikan, dan mengatasi data yang hilang (*missing value*) dengan metode yang sesuai, seperti pengisian menggunakan nilai rata-rata, median, atau pendekatan *predictive imputation* [8]. Selanjutnya, dilakukan pengecekan terhadap data yang tidak valid atau berada di luar jangkauan logis (*outliers*) untuk memastikan bahwa data yang tersisa benar-benar mewakili kondisi di lapangan. Tahap berikutnya adalah data *normalization* atau normalisasi data, yang bertujuan menyamakan skala antarvariabel sehingga analisis menjadi lebih akurat dan tidak bias akibat perbedaan skala pengukuran. Proses *preprocessing* ini sangat penting karena kualitas hasil analisis sangat bergantung pada kebersihan dan kelengkapan data. Dengan melakukan langkah ini secara sistematis, data yang dihasilkan akan memiliki reliabilitas tinggi, meminimalkan potensi kesalahan dalam analisis, dan memastikan bahwa model atau metode yang diterapkan, seperti algoritma *Decision Tree*, dapat bekerja secara optimal dalam menghasilkan kesimpulan yang akurat dan dapat dipertanggungjawabkan [13].

2.3. Penerapan Algoritma *Decision Tree*

Setelah seluruh data melalui tahap *preprocessing* dan dinyatakan siap digunakan, Langkah berikutnya adalah menerapkan algoritma *Decision Tree* untuk membangun model klasifikasi yang mampu mengidentifikasi pola preferensi pengunjung wisata berdasarkan atribut-atribut yang telah dikumpulkan, seperti data demografis (usia, jenis kelamin, pendidikan, pekerjaan), jenis destinasi yang diminati, serta faktor-faktor pendukung lainnya. Tahap awal dalam proses ini adalah melakukan pembagian dataset menjadi dua bagian utama, yaitu *training set* (data latih) dan *testing set* (data uji), dengan proporsi yang umum digunakan seperti 80% untuk pelatihan model dan 20% untuk pengujian model. *Training set* digunakan untuk melatih algoritma

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15171

Decision Tree dalam mengenali pola dan hubungan antarvariabel, sedangkan *testing set* digunakan untuk mengevaluasi sejauh mana model yang dihasilkan mampu melakukan prediksi terhadap data baru yang belum pernah dilihat sebelumnya. Proses pembangunan model Decision Tree melibatkan pembuatan struktur pohon keputusan yang terdiri dari simpul (*nodes*) dan cabang (*branches*) yang merepresentasikan aturan-aturan klasifikasi berdasarkan nilai atribut. Algoritma akan memilih atribut terbaik untuk dijadikan titik pemisahan (split point) berdasarkan ukuran seperti *Gini Index*, *Entropy*, atau *Information Gain*, dengan tujuan memaksimalkan homogenitas kelas pada setiap cabang pohon. Pemisahan ini dilakukan secara rekursif hingga memenuhi kriteria berhenti, seperti kedalaman pohon yang telah ditentukan atau jumlah data dalam *node* yang terlalu sedikit untuk dibagi lebih lanjut. Melalui penerapan algoritma *Decision Tree* ini, diharapkan dapat diperoleh pemahaman yang lebih mendalam mengenai faktor-faktor yang mempengaruhi pilihan destinasi wisata pengunjung. Hasil analisis ini tidak hanya bermanfaat dalam penyusunan strategi promosi pariwisata yang lebih efektif, tetapi juga dapat menjadi dasar pengambilan keputusan bagi pihak pengelola wisata dan pemerintah daerah dalam mengembangkan potensi pariwisata di kabupaten Pekalongan secara lebih terarah dan berkelanjutan [14].

2.4. Evaluasi Model

Tahap evaluasi model merupakan Langkah krusial untuk memastikan bahwa model klasifikasi yang dibangun menggunakan algoritma *Decision Tree* memiliki performa yang optimal sebelum diimplementasikan pada aplikasi nyata. Pada tahap ini, kinerja model dianalisis menggunakan sejumlah metrik evaluasi yang umum digunakan dalam pembelajaran mesin, antara lain akurasi untuk mengukur persentase prediksi yang benar dibandingkan dengan seluruh prediksi yang dilakukan oleh model. *Precision* untuk mengukur sejauh mana prediksi positif yang dihasilkan model benar-benar sesuai dengan kondisi sebenarnya, sehingga penting untuk meminimalkan kesalahan prediksi positif (*false positive*). *Recall* untuk mengukur kemampuan model dalam menemukan seluruh data yang termasuk ke dalam kelas positif, sehingga bermanfaat untuk mengurangi kemungkinan terlewatnya data yang seharusnya terklasifikasi dengan benar (*false negative*). *F1-score* merupakan rata-rata antara *precision* dan *recall* yang memberikan gambaran seimbang mengenai kinerja model, terutama jika terdapat ketidakseimbangan jumlah data antar kelas [11]. Evaluasi dilakukan dengan menggunakan testing set yang sebelumnya telah dipisahkan dari training set, sehingga model diuji pada data yang benar-benar baru bagi algoritma. Hal ini bertujuan untuk mengetahui kemampuan model dalam melakukan generalisasi terhadap data yang tidak pernah dilihat sebelumnya. Melalui evaluasi yang komprehensif, peneliti dapat menilai apakah model yang dihasilkan sudah cukup baik untuk diimplementasikan. Hanya model dengan performa yang konsisten tinggi dan kesalahan yang minimal yang layak untuk digunakan dalam pengambilan keputusan di dunia nyata, khususnya dalam konteks memprediksi preferensi pengunjung wisata di kabupaten Pekalongan [16-17].

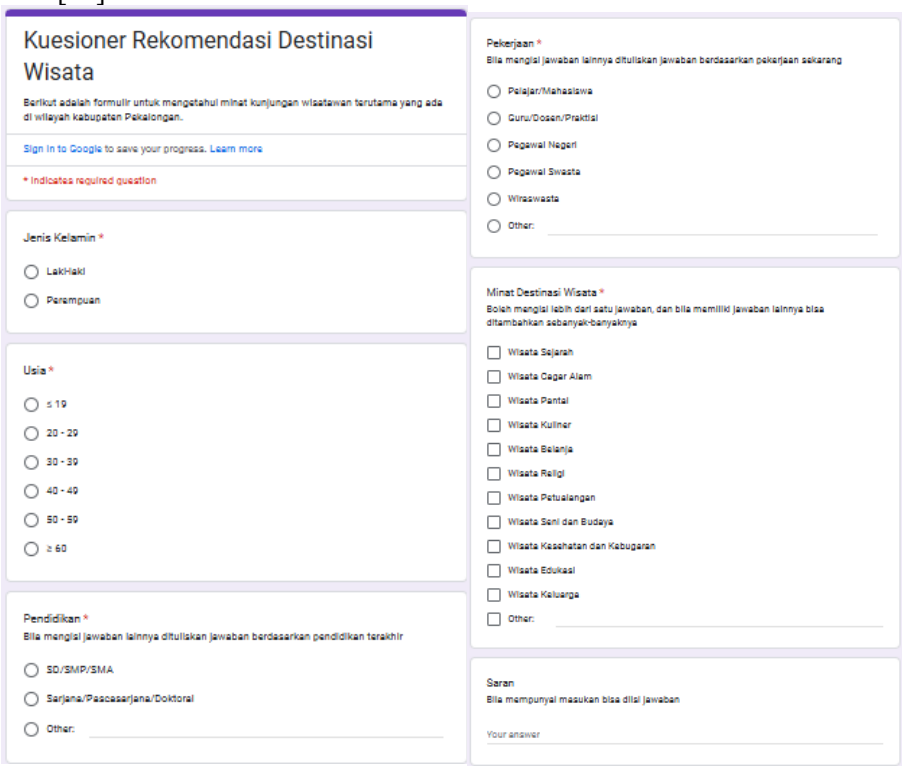
3. HASIL DAN PEMBAHASAN

3.1. Pengumpulan Data

Peneliti menjalankan proses pengumpulan data dengan pendekatan yang modern, efektif, serta adaptif terhadap perkembangan teknologi digital, yaitu melalui penyebaran kuesioner secara online. Pendekatan ini dipilih sebagai strategi yang paling efisien untuk mengumpulkan data dalam skala besar dengan waktu yang relatif singkat, sekaligus mengurangi keterbatasan yang biasanya ditemui dalam metode pengumpulan data konvensional. Selain itu, penggunaan teknologi berbasis daring juga mencerminkan upaya peneliti untuk menyesuaikan metode penelitian dengan tren digitalisasi masyarakat modern, di mana sebagian besar populasi telah terbiasa menggunakan internet dan perangkat mobile dalam aktivitas sehari-hari. Kuesioner dirancang dan disebarkan secara digital melalui platform Google Forms, yang memiliki

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15171

keunggulan dalam hal kemudahan akses, fleksibilitas, serta kompatibilitas lintas perangkat. Responden dapat mengisi kuesioner kapan saja dan di mana saja, baik melalui komputer, laptop, maupun ponsel pintar, sehingga meningkatkan kemungkinan partisipasi masyarakat secara lebih luas. Tautan kuesioner tersebut dapat diakses melalui alamat: https://forms.gle/AYPzjpuJRYs6TTy77, dan disebarluaskan secara strategis melalui berbagai kanal komunikasi digital. Untuk memastikan penyebaran yang efektif dan merata, tautan kuesioner dipromosikan melalui berbagai media sosial seperti Facebook, Instagram, dan WhatsApp, serta dibagikan di grup komunitas lokal, forum daring, dan jaringan personal yang berhubungan dengan pariwisata di Kabupaten Pekalongan. Strategi ini tidak hanya menjangkau masyarakat lokal, tetapi juga calon wisatawan dari luar daerah yang memiliki ketertarikan terhadap destinasi wisata Pekalongan. Selain itu, penggunaan media digital juga mempermudah proses pengumpulan, penyimpanan, dan pengolahan data secara otomatis, sehingga meminimalkan potensi kesalahan manusia (*human error*) yang sering terjadi dalam pengumpulan data manual [18].



Gambar 2. Kuesioner Rekomendasi Destinasi Wisata

Kuesioner dirancang secara komprehensif dan sistematis untuk mengumpulkan beragam informasi yang relevan dengan fokus penelitian. Data yang diminta mencakup atribut demografis seperti jenis kelamin, usia, tingkat pendidikan, jenis pekerjaan, serta minat dan preferensi terhadap berbagai kategori destinasi wisata. Selain itu, kuesioner juga memuat pertanyaan terbuka yang memungkinkan responden menyampaikan pendapat, saran, dan kritik terhadap fasilitas dan pengelolaan pariwisata di kabupaten Pekalongan.

Dengan strategi ini, peneliti berharap dapat membangun profil mendalam dan representatif mengenai karakteristik masyarakat serta kecenderungan minat mereka terhadap sektor wisata daerah. Data yang terkumpul nantinya tidak hanya akan berperan penting dalam tahap analisis untuk menemukan pola dan tren, tetapi juga akan menjadi masukan langsung bagi pihak terkait, seperti Dinas Pariwisata, dalam merancang strategi promosi dan pengembangan destinasi. Melalui pendekatan ini, masyarakat lokal turut dilibatkan secara aktif dalam proses penelitian, sehingga menumbuhkan rasa memiliki terhadap sektor pariwisata daerahnya.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15171

Diharapkan, hasil dari pengumpulan data ini dapat menjadi landasan kuat bagi tahapan analisis berikutnya, memberikan wawasan yang mendalam, dan mendorong terciptanya inovasi pengelolaan pariwisata yang berkelanjutan.

3.2. *Preprocessing* Data

Data yang berhasil dikumpulkan melalui kuesioner *online* berjumlah 500, yang mencakup informasi demografis, preferensi wisata, serta faktor-faktor penunjang lainnya. Seluruh data ini telah tersimpan dengan aman dalam format terstruktur yang memudahkan proses pengolahan selanjutnya. Namun, sebelum data tersebut dapat digunakan untuk analisis dan pemodelan, diperlukan serangkaian tahapan *preprocessing* data yang bertujuan untuk memastikan bahwa data memiliki kualitas, konsistensi, dan kelengkapan yang memadai [19].

Timestamp	Jenis Kelamin	Usia	Pendidikan	Pekerjaan	Minat Destinasi Wisata	Saran
12/06/2024 15:09:06	Laki-laki	20 - 29	Sarjana/Pascasarjana/D(	Pelajar/Mahasiswa	Wisata Cagar Alam	
12/06/2024 16:20:11	Laki-laki	20 - 29	SD/SMP/SMA	Pelajar/Mahasiswa	Wisata Cagar Alam, Wisata Pantai, Wisata Religi, Wisata Petualangan, Wisata Seni dan Bud	
12/06/2024 16:23:10	Laki-laki	20 - 29	SD/SMP/SMA	Pelajar/Mahasiswa	Wisata Pantai	
12/06/2024 16:27:53	Laki-laki	20 - 29	Sarjana/Pascasarjana/D(	Pelajar/Mahasiswa	Wisata Sejarah, Wisata F	Wisata merilekskan pikiran
12/06/2024 16:31:44	Perempuan	20 - 29	Sarjana/Pascasarjana/D(	Pelajar/Mahasiswa	Wisata Pantai	
12/06/2024 16:50:05	Laki-laki	20 - 29	Sarjana/Pascasarjana/D(	Pelajar/Mahasiswa	Wisata Keluarga	
12/06/2024 16:51:31	Perempuan	20 - 29	SD/SMP/SMA	Pelajar/Mahasiswa	Wisata Cagar Alam	
12/06/2024 17:35:37	Laki-laki	20 - 29	Sarjana/Pascasarjana/D(	Pelajar/Mahasiswa	Wisata Kuliner	
12/06/2024 18:52:39	Laki-laki	≤ 19	SD/SMP/SMA	Pelajar/Mahasiswa	Wisata Cagar Alam	
12/06/2024 20:33:54	Laki-laki	20 - 29	SD/SMP/SMA	Pelajar/Mahasiswa	Wisata Sejarah, Wisata (	Dalam pengelolaan wisata benar2 harus di perhatikan sampai jangk
12/06/2024 22:50:02	Laki-laki	20 - 29	SD/SMP/SMA	Pelajar/Mahasiswa	Wisata Kuliner, Wisata E	Tidak ada
19/06/2024 11:49:27	Laki-laki	20 - 29	SD/SMP/SMA	Pelajar/Mahasiswa	Wisata Pantai, Wisata Ku	Wisata kalo nggak sama pasangan nggak uasiq
19/06/2024 11:54:32	Perempuan	20 - 29	SD/SMP/SMA	Pelajar/Mahasiswa	Wisata Pantai, Wisata Kuliner, Wisata Petualangan	
19/06/2024 11:57:13	Laki-laki	20 - 29	SD/SMP/SMA	Pelajar/Mahasiswa	Wisata Sejarah, Wisata (	Rekomendasi harga tiket masuknya
19/06/2024 11:59:29	Laki-laki	20 - 29	Sarjana/Pascasarjana/D(	Wiraswasta	Wisata Sejarah, Wisata Pantai, Wisata Religi, Wisata Seni dan Budaya	
19/06/2024 12:01:24	Perempuan	20 - 29	Sarjana/Pascasarjana/D(	Pelajar/Mahasiswa	Wisata Pantai, Wisata Keluarga	
19/06/2024 12:01:51	Perempuan	20 - 29	SD/SMP/SMA	Pelajar/Mahasiswa	Wisata Sejarah, Wisata Pantai, Wisata Kuliner, Wisata Belanja, Wisata Kesehatan dan Kebug	
19/06/2024 12:02:01	Laki-laki	20 - 29	SD/SMP/SMA	Pelajar/Mahasiswa	Wisata Pantai, Wisata Petualangan	
19/06/2024 12:02:31	Laki-laki	20 - 29	SD/SMP/SMA	Pelajar/Mahasiswa	Wisata Cagar Alam, Wisata Kuliner	

Gambar 3. Jawaban Kuesioner

Dari data ini, dilakukan serangkaian tahapan ekstraksi fitur untuk memastikan bahwa data yang akan digunakan dalam proses analisis dan pemodelan lebih lanjut sudah siap dan memiliki kualitas yang optimal. Proses ini melibatkan beberapa langkah penting yang dilakukan dengan cermat untuk mengatasi berbagai potensi masalah yang mungkin ada pada data mentah. Langkah-langkah seperti berikut:

3.2.1. Memeriksa adanya data kosong atau yang hilang (*missing value*)

Langkah awal dalam proses *preprocessing* adalah melakukan pemeriksaan terhadap keberadaan data yang kosong atau hilang (*missing value*) pada setiap atribut yang terdapat dalam *dataset*. Data yang hilang dapat muncul karena berbagai faktor, salah satunya responden yang tidak mengisi seluruh pertanyaan pada kuesioner. Kehadiran data yang hilang ini, jika tidak ditangani dengan tepat, berpotensi menimbulkan masalah serius dalam proses analisis dan pemodelan, seperti menurunnya tingkat akurasi model atau munculnya bias pada hasil prediksi.

Untuk mengatasi hal tersebut, peneliti memutuskan untuk mengisi nilai yang hilang dengan menggunakan nilai modus dari masing-masing atribut. Nilai modus, yang merupakan kategori atau nilai yang paling sering muncul dalam suatu atribut, dipilih karena dianggap mampu mewakili karakteristik umum dari data tanpa menimbulkan perubahan signifikan pada distribusi nilai. Pendekatan ini dinilai lebih tepat dibandingkan menghapus baris data yang tidak lengkap, karena penghapusan dapat mengurangi jumlah data secara signifikan dan berisiko menghilangkan informasi penting.

Selain itu, penggunaan modus sebagai metode imputasi juga membantu mempertahankan konsistensi data, terutama untuk atribut yang bersifat kategorikal, seperti jenis kelamin, tingkat

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15171

pendidikan, dan jenis pekerjaan. Dengan demikian, proses pengisian nilai hilang ini tidak hanya mengembalikan kelengkapan data, tetapi juga meminimalkan potensi distorsi informasi yang dapat mempengaruhi performa algoritma *Decision Tree* pada tahap pemodelan berikutnya. Seluruh proses ini dilakukan secara hati-hati untuk memastikan bahwa kualitas dan keandalan dataset tetap terjaga sebelum melangkah ke tahap analisis yang lebih kompleks.

```
# Pembersihan Data
# Memeriksa data yang hilang
missing_data = df.isnull().sum()
print(missing_data)
```

Gambar 4. Pembersihan Data dan Cek Data Hilang

```
# Mengisi nilai yang hilang dengan nilai modus
df['usia'].fillna(df['usia'].mode()[0], inplace=True)
df['jenis_kelamin'].fillna(df['jenis_kelamin'].mode()[0], inplace=True)
df['pendidikan'].fillna(df['pendidikan'].mode()[0], inplace=True)
df['pekerjaan'].fillna(df['pekerjaan'].mode()[0], inplace=True)
```

Gambar 5. Isi Nilai Kosong

3.2.2. Memeriksa dan menangani data yang duplikat

Tahap berikutnya dalam proses *preprocessing* adalah melakukan pemeriksaan terhadap data yang duplikat. Data duplikat adalah entri yang tercatat lebih dari satu kali dalam *dataset* dengan nilai atribut yang sama persis. Duplikasi data dapat terjadi karena berbagai alasan, salah satunya responden yang mengisi kuesioner lebih dari sekali, kesalahan sistem saat proses penyimpanan data, atau terjadinya redudansi akibat penggabungan beberapa sumber data yang berbeda tanpa proses penyaringan yang memadai. Keberadaan data duplikat dalam *dataset* dapat menimbulkan sejumlah masalah, diantaranya adalah bias dalam analisis statistik dan *overfitting* pada model ini. Bias tersebut terjadi karena informasi yang sama dihitung lebih dari sekali, sehingga dapat memberikan gambaran yang keliru mengenai pola sebenarnya dalam data. Selain itu, data yang berulang juga dapat meningkatkan kompleksitas perhitungan dan memperlambat kinerja sistem, khususnya pada dataset berukuran besar.

Untuk menghindari dampak negatif tersebut, peneliti melakukan identifikasi data duplikat menggunakan fungsi deteksi bawaan pada perangkat lunak pengolahan data yang digunakan. Setiap baris data dibandingkan secara menyeluruh pada seluruh atribut untuk memastikan tidak ada nilai yang sama persis pada lebih dari satu entri. Setelah duplikasi terdeteksi, langkah selanjutnya adalah menghapus entri yang berulang sehingga hanya tersisa data yang unik, valid, dan merepresentasikan responden secara akurat. Proses penghapusan ini dilakukan dengan hati-hati untuk memastikan bahwa tidak ada data penting yang terhapus secara keliru. Dengan demikian, *dataset* yang digunakan dalam penelitian ini menjadi lebih bersih, akurat, dan representatif, sehingga dapat memberikan hasil analisis yang lebih valid dan meningkatkan performa algoritma *Decision Tree* pada tahap pemodelan selanjutnya.

```
# Memeriksa duplikasi data
duplicates = df.duplicated().sum()
# Menampilkan panjang DataFrame sebelum menghapus duplikat
print(f'Panjang DataFrame sebelum menghapus duplikat: {df.shape[0]}')
# Menghapus data duplikat
df = df.drop_duplicates()
# Menampilkan panjang DataFrame setelah menghapus duplikat
print(f'Panjang DataFrame setelah menghapus duplikat: {df.shape[0]}')
```

Gambar 6. Cek Duplikasi Data

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15171

3.2.3. Validasi data pada kolom

Langkah selanjutnya dalam proses *preprocessing* adalah melakukan validasi terhadap nilai pada kolom tertentu, khususnya kolom jenis_kelamin. Proses ini bertujuan untuk memastikan bahwa seluruh data yang tercatat pada kolom tersebut sesuai dengan kategori yang telah ditetapkan sebelumnya, yaitu hanya boleh berisi dua nilai yang valid, yaitu "Laki-laki" atau "Perempuan". Standarisasi kategori ini sangat penting untuk menjaga konsistensi data, menghindari perbedaan penulisan, dan memastikan bahwa proses analisis dapat berjalan tanpa gangguan akibat data yang tidak sesuai format.

Pada tahap ini, sistem akan memeriksa setiap entri pada kolom jenis_kelamin dan membandingkan dengan daftar kategori valid yang telah ditentukan. Apabila ditemukan entri yang tidak sesuai, salah satunya nilai kosong, maka entri tersebut akan diidentifikasi sebagai data yang tidak valid. Setelah proses identifikasi, langkah selanjutnya adalah melakukan penanganan terhadap data yang tidak valid. Penanganan ini dapat dilakukan dengan beberapa cara, seperti memperbaiki nilai berdasarkan data asli, mengganti nilai yang salah dengan kategori yang benar sesuai hasil verifikasi, atau menghapus data tersebut apabila perbaikannya tidak memungkinkan. Pendekatan yang dipilih bergantung pada tingkat kesalahan dan kelengkapan informasi tambahan yang dimiliki oleh peneliti.

Dengan menerapkan proses validasi ini, kualitas data menjadi lebih terjamin, sehingga analisis yang dilakukan pada tahap berikutnya dapat menghasilkan temuan yang lebih akurat dan dapat dipertanggungjawabkan. Selain itu, langkah ini juga berkontribusi pada peningkatan kinerja algoritma *Decision Tree*, karena model akan dilatih menggunakan data yang konsisten, bersih, dan sesuai dengan standar kategori yang telah ditentukan.

```
# Memeriksa duplikasi data
duplicates = df.duplicated().sum()
# Menampilkan panjang DataFrame sebelum menghapus duplikat
print(f'Panjang DataFrame sebelum menghapus duplikat: {df.shape[0]}')
# Menghapus data duplikat
df = df.drop_duplicates()
# Menampilkan panjang DataFrame setelah menghapus duplikat
print(f'Panjang DataFrame setelah menghapus duplikat: {df.shape[0]}')
```

Gambar 7. Cek Nilai Tidak Valid

3.2.4. Transformasi data pada kolom

Tahap selanjutnya dalam proses *preprocessing* adalah melakukan transformasi data pada kolom kategori_wisata agar dapat digunakan oleh algoritma ini. Kolom kategori_wisata pada *dataset* awal berisi nilai dalam bentuk teks (*string*), seperti "Wisata Alam", "wisata Religi", "Wisata Kuliner, atau "Wisata Sejarah". Meskipun mudah dipahami oleh manusia, data dalam bentuk teks ini tidak dapat diproses secara langsung oleh Sebagian besar algoritma pemodelan, khususnya algoritma seperti *Decision Tree* yang memerlukan dalam format numerik untuk melakukan perhitungan dan pembentukan model prediksi.

Untuk mengatasi hal tersebut, digunakan teknik Label *Encoding* yang diimplementasikan melalui metode *LabelEncoder*. Proses ini bekerja dengan cara memberikan representasi numerik unik untuk setiap kategori pada kolom kategori_wisata. Sebagai contoh, "Wisata Alam" dapat diubah menjadi angka 0, "Wisata Religi" menjadi 1, "Wisata Kuliner" menjadi 2, dan "Wisata Sejarah" menjadi 3. Angka-angka ini tidak merepresentasikan urutan atau peringkat, melainkan hanya sebagai kode identifikasi yang memudahkan komputer dalam mengolah data.

Keuntungan dari penggunaan *LabelEncoder* adalah proses transformasi dapat dilakukan secara otomatis dan konsisten, sehingga setiap kategori teks akan selalu dikonversi ke nilai numerik yang sama selama proses *training* dan *testing*. Hal ini memastikan bahwa model tidak mengalami kebingungan dalam mengenali kategori yang sama pada data baru.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15171

Selain itu, transformasi ini juga dapat mengurangi potensi kesalahan yang muncul akibat penulisan kategori yang tidak konsisten. Sebelum dilakukan *encoding*, kategori terlebih dahulu divalidasi dan dibersihkan sehingga tidak ada variasi penulisan yang menyebabkan pembentukan label ganda. Dengan menerapkan Langkah transformasi ini, *dataset* menjadi lebih siap untuk diproses oleh algoritma *Decision Tree* maupun model lain yang berbasis perhitungan numerik. Hasil akhirnya adalah data yang terstruktur dengan baik, konsisten, dan kompatibel dengan berbagai metode analisis maupun prediksi yang akan dilakukan pada tahap pemodelan.

```
# Mengonversi data kategori menjadi data numerik
le = LabelEncoder()
df['kategori_wisata'] = le.fit_transform(df['kategori_wisata'])
```

Gambar 8. Konversi Kategori Menjadi Numerik

3.2.5. Mengganti atribut pada kolom

Setelah tahap transformasi kategori_wisata menjadi data numerik, proses *preprocessing* selanjutnya adalah mengubah atribut-atribut kategorikal lainnya, seperti pendidikan, pekerjaan, dan jenis_kelamin, menjadi format yang dapat diproses secara optimal oleh algoritma ini. Ketiga atribut ini pada awalnya berbentuk data kategorikal non-numerik, seperti kolom pendidikan berisi kategori "SMA", "Diploma", "Sarjana", atau "Pascasarjana"; kolom pekerjaan berisi kategori "Pelajar", "karyawan Swasta", "PNS", atau "Wirausaha"; dan kolom jenis_kelamin berisi kategori "Laki-laki" atau "Perempuan".

Untuk menghindari kesalahan interpretasi oleh model, atribut-atribut ini perlu dikonversi menggunakan metode *One-Hot Encoding*. Metode ini bekerja dengan cara membuat variabel *dummy* untuk setiap kategori unik yang ada dalam kolom. Setiap variabel *dummy* berbentuk kolom baru yang bernilai 1 jika data tersebut termasuk dalam kategori tersebut, dan bernilai 0 jika tidak. Seperti, kolom pendidikan akan dihasilkan kolom tambahan yaitu pendidikan_SMA, pendidikan _Diploma, pendidikan_Sarjana, dan pendidikan_Pascasarjana. Jika seseorang memiliki pendidikan "Sarjana", maka kolom pendidikan_Sarjana akan bernilai 1, sedangkan kolom lainnya bernilai 0.

Keunggulan dari penggunaan *One-Hot Encoding* adalah dapat menghindari asumsi urutan atau hierarki pada data kategorikal. Berbeda dengan Label *Encoding* yang menghasilkan nilai numerik tunggal, *One-Hot Encoding* tidak mengisyaratkan bahwa kategori dengan angka lebih tinggi memiliki tingkat yang lebih besar atau lebih penting dibandingkan yang lain. Hal ini penting untuk menjaga netralitas data sehingga model tidak bias dalam memproses kategori.

Proses ini juga meningkatkan kemampuan model untuk mengidentifikasi pola unik pada setiap kategori secara independen. Seperti model dapat mengenali bahwa responden dengan pekerjaan "Wirausaha" memiliki kecenderungan preferensi wisata yang berbeda dibandingkan "Pelajar" atau "PNS" tanpa terpengaruh oleh nilai numerik yang melekat. Dengan menerapkan *One-Hot Encoding*, *dataset* menjadi lebih terstruktur dan siap digunakan pada proses pelatihan model, khususnya algoritma *Decision Tree* yang dapat memanfaatkan informasi biner ini secara optimal dalam proses pemisahan node dan penentuan aturan klasifikasi. Langkah ini juga membantu memaksimalkan akurasi prediksi serta menjaga integritas analisis data, karena semua kategori direpresentasikan secara adil dan setara.

```
# One-hot encoding untuk kolom usia, pendidikan, dan pekerjaan
df = pd.get_dummies(df, columns=['usia', 'pendidikan', 'pekerjaan', 'jenis_kelamin'], drop_first=True)
```

Gambar 9. Mengubah Atribut Menjadi Variabel Dummy

3.2.6. Menangani teknik *oversampling*

Salah satu tantangan yang sering muncul dalam analisis data, khususnya pada permasalahan klasifikasi, adalah ketidakseimbangan distribusi kelas (*class imbalance*) pada

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15171

dataset. Ketidakseimbangan ini terjadi ketika jumlah data pada satu kelas jauh lebih banyak dibandingkan kelas lainnya. Dalam konteks penelitian ini, kondisi tersebut dapat menyebabkan pemodelan menjadi bias terhadap kelas mayoritas, sehingga lebih sering memprediksi kategori tersebut dan mengabaikan keberadaan kelas minoritas. Akibatnya, meskipun model tampak memiliki akurasi keseluruhan yang tinggi, kinerjanya dalam memprediksi kelas minoritas menjadi buruk.

Untuk mengatasi permasalahan ini, dilakukan teknik *oversampling* pada kelas minoritas. *Oversampling* adalah metode yang bertujuan menambah jumlah sampel pada kelas yang jumlah datanya lebih sedikit hingga setara dengan kelas mayoritas. Dalam penelitian ini, digunakan pendekatan resampling pada data minoritas. Proses ini melibatkan penggandaan sampel yang ada pada kelas minoritas secara acak, sehingga jumlah total sampel di kedua kelas menjadi seimbang.

Langkah ini dilakukan dengan hati-hati untuk memastikan bahwa proses penggandaan data tidak menimbulkan *overfitting*, yaitu kondisi di mana model terlalu menghafal data pelatihan dan gagal menggeneralisasi pada data baru. Oleh karena itu, metode resampling yang dipilih mempertahankan distribusi asli data sembari memastikan bahwa setiap sampel tambahan tetap merepresentasikan karakteristik kelas minoritas secara konsisten.

```
# Menangani kelas tidak seimbang (jika diperlukan)
# Contoh menggunakan teknik oversampling
X = df.drop(columns=['kategori_wisata'])
y = df['kategori_wisata']
X_resampled, y_resampled = resample(X, y, random_state=42)
```

Gambar 10. Teknik Oversampling

3.2.7. Tahap akhir *preprocessing* data

Tahap akhir dari proses *preprocessing* data merupakan langkah krusial yang menandai berakhirnya seluruh kegiatan persiapan data sebelum digunakan dalam tahap pemodelan. Setelah seluruh proses sebelumnya, mulai dari pembersihan data, penanganan *missing value*, penghapusan duplikasi, validasi nilai pada kolom, transformasi data menjadi format numerik, pembuatan variabel *dummy*, hingga penanganan ketidakseimbangan kelas dengan teknik *oversampling* selesai dilakukan, maka semua hasil transformasi tersebut dikompilasi menjadi satu kesatuan dataset yang utuh dan terstruktur dengan baik.

Dataset hasil akhir ini kemudian disimpan dalam sebuah tabel baru yang diberi nama *new_master_training*. Tabel ini menjadi versi final dari data pelatihan yang sudah bebas dari ketidakkonsistenan, memiliki format yang sesuai untuk analisis, dan memuat fitur-fitur yang telah diolah sedemikian rupa agar kompatibel dengan algoritma yang akan digunakan. Penyimpanan dalam tabel baru juga bertujuan untuk menjaga integritas data mentah (raw data) agar tetap utuh.

Dengan selesainya tahap ini, peneliti telah memiliki dataset yang siap pakai untuk implementasi model. Data yang ada kini telah memenuhi kriteria kualitas, kelengkapan, dan konsistensi yang diperlukan untuk menghasilkan pemodelan yang akurat dan andal. Tahap ini juga menjadi penutup dari fase data *preparation*, sekaligus menjadi jembatan menuju fase berikutnya, yaitu penerapan algoritma *Decision Tree* dan evaluasi kinerjanya. Dengan *dataset* seperti berikut:

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15171

A	B	C	D	E	F
id	jenis_kela	usia	pekerjaan	pendidikan	kategori_wisata
1	Laki-laki	20 - 29	Pelajar/Mahasiswa	Sarjana/Pascasarjana/Doktoral	Wisata Belanja
2	Laki-laki	20 - 29	Pelajar/Mahasiswa	Sarjana/Pascasarjana/Doktoral	Wisata Cagar Alam
3	Perempua	50 - 59	Guru/Dosen/Praktisi	Sarjana/Pascasarjana/Doctoral	Wisata Kesehatan dan Kebugaran
4	Perempua	50 - 59	Guru/Dosen/Praktisi	Sarjana/Pascasarjana/Doctoral	Wisata Edukasi
5	Laki-laki	20 - 29	Pelajar/Mahasiswa	SD/SMP/SMA	Wisata Seni dan Budaya
6	Laki-laki	20 - 29	Pelajar/Mahasiswa	SD/SMP/SMA	Wisata Seni dan Budaya
7	Laki-laki	20 - 29	Pelajar/Mahasiswa	Sarjana/Pascasarjana/Doktoral	Wisata Edukasi
8	Laki-laki	20 - 29	Pelajar/Mahasiswa	SD/SMP/SMA	Wisata Seni dan Budaya
9	Laki-laki	20 - 29	Guru/Dosen/Praktisi	Sarjana/Pascasarjana/Doctoral	Wisata Kuliner
10	Laki-laki	20 - 29	Pelajar/Mahasiswa	Sarjana/Pascasarjana/Doktoral	Wisata Keluarga
11	Laki-laki	20 - 29	Pelajar/Mahasiswa	Sarjana/Pascasarjana/Doktoral	Wisata Keluarga
12	Laki-laki	20 - 29	Pelajar/Mahasiswa	Sarjana/Pascasarjana/Doktoral	Wisata Kesehatan dan Kebugaran
13	Laki-laki	20 - 29	Pelajar/Mahasiswa	SD/SMP/SMA	Wisata Seni dan Budaya
14	Laki-laki	20 - 29	Pelajar/Mahasiswa	Sarjana/Pascasarjana/Doktoral	Wisata Kesehatan dan Kebugaran
15	Laki-laki	20 - 29	Wiraswasta	Sarjana/Pascasarjana/Doktoral	Wisata Cagar Alam
16	Laki-laki	20 - 29	Guru/Dosen/Praktisi	Sarjana/Pascasarjana/Doctoral	Wisata Religi
17	Laki-laki	20 - 29	Guru/Dosen/Praktisi	Sarjana/Pascasarjana/Doctoral	Wisata Petualangan
18	Laki-laki	20 - 29	Wiraswasta	Sarjana/Pascasarjana/Doktoral	Wisata Pantai

Gambar 11. Hasil Preprocessing Data

3.3. Penerapan Algoritma *Decision Tree*

Pada tahapan ini, peneliti memasuki fase krusial dalam siklus pengembangan sistem rekomendasi, yaitu penerapan algoritma ini untuk membangun model prediksi. Proses dimulai dengan membagi *dataset* yang telah melalui tahapan *preprocessing* menjadi dua bagian utama, yakni set data latih (X_train, Y_train) dan set data uji (X_test, Y_test). Pembagian ini memiliki peranan yang sangat penting dalam memastikan kualitas model, karena memungkinkan model untuk mempelajari pola-pola dari data pada tahap pelatihan (*training*), sementara bagian data lainnya sengaja dipisahkan untuk digunakan dalam tahap pengujian (*testing*) guna mengukur kemampuan generalisasi model terhadap data baru yang belum pernah dilihat sebelumnya.

Selanjutnya, peneliti memilih dan memanfaatkan model *RandomForestClassifier* sebagai implementasi dari metode *Decision Tree*. Pemilihan algoritma ini bukan tanpa alasan, *Random Forest* dikenal memiliki kemampuan yang baik dalam menangani *dataset* dengan banyak variabel prediktor, mampu meminimalkan risiko *overfitting* melalui penggunaan teknik *ensemble learning*, serta memiliki ketahanan terhadap data yang mengandung *noise*. Algoritma ini bekerja dengan membangun sekumpulan pohon keputusan (*decision trees*) pada berbagai subset data latih, kemudian menggabungkan hasil prediksi dari seluruh pohon tersebut untuk menghasilkan keputusan akhir yang lebih stabil dan akurat.

Data yang digunakan pada tahap pelatihan telah melalui proses resampling dengan tujuan menyeimbangkan distribusi kelas dalam variabel target. Langkah ini sangat penting, mengingat data yang tidak seimbang dapat menyebabkan model cenderung memberikan prediksi yang bias ke arah kelas mayoritas, sehingga mengabaikan kelas minoritas yang justru seringkali menjadi bagian penting dalam rekomendasi wisata. Dengan proses resampling, jumlah sampel pada setiap kelas dibuat relatif seimbang, sehingga algoritma dapat mempelajari karakteristik masing-masing kategori wisata secara proporsional.

Melalui penerapan strategi ini, diharapkan model *Decision Tree* yang dihasilkan tidak hanya mampu memberikan tingkat akurasi yang tinggi, tetapi juga bersifat adil (*fair*) dalam melakukan klasifikasi terhadap semua kategori wisata yang ada. Pada akhirnya, model ini akan menjadi inti dari sistem rekomendasi yang dapat memberikan saran destinasi wisata kepada pengguna berdasarkan preferensi dan karakteristik mereka, dengan tingkat ketetapan yang optimal.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15171

```
# Membagi data menjadi data latih dan data uji
X_train, X_test, y_train, y_test = train_test_split(X_resampled, y_resampled, test_size=0.2, random_state=42)

# Melatih model menggunakan RandomForestClassifier dengan pengaturan yang lebih hati-hati
clf = RandomForestClassifier(n_estimators=100, max_depth=10, random_state=42)
clf.fit(X_train, y_train)
```

Gambar 12. Membagi Data Latih dan Data Uji

Setelah proses pelatihan selesai, langkah selanjutnya adalah melakukan evaluasi model.

3.4. Evaluasi Model

Evaluasi dilakukan dengan menggunakan metrik akurasi serta laporan klasifikasi yang mencakup metrik presisi, *recall*, dan *f1-score*. Namun, peneliti menyadari bahwa evaluasi yang dilakukan pada data yang sama dengan data latih dapat memberikan hasil yang terlalu optimis, karena model sudah terbiasa dengan pola pada data. Oleh karena itu, meskipun metrik memberikan gambaran awal tentang kinerja model, evaluasi lebih lanjut dengan data yang belum pernah dilihat model sebelumnya sangat penting untuk mengukur performa sebenarnya.

```
# Evaluasi model pada data uji
y_pred = clf.predict(X_resampled)
accuracy = accuracy_score(y_resampled, y_pred)
print(f'Akurasi: {accuracy * 100:.2f}%')
print(classification_report(y_resampled, y_pred))
```

Gambar 13. Evaluasi Model

Hasil pengujian model yang dilakukan menunjukkan bahwa model yang diterapkan memiliki tingkat akurasi sebesar 76,92%. Menunjukkan model mampu memperdiksi dengan tingkat ketetapan yang tinggi, memberikan prediksi yang dapat diandalkan bagi pengguna dalam konteks rekomendasi wisata yang diharapkan. Dimana model Ini mampu menangkap pola yang lebih kompleks dari data dan mengurangi risiko overfitting pada penggabungan beberapa *Decision Trees*. Meski demikian, masih ada ruang untuk meningkatkan performa lebih lanjut dengan mengoptimalkan dalam menerapkan teknik resampling pada data untuk meningkatkan kinerja pada kelas yang mungkin tidak seimbang.

```
jenis_kelamin      0
usia               1
pekerjaan          0
pendidikan         0
kategori_wisata    0
dtype: int64
Panjang DataFrame sebelum menghapus duplikat: 61
Panjang DataFrame setelah menghapus duplikat: 61
Akurasi pada data uji: 76.92%
              precision    recall  f1-score   support

           0       0.50      0.67      0.57         3
           1       1.00      0.67      0.80         3
           2       0.80      0.80      0.80         5
           3       1.00      1.00      1.00         2

    accuracy                           0.77        13
   macro avg       0.82      0.78      0.79        13
weighted avg       0.81      0.77      0.78        13
```

Gambar 14. Hasil Akurasi

3.5. Analisa Hasil Model *Decision Tree*

3.5.1. Evaluasi kinerja model

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15171

Setelah model *Decision Tree* dibangun dan diuji menggunakan data hasil preprocessing. Evaluasi terhadap data uji menunjukkan bahwa model mencapai akurasi sebesar 76,92%. Tingkat akurasi ini menunjukkan bahwa model mampu memprediksi kategori wisata hingga hampir 77 persen secara tepat pada data uji, yang menandakan model memiliki kemampuan klasifikasi yang cukup baik dalam konteks preferensi wisata di Kabupaten Pekalongan.

3.5.2. Interpretasi pola keputusan model

Pada penelitian ini, hasil model menunjukkan bahwa beberapa atribut demografis memiliki peran penting dalam menentukan preferensi wisata pengunjung, yaitu:

- Usia merupakan salah satu atribut yang terdapat nilai "1", sehingga menunjukkan bahwa kelompok usia berbeda cenderung memilih jenis wisata yang berbeda pula.
- Tingkat pendidikan juga mempengaruhi pilihan tertentu, seperti pengunjung dengan pendidikan tinggi cenderung mengarah ke kategori wisata edukatif atau sejarah.
- Jenis kelamin dan pekerjaan memberikan pola segmen yang berbeda, namun pengaruhnya relatif lebih rendah dibanding usia dan Pendidikan.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat ditarik beberapa poin penting yang menggambarkan keberhasilan dan kontribusi penelitian ini terhadap analisis preferensi pengunjung wisata di Kabupaten Pekalongan.

4.1. Pengumpulan data efektif

Penelitian ini berhasil mengimplementasikan metode pengumpulan data yang efisien melalui penyebaran kuesioner online menggunakan platform *Google Forms*. Pendekatan ini memungkinkan peneliti menjangkau responden secara luas tanpa batasan geografis serta mempermudah proses akuisisi data secara cepat dan terstruktur. Sebanyak 500 responden berpartisipasi dalam penelitian ini, yang terdiri dari masyarakat lokal dengan latar belakang demografis yang beragam, mencakup atribut seperti jenis kelamin, usia, tingkat pendidikan, dan pekerjaan. Data yang diperoleh sangat relevan untuk mengidentifikasi pola dan preferensi pengunjung terhadap berbagai kategori wisata yang ada di Kabupaten Pekalongan, seperti wisata alam, budaya, kuliner, religi, hingga edukasi.

4.2. Preprocessing data yang komprehensif

Tahapan *preprocessing* dilakukan dengan cermat untuk memastikan kualitas data yang digunakan dalam proses pemodelan berada dalam kondisi terbaik. Langkah-langkah yang dilakukan meliputi pemeriksaan dan pengisian nilai yang hilang (*missing value imputation*) menggunakan metode modus, penghapusan data duplikat agar tidak terjadi bias, serta validasi nilai kategori pada atribut jenis kelamin untuk menjaga konsistensi data. Selanjutnya, dilakukan proses *Label Encoding* dan *One-Hot Encoding* untuk mengubah data kategorikal menjadi bentuk numerik yang dapat dipahami oleh algoritma *machine learning*. Selain itu, teknik *oversampling* diterapkan guna menyeimbangkan distribusi kelas yang tidak merata, sehingga model dapat belajar secara optimal dan memberikan hasil prediksi yang lebih akurat terhadap setiap kategori wisata.

4.3. Implementasi algoritma *Decision Tree*

Pada tahap implementasi, algoritma *Decision Tree* digunakan sebagai metode klasifikasi utama karena kemampuannya dalam menghasilkan model yang mudah diinterpretasikan serta efektif dalam menangani data dengan berbagai tipe atribut. Proses pelatihan model dilakukan dengan membagi data menjadi *training set* dan *testing set*. Hasil evaluasi menunjukkan tingkat akurasi sebesar 76,92%, yang menandakan bahwa model memiliki performa yang cukup baik dalam memprediksi preferensi wisata pengunjung. Model ini mampu mengidentifikasi atribut demografis yang paling berpengaruh terhadap pilihan wisata, seperti hubungan antara tingkat pendidikan dan minat terhadap wisata edukatif, atau pengaruh usia terhadap kecenderungan memilih wisata petualangan. Dengan demikian, hasil penelitian ini dapat menjadi dasar bagi

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15171

pengambil kebijakan dan pengelola destinasi wisata untuk menyusun strategi promosi dan pengembangan wisata yang lebih tepat sasaran dan berbasis data.

DAFTAR PUSTAKA

[1] M. Rizaludin, H. Hidayat, dan M. Hakim, "Analisis Tren Produk UMKM Di Desa Bugangan Menggunakan Algoritma Iterative Dichotomiser 3 (ID3)," Teknomatika, vol. 14, no. 2, pp. 21-8, 2024.

[2] U.L. Mahmudah, "Latar Belakang Unesco Menetapkan Pekalongan Sebagai Jejaring Kota Kreatif Tahun 2014," Jurusan Ilmu Hubungan Internasional, Fakultas Ilmu Sosial dan Ilmu Politik, Universitas Pembangunan Nasional Veteran Yogyakarta, Yogyakarta, Indonesia, 2017.

[3] K. Hyeyoung, L. Suyeon, P. Yoonseo, dan C. Anna, "A Survey of Recommendation Models, Techniques, and Application Fields," Electronics, vol.11, no.141, pp. 1-48, 2022.

[4] Tim Komunikasi Publik, "Kunjungan Wisata di Kota Pekalongan Semester Pertama Capai 35,38 Persen, Pantai Pasir Kencana Masih Jadi Primadona [internet]," Pemerintah Kota Pekalongan, 2025 [Diakses pada 8 Agustus 2025], dari https://pekalongankota.go.id/berita/kunjungan-wisata-di-kota-pekalongan-semester-pertama-capai-3538-persen-pantai-pasir-kencana-masih-jadi-primadona.html.

[5] A.K. Laturiuw dan Y.A. Singgalen, "Sentiment Analysis of Raja Ampat Tourism Destination Using CRISP-DM: SVM, NBC, DT, and k-NN Algorithm," J Inf Syst Informatics, vol. 5, no. 2, pp. 518-35, 2023.

[6] M. Rizaludin dan F. Fikriah, "Prediksi Perilaku Pelanggan Pada Produk UMKM Batik Dengan Menggunakan Algoritma Decision Tree," Teknomatika, vo. 13, no. 2, pp. 8-16, 2023.

[7] Z. Hidayatullah, K. Anwariyah, Y. Sa'adati, dan F. Syuhada, "Sistem Pakar Dengan Metode Decision Tree Untuk Rekomendasi Wisata Di Kabupaten Lombok Timur," SainsTech Innovation Journal, vol. 5, no. 2, pp. 219-227, 2022.

[8] H.K.N Iman, N. Latifah, S. Supriyono, dan F. Nugraha, "Pemanfaatan Metode Decision Tree dengan Algoritma C4.5 Untuk Prediksi Potensi Kunjungan Wisatawan." Jurnal Sistem Komputer dan Informatika (JSON), vol. 5, no. 3, pp. 611, 2024.

[9] A.H. Wijaya, A. Vandaningtyas, dan R.K. Hapsari, "Penentuan Destinasi Wisata Favorit di Yogyakarta dengan Menggunakan Algoritma Decision Tree," Jurnal Riset Inovasi Bidang Informatika dan Pendidikan Informatika, vol. 4, no. 2, pp. 102-110, 2023.

[10] R. Oktafiani dan R. Rianto, "Perbandingan Algoritma Support Vector Machine (SVM) dan Decision Tree Untuk Sistem Rekomendasi Tempat Wisata," Jurnal Nasional Teknologi dan Sistem Informasi, vol. 9, no. 2, pp. 113-121, 2023.

[11] F.J. Pa'o dan H. Hendry, "Decision Tree dalam Menganalisis Data Pengunjung Wisata Danau Poso untuk Pengambilan Keputusan," Jurnal Sistem Komputer dan Informatika (JSON), vol. 2, no. 3, 2021.

[12] M. Rizaludin, N. Hadian, dan N. Hayati, "Integrating Augmented Reality with C4.5 Algorithm to Enhance Tourism Experience in Pekalongan," Jurnal Ilmu Pengetahuan dan Teknologi Komputer (JITK), vol. 10, no. 4, pp. 970-979, 2025.

[13] M. Rizaludin, M.R. Fanani, H. Hidayat, dan N. Hadian, "Perbandingan Teknik Data Mining Untuk Prediksi Penjualan Produk UMKM Batik Karangdowo," Teknomatika, vol. 14, no. 1, pp.32-9, 2024.

[14] I. Zulfahmi, H. Syahputra, S.I. Naibaho, M.A. Maulana, dan E.P. Sinaga, "Perbandingan Algoritma Support Vector Machine (SVM) dan Decision Tree Untuk Deteksi Tingkat Depresi Mahasiswa," Bina Insa Ict J. vol. 10, no. 1, pp. 52, 2023.

[15] A.H. Wijaya, A. Vandaningtyas, dan K. Hapsari, "Penentuan Destinasi Wisata Favorit di Yogyakarta Dengan Menggunakan Algoritm Decision Tree," vol. 4, no. 2, pp. 102-10, 2023.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15171

[16] N. Hadian, M. Rizaludin, dan H. Hidayat, "Pengembangan Aplikasi Rekomendasi Wisata Berbasis Augmented Reality Di Kabupaten Pekalongan," Jurnal Teknologi Informasi (Djtechno), vol. 6, no. 1, pp. 180-192, 2025.
[17] M.R. Fanani, M. Rizaludin, dan H. Hidayat, "Sosialisasi Produksi Penjualan Produk UMKM Yang Diminati Konsumen Menggunakan Algoritma C4.5 di Desa Karangdowo Kec. Kedungwuni Kab. Pekalongan," Jurnal Pengabdian Kepada Masyarakat (Sorot), vol. 3, no. 1, pp. 22-25, 2024.
[18] M. Rizaludin, H. Hidayat, dan F.K. Fikriah, "Perancangan Sistem Informasi Perpustakaan Berbasis Web Pada ITSNU Pekalongan," Bulletin of Information Technology (BIT), vol. 3, no. 4, pp. 350-354, 2022.
[19] M. Rizaludin, F.K. Fikriah, dan H. Hidayat, "Pengenalan Augmented Reality (AR) Sebagai Media Pembelajaran di SMK NU Kesesi," Jurnal Pengabdian Masyarakat Tekno, vol. 3, no. 2, pp. 77-83, 2022.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i1.15171