PERBANDINGAN METODE *RANDOM FOREST* DAN *NAIVE BAYES* DALAM KLASIFIKASI PREFERENSI DESTINASI WISATA BERBASIS ALAM

Feriantano Sundang Pranata, Indra Saputra
Universitas Negeri Padang
Jalan Prof. Dr. Hamka, Air Tawar Padang, Sumatera Barat, Indonesia
feriantano@unp.ac.id

ABSTRAK

Prediksi preferensi wisatawan pada destinasi berbasis alam merupakan isu penting dalam analitik pariwisata karena perbedaan minat pengunjung memengaruhi segmentasi, promosi, perencanaan layanan, dan pengembangan destinasi. Tantangan utama penelitian ini adalah menentukan model klasifikasi yang mampu merepresentasikan pola preferensi wisatawan secara akurat. Penelitian ini membandingkan *Random Forest* dan *Naive Bayes* untuk mengklasifikasikan preferensi wisatawan terhadap destinasi pegunungan dan pantai. Perbandingan kedua metode diperlukan karena data preferensi wisatawan memuat atribut demografis dan perilaku yang berpotensi membentuk pola klasifikasi berbeda, sehingga perlu diuji menggunakan algoritma dengan karakteristik pembelajaran yang berbeda. *Random Forest* dipilih karena mampu menangkap interaksi fitur kompleks melalui mekanisme *ensemble* berbasis *decision tree*, sedangkan *Naive Bayes* digunakan sebagai pembanding karena sederhana, efisien, dan relevan sebagai *baseline* probabilistik dalam tugas klasifikasi. Penelitian ini menggunakan desain kuantitatif berbasis *machine learning* dengan *dataset Mountains vs. Beaches Preferences* dari *Kaggle*. Tahapan penelitian meliputi eksplorasi data, *praproses*, pembangunan model, dan evaluasi komparatif. Data dibagi dengan proporsi 80:20. Evaluasi dilakukan menggunakan *confusion matrix* serta metrik akurasi, presisi, *recall*, dan *F1-score*. Hasil menunjukkan *Random Forest* unggul atas *Naive Bayes* pada seluruh metrik. *Random Forest* memperoleh akurasi 0,994596, presisi 0,994626, *recall* 0,994596, dan *F1-score* 0,994604, sedangkan *Naive Bayes* memperoleh akurasi 0,919226, presisi 0,930133, *recall* 0,919226, dan *F1-score* 0,921664. Temuan ini menunjukkan bahwa *Random Forest* lebih sesuai untuk *dataset* yang dianalisis *dan machine learning* dapat mendukung pengambilan keputusan manajerial dalam pengelolaan destinasi wisata alam.

Kata kunci : *Machine Learning, Random Forest, Naive Bayes, Preferensi Wisata, Destinasi Wisata Alam*

1. PENDAHULUAN

Pariwisata merupakan salah satu sektor yang berkontribusi terhadap pertumbuhan ekonomi, mobilitas masyarakat, serta pertukaran sosial dan budaya [1]. Di antara berbagai bentuk pariwisata, wisata berbasis alam terus menunjukkan daya tarik yang tinggi karena menawarkan pengalaman rekreasi, relaksasi, dan petualangan di lingkungan terbuka [2]. Destinasi pegunungan dan pantai menjadi dua pilihan yang paling umum dalam kategori ini, namun preferensi wisatawan terhadap kedua jenis destinasi tersebut tidak bersifat seragam. Perbedaan preferensi dapat dipengaruhi oleh karakteristik demografis, gaya hidup, pengalaman perjalanan, serta orientasi aktivitas rekreasi, sehingga pemahaman terhadap preferensi wisatawan menjadi penting bagi pengelola destinasi dan pemasar pariwisata dalam merancang strategi yang lebih tepat sasaran [3].

Perkembangan data digital dan pendekatan analitik modern membuka peluang yang lebih besar untuk memahami perilaku wisatawan secara sistematis. Dalam sektor pariwisata, data mengenai karakteristik wisatawan, preferensi aktivitas, anggaran perjalanan, dan pilihan destinasi dapat diolah menjadi informasi prediktif yang berguna bagi perumusan keputusan manajerial. Dalam konteks ini, preferensi destinasi dapat dipandang sebagai permasalahan klasifikasi, yaitu pengelompokan wisatawan berdasarkan kecenderungan mereka terhadap tipe destinasi tertentu. *Machine learning* menjadi pendekatan yang relevan karena mampu mengidentifikasi pola tersembunyi, hubungan antarvariabel, serta kecenderungan preferensi yang sulit ditangkap melalui analisis deskriptif konvensional. Informasi yang dihasilkan dari model prediktif dapat mendukung pengambilan keputusan dalam sektor pariwisata, seperti segmentasi pasar, personalisasi promosi, perencanaan layanan, pengelolaan kapasitas destinasi, dan penentuan prioritas pengembangan wisata berbasis data [4]–[6]. Penerapan *machine learning* dalam penelitian pariwisata telah berkembang pada berbagai bidang, seperti analisis sentimen, klasifikasi ulasan, sistem rekomendasi, dan prediksi kunjungan wisatawan [7]–[11].

Meskipun demikian, kajian yang secara khusus membandingkan performa algoritma klasifikasi dalam memprediksi preferensi wisatawan terhadap destinasi wisata berbasis alam, khususnya antara pegunungan dan pantai, masih relatif terbatas [12]. Sebagian besar penelitian terdahulu lebih berfokus pada kepuasan wisatawan, ulasan daring, kualitas layanan, atau pola kunjungan, sehingga ruang kajian mengenai klasifikasi preferensi destinasi alam masih terbuka untuk diteliti lebih lanjut. Kesenjangan ini menunjukkan perlunya penelitian yang tidak hanya menerapkan *machine learning* dalam konteks pariwisata, tetapi juga membandingkan kemampuan

model yang berbeda dalam menangkap pola preferensi wisatawan.

Berdasarkan hal tersebut, penelitian ini bertujuan untuk membandingkan kinerja *Random Forest* dan *Naive Bayes* dalam mengklasifikasikan preferensi wisatawan terhadap destinasi wisata berbasis alam. Kedua algoritma dipilih karena mewakili pendekatan klasifikasi yang berbeda, yaitu model *ensemble* berbasis pohon keputusan dan model probabilistik sederhana. Perbandingan dilakukan untuk mengidentifikasi model yang memiliki kemampuan prediktif lebih baik pada dataset yang digunakan. Hasil penelitian ini diharapkan memberikan kontribusi teoritis dalam pengembangan penerapan *machine learning* pada studi pariwisata, serta kontribusi praktis bagi pengelola destinasi, perencana pariwisata, dan lembaga pemasaran dalam menyusun strategi promosi dan pengembangan destinasi berbasis data.

2. TINJAUAN PUSTAKA

Preferensi wisata merupakan kecenderungan wisatawan dalam memilih jenis destinasi, aktivitas, dan pengalaman perjalanan yang sesuai dengan minat, kebutuhan, serta karakteristik personal. Dalam konteks wisata berbasis alam, preferensi terhadap destinasi pegunungan dan pantai mencerminkan perbedaan orientasi rekreasi, seperti pencarian ketenangan, petualangan, kenyamanan, dan hiburan. Literatur menunjukkan bahwa keputusan wisata tidak bersifat homogen, melainkan dipengaruhi oleh berbagai faktor, seperti latar belakang demografis, gaya hidup, pengalaman perjalanan, dan persepsi terhadap daya tarik destinasi [13], [14]. Oleh karena itu, preferensi wisatawan dapat dipahami sebagai pola perilaku yang layak dianalisis secara sistematis untuk mendukung segmentasi pasar dan perumusan strategi pengembangan destinasi yang lebih tepat sasaran.

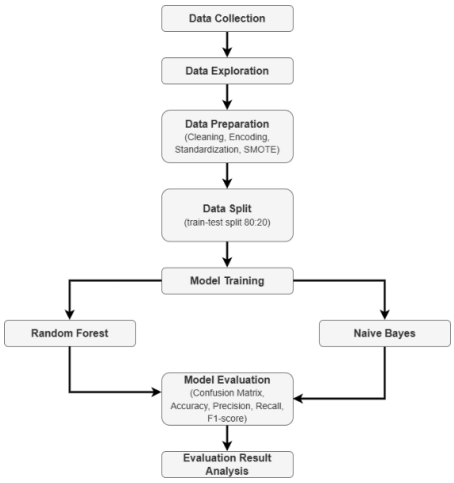
Perkembangan data digital dan teknologi analitik telah mendorong pemanfaatan *machine learning* dalam penelitian pariwisata. Pendekatan ini memungkinkan identifikasi pola dari data dalam jumlah besar serta mendukung proses prediksi dan klasifikasi secara lebih objektif dan efisien. Dalam kajian pariwisata, *machine learning* telah digunakan untuk analisis perilaku wisatawan, prediksi kunjungan, klasifikasi ulasan, dan pengembangan sistem rekomendasi destinasi [15], [16]. Hal ini menunjukkan bahwa perilaku dan preferensi wisatawan tidak lagi hanya dianalisis melalui pendekatan deskriptif konvensional, tetapi juga melalui model prediktif yang mampu mengekstraksi hubungan antar variabel secara lebih mendalam [17]–[20].

Random Forest dan *Naive Bayes* merupakan dua algoritma klasifikasi yang relevan untuk dibandingkan dalam konteks tersebut karena keduanya mewakili pendekatan pemodelan yang berbeda. *Random Forest* merupakan metode *ensemble* yang menggabungkan banyak *decision tree* untuk menghasilkan prediksi yang lebih stabil, akurat, dan tahan terhadap *overfitting*, serta efektif dalam menangani data berdimensi banyak dan hubungan fitur yang kompleks [21], [22]. Sebaliknya, *Naive Bayes* adalah algoritma probabilistik yang dibangun atas asumsi independensi kondisional antar fitur, dengan keunggulan pada kesederhanaan, efisiensi komputasi, dan kemampuan memberikan kinerja yang memadai pada berbagai tugas klasifikasi [23]. Perbedaan karakteristik tersebut menjadikan keduanya layak diuji secara komparatif pada klasifikasi preferensi wisatawan.

Berdasarkan kajian terdahulu, penerapan *machine learning* dalam bidang pariwisata telah berkembang luas, tetapi sebagian besar penelitian masih berfokus pada analisis sentimen, prediksi kunjungan, sistem rekomendasi, dan klasifikasi ulasan wisatawan [15]–[20]. Penelitian yang secara khusus membandingkan performa algoritma klasifikasi untuk memprediksi preferensi wisatawan terhadap destinasi wisata berbasis alam, terutama antara pegunungan dan pantai, masih relatif terbatas. Kesenjangan tersebut menunjukkan perlunya penelitian yang menelaah secara langsung perbandingan kinerja *Random Forest* dan *Naive Bayes* dalam konteks klasifikasi preferensi destinasi. Dengan demikian, penelitian ini memiliki posisi penting baik secara teoritis, karena memperluas penerapan *machine learning* dalam studi pariwisata, maupun secara praktis, karena menyediakan dasar analitis bagi pengelola destinasi dalam merancang strategi promosi dan pengembangan wisata berbasis data.

3. METODE PENELITIAN

Penelitian ini menggunakan desain kuantitatif dengan memanfaatkan kerangka klasifikasi *machine learning* untuk mengidentifikasi preferensi wisatawan terhadap destinasi wisata berbasis alam, khususnya wisata pegunungan dan pantai. Proses metodologis penelitian mencakup beberapa tahapan, yaitu pengumpulan data, analisis eksploratif, persiapan data, pembagian *dataset*, pembangunan model, evaluasi, serta interpretasi hasil. Penelitian ini membandingkan dua algoritma klasifikasi, yaitu *Random Forest* dan *Naive Bayes*. Alur umum proses penelitian ini disajikan pada Gambar 1.



Gambar 1. Diagram alur penelitian

Random Forest digunakan karena struktur *ensemble* yang dimilikinya mampu menangani interaksi fitur yang kompleks, mengurangi risiko *overfitting*, serta menghasilkan performa prediktif yang relatif stabil. Sementara itu, *Naive Bayes* dipilih sebagai model pembanding karena memiliki struktur yang lebih sederhana, kebutuhan komputasi yang relatif rendah, serta relevan untuk klasifikasi berbasis probabilitas. Melalui rancangan tersebut, penelitian ini bertujuan memperoleh evaluasi yang objektif terhadap kemampuan prediktif kedua algoritma dalam mengklasifikasikan preferensi wisatawan terhadap destinasi wisata berbasis alam.

3.1. Data Collection

Penelitian ini menggunakan data sekunder yang diperoleh dari platform *Kaggle*, yaitu dataset berjudul "*Mountains vs. Beaches Preferences*" [24]. *Dataset* tersebut disusun untuk menggambarkan perbedaan preferensi wisata antara dua jenis destinasi utama, yakni pegunungan dan pantai, berdasarkan informasi demografis serta karakteristik responden. Data tersedia dalam format *CSV* dan dirancang untuk mendukung prediksi terhadap kecenderungan pilihan destinasi wisata wisatawan. Pemilihan dataset dari *Kaggle* didasarkan pada sifatnya yang terbuka, mudah diakses, serta luas digunakan dalam penelitian berbasis *machine learning*. Lima contoh data dari dataset *Mountains vs. Beaches Preferences* yang digunakan dalam proses klasifikasi disajikan pada Tabel 1.

Tabel 1. Contoh data untuk kumpulan data preferensi pegunungan vs. pantai

No	*Age*	*Gender*	*Preferred Activities*	*Vacation Budget*	*Preference*
1	56	*male*	*skiing*	2477	1
2	69	*male*	*swimming*	4777	0
3	46	*female*	*skiing*	1469	1
4	32	*non-binary*	*hiking*	1482	1
5	60	*female*	*sunbathing*	516	0

3.2. *Data Exploration*

Tahap eksplorasi data merupakan pemeriksaan awal terhadap *dataset* sebelum dilakukan proses persiapan data dan pembangunan model. Pada tahap ini, analisis difokuskan pada pemahaman terhadap struktur data, yang mencakup jumlah observasi dan atribut, jenis variabel, konsistensi label kelas, serta keberadaan *missing values*, data duplikat, dan pola distribusi kelas. Pemeriksaan ini bertujuan untuk memastikan bahwa *dataset* layak digunakan dalam tugas klasifikasi dan mampu merepresentasikan preferensi wisatawan pada destinasi wisata berbasis alam, yang dalam penelitian ini dibedakan ke dalam dua kelompok, yaitu pegunungan dan pantai [5], [24].

3.3. *Data Preparation*

Sebelum model dibangun, dataset terlebih dahulu melalui tahap persiapan untuk memastikan bahwa data siap digunakan dalam analisis klasifikasi. Hasil eksplorasi menunjukkan bahwa dataset tidak memiliki *missing values* maupun data duplikat, sehingga tidak diperlukan proses imputasi maupun penghapusan data ganda. Entri *nonbiner* pada variabel *Gender* dikeluarkan dari dataset guna menjaga konsistensi tugas klasifikasi. Selanjutnya, dilakukan proses *encoding* untuk mengubah variabel kategorikal menjadi bentuk numerik agar dapat digunakan dalam pemodelan *machine learning* [25]. Setelah itu, fitur distandardisasi menggunakan *StandardScaler* untuk mengurangi perbedaan skala antarkomponen variabel dan mendukung proses pembelajaran model yang lebih konsisten [26]. Sebelum model dilatih, diterapkan pula *Synthetic Minority Oversampling Technique (SMOTE)* untuk mengatasi ketidakseimbangan kelas, karena data latih menunjukkan distribusi yang tidak merata antara kelas pantai dan pegunungan [27]. Rangkaian tahapan ini dilakukan agar dataset menjadi lebih konsisten dan siap digunakan pada proses pelatihan serta pengujian model.

3.4. *Data Split*

Dataset dibagi ke dalam *subset* data latih dan data uji dengan rasio 80:20 untuk membangun sekaligus mengevaluasi model klasifikasi. *Subset* data latih digunakan untuk melatih model, sedangkan subset data uji digunakan untuk menilai kemampuan model dalam melakukan prediksi terhadap data baru. Dalam proses pembagian tersebut, digunakan *random state* tetap sebesar 42 agar partisi data yang dihasilkan dapat direproduksi secara konsisten pada eksperimen yang berbeda. Pemisahan ini memungkinkan proses pelatihan dan pengujian dilakukan secara terpisah, sehingga evaluasi terhadap kinerja klasifikasi dapat dilakukan secara lebih objektif [28], [29]. Rasio 80:20 dipilih karena umum digunakan dalam penelitian klasifikasi dan dinilai mampu memberikan keseimbangan yang memadai antara kebutuhan pembelajaran model dan kebutuhan pengujian pada data yang tidak termasuk dalam proses pelatihan [30].

3.5. *Model Training*

Dua algoritma klasifikasi, yaitu *Random Forest* dan *Naive Bayes*, dilatih untuk memprediksi jenis destinasi wisata berbasis alam yang cenderung dipilih oleh wisatawan. Untuk mengurangi pengaruh ketidakseimbangan kelas terhadap proses pembelajaran model, data latih terlebih dahulu diproses menggunakan *StandardScaler* untuk menstandardisasi fitur dan *SMOTE* untuk menyeimbangkan distribusi kelas. Selanjutnya, model klasifikasi dibangun berdasarkan data hasil prapemrosesan tersebut untuk merepresentasikan hubungan antara karakteristik responden dan preferensi mereka terhadap destinasi pegunungan maupun pantai. *Random Forest* digunakan karena mekanisme *ensemble* yang dimilikinya memungkinkan penggabungan sejumlah *decision tree*, sehingga model menjadi lebih stabil dan memiliki kecenderungan *overfitting* yang lebih rendah [21].

Adapun *Naive Bayes* dipilih sebagai model pembanding karena memiliki struktur probabilistik yang sederhana, komputasi yang cepat, serta efektif digunakan untuk klasifikasi berdasarkan pola probabilitas fitur [23]. Kedua algoritma dijalankan dengan pengaturan parameter dasar yang hanya disesuaikan secara terbatas agar tetap sejalan dengan desain eksperimen. Model yang telah dilatih tersebut kemudian digunakan sebagai dasar untuk membandingkan kinerja masing-masing algoritma pada tahap evaluasi.

3.6. *Model Evaluation*

Data uji digunakan untuk membandingkan seberapa baik *Random Forest* dan *Naive Bayes* mampu memprediksi wisatawan yang cenderung memilih pegunungan atau pantai sebagai destinasi wisata. Label hasil prediksi kemudian dibandingkan dengan label aktual, dan hasilnya disajikan dalam bentuk *confusion matrix* untuk memperlihatkan kinerja klasifikasi secara lebih jelas. Dari perbandingan tersebut, diperoleh empat ukuran kinerja, yaitu *accuracy*, *precision*, *recall*, dan *F1-score*. Keempat ukuran ini digunakan untuk memberikan gambaran yang lebih komprehensif mengenai hasil klasifikasi biner, karena accuracy secara keseluruhan saja belum memadai untuk menunjukkan kualitas model secara utuh [31], [32]. Dalam penelitian ini, *accuracy* digunakan untuk menunjukkan persentase keseluruhan klasifikasi yang benar, *precision* untuk menunjukkan ketepatan prediksi positif, *recall* untuk menunjukkan kemampuan model dalam mengidentifikasi kasus positif, dan *F1-score* untuk menggabungkan *precision* dan *recall* ke dalam satu nilai tunggal. Untuk mempertimbangkan proporsi kelas dalam data uji, nilai *precision*, *recall*, dan *F1-score* disajikan dalam bentuk *weighted average*. Dengan menggunakan beberapa ukuran evaluasi sebelum penyajian rumus, kinerja model dapat dibandingkan dari berbagai sudut pandang secara lebih memadai [31], [33].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1-Score = 2x\frac{Precision\ x\ Recall}{Precision+Recall} \quad (4)$$

3.7. *Evaluation Result Analysis*

Tahap analisis hasil dilakukan untuk mengevaluasi secara menyeluruh kinerja *Random Forest* dan *Naive Bayes* dengan memanfaatkan *confusion matrix* serta metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-score*. Pada tahap ini, nilai dari setiap metrik dibandingkan antar model untuk menilai tingkat konsistensi kinerjanya. Analisis ini juga mencakup penelaahan terhadap kecenderungan masing-masing model dalam menghasilkan kesalahan klasifikasi pada kelas tertentu. Evaluasi tidak hanya bergantung pada *accuracy*, tetapi juga mempertimbangkan *precision* dan *recall* guna menghindari bias, terutama ketika distribusi kelas tidak sepenuhnya seimbang. Adapun *F1-score* digunakan sebagai ukuran kompromi antara *precision* dan *recall*. Seluruh proses evaluasi dilakukan dengan menggunakan data uji yang sama agar kesimpulan komparatif yang dihasilkan menjadi lebih kuat. Dengan demikian, perbedaan performa yang diamati dapat diatribusikan pada perbedaan perilaku model, bukan pada perbedaan sampel pengujian.

4. HASIL DAN PEMBAHASAN

Bagian ini menyajikan temuan penelitian sekaligus membahas kinerja model klasifikasi yang digunakan pada data preferensi wisatawan. Pembahasan dibagi ke dalam tiga bagian, yaitu hasil tahap eksplorasi data, kinerja model *Random Forest* dan *Naive Bayes*, serta analisis perbandingan kinerja kedua algoritma dalam mengklasifikasikan preferensi terhadap destinasi wisata berbasis alam.

4.1. Hasil Eksplorasi Data

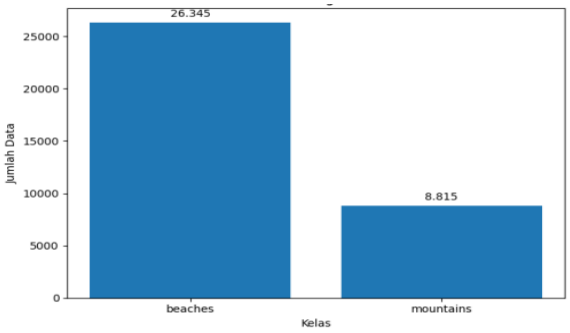
Hasil pemeriksaan awal terhadap dataset *Mountains vs. Beaches Preferences* menunjukkan bahwa dataset tersebut terdiri atas 52.444 rekaman dan 14 atribut. *Dataset* ini memuat 9 variabel numerik dan 5 variabel kategorikal. Variabel *Preference* berperan sebagai variabel target dalam penelitian ini, sedangkan atribut lainnya digunakan sebagai variabel prediktor. Hasil inspeksi *dataset* menunjukkan bahwa seluruh atribut memiliki informasi yang lengkap dan tidak mengandung *missing values*. Pemeriksaan terhadap data duplikat juga menunjukkan bahwa tidak terdapat rekaman ganda. Setelah melalui proses pembersihan data, *dataset* yang digunakan dalam analisis berjumlah 35.160 rekaman. Ringkasan *dataset* setelah proses eksplorasi dan pembersihan disajikan pada Tabel 2.

Tabel 2. Ringkasan *dataset* setelah eksplorasi dan pembersihan

Description	***Value***
Initial data	52,444
Final data	35,160
Number of attributes	14
Missing values	0
Duplicate data	0
Beaches class	26,345
Mountains class	8,815

Untuk menjaga konsistensi proses *encoding* pada variabel kategorikal, rekaman dengan kategori *nonbiner* pada variabel *Gender* dihapus pada tahap pembersihan data. Langkah ini menyebabkan jumlah rekaman berkurang menjadi 35.160. Setelah pembersihan, distribusi kelas menunjukkan bahwa kelas *Beaches* berjumlah 26.345 rekaman, sedangkan kelas *Mountains* berjumlah 8.815 rekaman. Kondisi ini menunjukkan adanya ketidakseimbangan kelas yang berpotensi memengaruhi kinerja klasifikasi. Untuk mengatasi ketimpangan distribusi tersebut, diterapkan *SMOTE* pada data latih. Analisis korelasi

antarvariabel juga menunjukkan bahwa sebagian besar atribut tidak memiliki hubungan linear yang kuat dengan variabel target. Kondisi ini mengindikasikan bahwa *dataset* kemungkinan memiliki hubungan antar fitur yang lebih kompleks. Oleh karena itu, perbandingan algoritma klasifikasi dengan karakteristik pemodelan yang berbeda, seperti *Random Forest* dan *Naive Bayes*, menjadi relevan untuk dilakukan. Visualisasi distribusi kelas target disajikan pada Gambar 2.



Gambar 2. Distribusi kelas target setelah tahap pembersihan data

4.2. Hasil Evaluasi Model

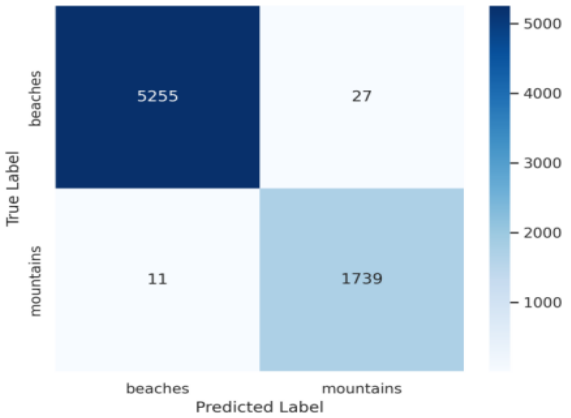
Subbagian ini membahas hasil perbandingan antara *Random Forest* dan *Naive Bayes* dalam mengklasifikasikan preferensi wisatawan terhadap destinasi wisata berbasis alam. Evaluasi kinerja model dilakukan menggunakan empat ukuran utama, yaitu *accuracy*, *precision*, *recall*, dan *F1-score*. Keempat ukuran tersebut digunakan untuk memberikan gambaran yang lebih komprehensif mengenai kemampuan prediktif dan klasifikasi masing-masing model. Berdasarkan Tabel 3, *Random Forest* secara konsisten menunjukkan nilai kinerja yang lebih tinggi dibandingkan *Naive Bayes* pada seluruh metrik evaluasi.

Tabel 3. Hasil evaluasi *Random Forest* dan *Naive Bayes*

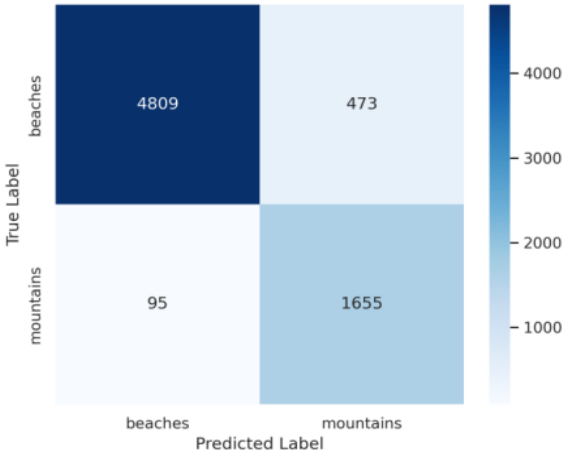
Model	***Accuracy***	***Precision***	***Recall***	***F1-Score***
Random Forest	0.994596	0.994626	0.994596	0.994604
Naive Bayes	0.919226	0.930133	0.919226	0.921664

Model *Random Forest* menghasilkan nilai *accuracy* sebesar 0,994596, *precision* sebesar 0,994626, *recall* sebesar 0,994596, dan *F1-score* sebesar 0,994604. Sebaliknya, model *Naive Bayes* memperoleh nilai *accuracy* sebesar 0,919226, *precision* sebesar 0,930133, *recall* sebesar 0,919226, dan *F1-score* sebesar 0,921664. Hasil tersebut menunjukkan bahwa *Random Forest* memiliki kemampuan yang jauh lebih baik dalam mengklasifikasikan *dataset* yang digunakan dalam penelitian ini dibandingkan dengan *Naive Bayes*.

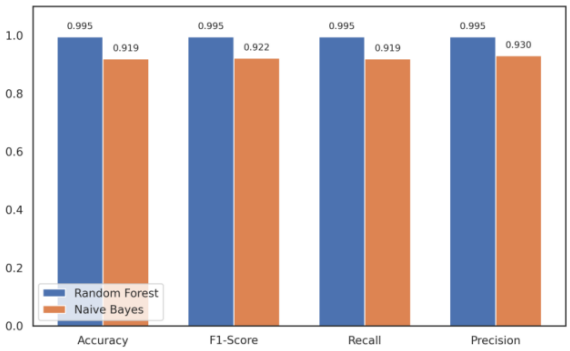
Temuan pada *confusion matrix* memperjelas perbedaan pola klasifikasi antara kedua model. Berdasarkan Gambar 3, model *Random Forest* mengklasifikasikan 5.255 data kelas *beaches* dengan benar sebagai *beaches*, sedangkan 27 data *beaches* salah diklasifikasikan sebagai *mountains*. Pada kelas *mountains*, sebanyak 1.739 data diklasifikasikan dengan benar sebagai *mountains*, sedangkan 11 data *mountains* salah diklasifikasikan sebagai *beaches*. Dengan demikian, *Random Forest* hanya menghasilkan 38 kesalahan klasifikasi dari keseluruhan data uji. Sebaliknya, berdasarkan Gambar 4, *Naive Bayes* menghasilkan jumlah kesalahan yang lebih besar. Model ini mengklasifikasikan 4.809 data *beaches* dengan benar, tetapi 473 data *beaches* salah diklasifikasikan sebagai *mountains*. Pada kelas *mountains*, sebanyak 1.655 data diklasifikasikan dengan benar, sedangkan 95 data *mountains* salah diklasifikasikan sebagai *beaches*. Secara keseluruhan, *Naive Bayes* menghasilkan 568 kesalahan klasifikasi. Perbedaan jumlah kesalahan tersebut menunjukkan bahwa *Random Forest* tidak hanya unggul pada metrik agregat seperti akurasi, presisi, *recall*, dan *F1-score*, tetapi juga lebih konsisten dalam membedakan preferensi wisatawan terhadap destinasi *beaches* dan *mountains*. Perbandingan visual antar metrik evaluasi pada kedua model disajikan pada Gambar 5, yang secara umum menegaskan keunggulan performa *Random Forest*.



Gambar 3. *Confusion matrix* untuk *Random Forest*



Gambar 4. *Confusion matrix* untuk *Naive Bayes*



Gambar 5. Perbandingan metrik evaluasi model

4.3. Analisis dan Perbandingan Hasil

Hasil evaluasi menunjukkan bahwa *Random Forest* memiliki performa lebih unggul dibandingkan *Naive Bayes* dalam mengklasifikasikan preferensi wisatawan terhadap destinasi wisata berbasis alam. Keunggulan ini tercermin secara konsisten pada Tabel 3, *confusion matrix* pada Gambar 3 dan Gambar 4, serta perbandingan metrik evaluasi pada Gambar 5. *Random Forest* memperoleh akurasi, presisi, *recall*, dan *F1-score* yang lebih tinggi dibandingkan *Naive Bayes*, sehingga menunjukkan kemampuan klasifikasi yang lebih stabil pada data uji. Faktor utama yang menjelaskan keunggulan tersebut adalah kemampuan *Random Forest* dalam menangkap pola *nonlinier* dan interaksi antar fitur melalui mekanisme *ensemble* berbasis banyak *decision tree*. Karakteristik ini relevan dengan *dataset* preferensi wisatawan yang memuat kombinasi atribut demografis, aktivitas, anggaran perjalanan, dan kecenderungan pilihan destinasi. Hubungan antar atribut dalam konteks pariwisata tidak selalu bersifat linear atau independen, karena preferensi wisatawan dapat terbentuk dari kombinasi beberapa faktor secara simultan. Hasil eksplorasi data juga menunjukkan bahwa sebagian besar atribut tidak memiliki hubungan linear yang kuat dengan variabel target, sehingga model yang mampu membentuk pemisahan kelas secara lebih fleksibel menjadi lebih sesuai.

Selain itu, adanya ketidakseimbangan kelas antara *beaches* dan *mountains* menunjukkan perlunya model yang mampu mengenali pola pada kedua kelas secara lebih seimbang. Penerapan *SMOTE* pada data latih membantu memperbaiki representasi kelas minoritas, sedangkan mekanisme *Random Forest* memungkinkan proses klasifikasi menjadi lebih stabil karena setiap pohon dibangun dari kombinasi observasi dan fitur yang berbeda. Dengan demikian, risiko kesalahan yang berasal dari satu pola pembelajaran tunggal dapat dikurangi. Sebaliknya, *Naive Bayes* bergantung pada asumsi independensi kondisional antar fitur. Asumsi ini cenderung sulit dipenuhi pada data preferensi wisata, karena variabel seperti aktivitas yang disukai, anggaran liburan, frekuensi perjalanan, dan karakteristik responden berpotensi saling berkaitan. Ketidaksesuaian antara asumsi model dan karakteristik data tersebut dapat menjelaskan mengapa *Naive Bayes* menghasilkan kesalahan klasifikasi yang lebih besar dibandingkan *Random Forest*.

Temuan penelitian ini secara umum sejalan dengan penelitian terdahulu dalam analitik pariwisata yang menegaskan efektivitas *machine learning* untuk fungsi prediksi, segmentasi, dan rekomendasi [5]. Studi sebelumnya juga menunjukkan bahwa *Random Forest* cenderung unggul pada data dengan struktur yang relatif kompleks, sedangkan *Naive Bayes* lebih sesuai digunakan pada konteks klasifikasi probabilistik yang lebih sederhana [11]. Dari sudut pandang praktis, hasil ini menunjukkan bahwa *Random Forest* merupakan model yang lebih tepat untuk memprediksi preferensi wisatawan terhadap destinasi dalam penelitian ini. Model tersebut berpotensi dimanfaatkan oleh pemangku kepentingan pariwisata untuk mengidentifikasi segmen pasar dan menyusun strategi pemasaran yang lebih terarah. Meskipun demikian, beberapa keterbatasan tetap perlu dicatat. Analisis ini hanya menggunakan satu *dataset* publik, membandingkan dua algoritma klasifikasi, dan menerapkan pengaturan parameter dasar tanpa optimasi *hyperparameter* yang mendalam. Oleh karena itu, penelitian selanjutnya dapat memperluas temuan ini melalui penggunaan beberapa *dataset*, penerapan *cross-validation*, serta integrasi teknik optimasi yang lebih lanjut agar hasil yang diperoleh menjadi lebih kuat dan lebih luas daya generalisasinya.

5. KESIMPULAN DAN SARAN

Analisis komparatif menunjukkan bahwa *Random Forest* dan *Naive Bayes* sama-sama mampu mengklasifikasikan preferensi wisatawan terhadap destinasi wisata berbasis alam. Namun, *Random Forest* menunjukkan kinerja yang lebih unggul pada seluruh metrik evaluasi. *Random Forest* memperoleh akurasi sebesar 0,994596, presisi sebesar 0,994626, *recall* sebesar 0,994596, dan *F1-score* sebesar 0,994604. Sementara itu, *Naive Bayes* memperoleh akurasi sebesar 0,919226, presisi sebesar 0,930133, *recall* sebesar 0,919226, dan *F1-score* sebesar 0,921664. Perbedaan nilai tersebut menunjukkan bahwa *Random Forest* memiliki kemampuan klasifikasi yang lebih konsisten dan lebih efektif dalam membedakan preferensi wisatawan terhadap destinasi pantai dan pegunungan pada *dataset* yang dianalisis. Kontribusi utama penelitian ini terletak pada penyediaan bukti empiris mengenai efektivitas komparatif dua algoritma klasifikasi dalam memprediksi preferensi wisata berbasis alam. Selain itu, penelitian ini memperkuat pemanfaatan *machine learning* dalam riset pariwisata dengan menunjukkan bahwa pemodelan prediktif dapat digunakan untuk mendukung klasifikasi berbasis preferensi dan analitik pariwisata. Dari sisi praktis, temuan ini berpotensi membantu pengelola destinasi, perencana pariwisata, dan lembaga pemasaran dalam menyusun segmentasi pasar, meningkatkan ketepatan strategi promosi, serta mendukung pengambilan keputusan berbasis data dalam pengelolaan pariwisata alam. Meskipun

demikian, penelitian ini memiliki sejumlah keterbatasan. Analisis hanya bertumpu pada satu *dataset* publik, sehingga daya generalisasi hasil penelitian masih terbatas. Di samping itu, perbandingan hanya dilakukan terhadap dua algoritma, dan model diuji menggunakan pengaturan parameter yang sederhana tanpa prosedur optimasi yang lebih mendalam. Oleh karena itu, penelitian selanjutnya perlu mempertimbangkan penggunaan *dataset* yang lebih besar dan lebih heterogen, memasukkan metode klasifikasi tambahan, serta menerapkan teknik optimasi model yang lebih ketat, termasuk *cross-validation* dan *hyperparameter tuning*, agar diperoleh temuan yang lebih kuat dan lebih dapat diterapkan pada beragam konteks pariwisata.

DAFTAR PUSTAKA

[1] R. Robina-Ramírez, J. Torrecilla-Pinero, A. Leal-Solís, and J. A. Pavón-Pérez, "Tourism as a driver of economic and social development in underdeveloped regions," *Regional Science Policy & Practice*, vol. 16, no. 1, p. 12639, Jan. 2024, doi: 10.1111/rsp3.12639.

[2] P. L. Winter, S. Selin, L. Cerveny, and K. Bricker, "Outdoor Recreation, Nature-Based Tourism, and Sustainability," *Sustainability*, vol. 12, no. 1, p. 81, Dec. 2019, doi: 10.3390/su12010081.

[3] C. Jönsson and D. Devonish, "Does Nationality, Gender, and Age Affect Travel Motivation? a Case of Visitors to The Caribbean Island of Barbados," *Journal of Travel & Tourism Marketing*, vol. 25, no. 3–4, pp. 398–408, Dec. 2008, doi: 10.1080/10548400802508499.

[4] S. J. Miah, H. Q. Vu, J. Gammack, and M. McGrath, "A Big Data Analytics Method for Tourist Behaviour Analysis," *Information & Management*, vol. 54, no. 6, pp. 771–785, Sep. 2017, doi: 10.1016/j.im.2016.11.011.

[5] J. C. S. Núñez, J. A. Gómez-Pulido, and R. R. Ramírez, "Machine learning applied to tourism: A systematic review," *WIREs Data Mining and Knowledge Discovery*, vol. 14, no. 5, Sep. 2024, doi: 10.1002/widm.1549.

[6] S. Kongpeng and A. Hanskunatai, "Tourist Destination Recommendation System based on Machine Learning," in *2024 9th International Conference on Big Data and Computing*, New York, NY, USA: ACM, May 2024, pp. 58–67. doi: 10.1145/3695220.3695229.

[7] V. Taecharungroj and B. Mathayomchan, "Analysing TripAdvisor reviews of tourist attractions in Phuket, Thailand," *Tour Manag*, vol. 75, pp. 550–568, Dec. 2019, doi: 10.1016/j.tourman.2019.06.020.

[8] N. Leelawat *et al.*, "Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning," *Heliyon*, vol. 8, no. 10, p. e10894, Oct. 2022, doi: 10.1016/j.heliyon.2022.e10894.

[9] X. Zhou, J. Peng, B. Wen, and M. Su, "Tour Route Recommendation Model by the Improved Symmetry-Based Naive Bayes Mining and Spatial Decision Forest Search," *Symmetry (Basel)*, vol. 15, no. 12, p. 2168, Dec. 2023, doi: 10.3390/sym15122168.

[10] J. Chen, J. Cong, M. Li, Y. Sun, and J. Zhang, "T-ECBM: a deep learning-based text-image multimodal model for tourist attraction recommendation," *Sci Rep*, vol. 15, no. 1, p. 41638, Nov. 2025, doi: 10.1038/s41598-025-25630-z.

[11] C. Li and W. Zheng, "Nipping trouble in the bud: A proactive tourism recommender system," *Information & Management*, vol. 62, no. 1, p. 104062, Jan. 2025, doi: 10.1016/j.im.2024.104062.

[12] A. Tkaczynski, S. R. Rundle-Thiele, and N. Beaumont, "Segmentation: A tourism stakeholder view," *Tour Manag*, vol. 30, no. 2, pp. 169–175, Apr. 2009, doi: 10.1016/j.tourman.2008.05.010.

[13] V. Chang, M. R. Islam, A. Ahad, M. J. Ahmed, and Q. A. Xu, "Machine learning for predicting tourist spots' preference and analysing future tourism trends in Bangladesh," *Enterp Inf Syst*, vol. 18, no. 12, Dec. 2024, doi: 10.1080/17517575.2024.2415568.

[14] A. Solano-Barliza, A. Valls, A. Moreno, M. Acosta-Coll, and J. Escorcia-Gutierrez, "A logic multi-criteria recommender system for tourist services: Jimataa," *International Journal of Information Technology*, vol. 18, no. 1, pp. 29–53, Jan. 2026, doi: 10.1007/s41870-025-02742-3.

[15] J. C. S. Núñez, J. A. Gómez-Pulido, and R. R. Ramírez, "Machine learning applied to tourism: A systematic review," *WIREs Data Mining and Knowledge Discovery*, vol. 14, no. 5, Sep. 2024, doi: 10.1002/widm.1549.

[16] M. Á. Álvarez-Carmona *et al.*, "Natural language processing applied to tourism research: A systematic review and future research directions," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, pp. 10125–10144, Nov. 2022, doi: 10.1016/j.jksuci.2022.10.010.

[17] A. Derdouri and T. Osaragi, "A machine learning-based approach for classifying tourists and locals using geotagged photos: the case of Tokyo," *Information Technology & Tourism*, vol. 23, no. 4, pp. 575–609, Dec. 2021, doi: 10.1007/s40558-021-00208-3.

[18] M. Mor, S. Dalyot, and Y. Ram, "Who is a tourist? Classifying international urban tourists using machine learning," *Tour Manag*, vol. 95, p. 104689, Apr. 2023, doi: 10.1016/j.tourman.2022.104689.

[19] J. Bravo, R. Alarcón, C. Valdivia, and O. Serquén, "Application of Machine Learning

Techniques to Predict Visitors to the Tourist Attractions of the Moche Route in Peru," *Sustainability*, vol. 15, no. 11, p. 8967, Jun. 2023, doi: 10.3390/su15118967.

[20] M. Camacho-Ruiz, R. A. Carrasco, G. Fernández-Avilés, and A. LaTorre, "Tourism destination events classifier based on artificial intelligence techniques," *Appl Soft Comput*, vol. 148, p. 110914, Nov. 2023, doi: 10.1016/j.asoc.2023.110914.

[21] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[22] A. Utami, D. F. H. Permadi, Y. D. Rosita, and J. Unjung, "Performance Comparison of Random Forest (RF) and Classification and Regression Trees (CART) for Hotel Star Rating Prediction," *Scientific Journal of Informatics*, vol. 11, no. 3, pp. 733–748, Oct. 2024, doi: 10.15294/sji.v11i3.11068.

[23] P. Domingos and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss," *Mach Learn*, vol. 29, no. 2–3, pp. 103–130, Nov. 1997, doi: 10.1023/A:1007413511361.

[24] J. Paliwal, "Mountains vs. Beaches Preferences," *Kaggle*, 2023. [Online]. Available: https://www.kaggle.com/datasets/jahnavipaliwal/mountains-vs-beaches-preference.

[25] F. Bolikulov, R. Nasimov, A. Rashidov, F. Akhmedov, and Y.-I. Cho, "Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms," *Mathematics*, vol. 12, no. 16, p. 2553, Aug. 2024, doi: 10.3390/math12162553.

[26] M. Ahsan, M. Mahmud, P. Saha, K. Gupta, and Z. Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *Technologies (Basel)*, vol. 9, no. 3, p. 52, Jul. 2021, doi: 10.3390/technologies9030052.

[27] N. v. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[28] Y. Xu and R. Goodacre, "On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning," *J Anal Test*, vol. 2, no. 3, pp. 249–262, Jul. 2018, doi: 10.1007/s41664-018-0068-2.

[29] T. J. Bradshaw, Z. Huemann, J. Hu, and A. Rahmim, "A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging," *Radiol Artif Intell*, vol. 5, no. 4, Jul. 2023, doi: 10.1148/ryai.220232.

[30] M. Sivakumar, S. Parthasarathy, and T. Padmapriya, "Trade-off between training and testing ratio in machine learning for medical image processing," *PeerJ Comput Sci*, vol. 10, p. e2245, Sep. 2024, doi: 10.7717/peerj-cs.2245.

[31] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf Process Manag*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[32] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv:2010.16061*, 2011. doi: 10.48550/arXiv.2010.16061.

[33] M. Conciatori, A. Valletta, and A. Segalini, "Improving the quality evaluation process of machine learning algorithms applied to landslide time series analysis," *Comput Geosci*, vol. 184, p. 105531, Feb. 2024, doi: 10.1016/j.cageo.2024.105531.