Jurnal Teknik Informatika dan Teknologi Informasi
Volume. 5 Nomor. 1 April 2025
E-ISSN: 2827-9387, P- ISSN: 2827-9379, Hal 441-455
DOI: https://doi.org/10.55606/jutiti.v5i1.5364
Available Online at : https://journalshub.org/index.php/jutiti

Prediksi Preferensi Peserta Event Marathon terhadap Kategori Lomba menggunakan Algoritma Machine Learning

Dominic Dinand Aristo[1*], Satria Dwi Nurwicaksana[2], Dendi Putra Prakoso[3], M Afrian Maulana[4], Nurfaizah[5]
[1-5]Program Studi Sistem Informasi,Fakultas Ilmu Komputer, Universitas Amikom Purwokerto, Indonesia

Jl. Letjend Pol.Soemarto No.127, Watumas, Purwanegara, Kecamatan Purwokerto Utara, Kabupaten Banyumas, Jawa Tengah, 53127
Korespondensi penulis: domicdinand@gmail.com[1]*

Abstract. *Marathons are becoming an increasingly popular form of exercise and social interaction. Participants who choose race categories based on the mileage provided, such as 6K, 7.9K, and 11K, according to personal preference. However, this category selection has not been analyzed based on participant characteristics, even though this information is important for organizers to support promotional strategies, and segmentation of participants. This study aims to predict marathon category selection based on demographic characteristics, namely age and gender, by applying Decision Tree and Random Forest machine learning algorithms. The dataset used is primary data from two events, namely RSDK Berlari with a total of 1091 data and Skybridgefunrun with 1519 data. The results show that the Decision Tree algorithm gets an accuracy of 56.81%, and the Random Forest algorithm is 57.38%. With these results, it shows that the Random Forest algorithm has higher accuracy than the Decision Tree algorithm, with accuracy reaching 57.38%. However, the model tends to be biased towards the 7.9K category, with recall reaching 94%, while the 6K and 11K categories are very low. Then, feature importance analysis shows that the most influential factor on category selection is age, while gender is smaller. This research provides insight for event organizers in designing promotional strategies and participant segmentation more precisely.*

Keywords: *Decision Tree, Machine Learning, Prediction, Random Forest , Sport*

Abstrak. Maraton menjadi ajang olahraga sekaligus interaksi sosial yang semakin diminati. Para peserta yang memilih kategori lomba berdasarkan jarak tempuh yang sudah disediakan, seperti 6K, 7.9K, dan 11K, sesuai preferensi pribadi. Namun, pemilihan kategori ini belum dianalisis berdasarkan karakteristik peserta, padahal informasi tersebut penting bagi penyelenggara untuk mendukung strategi promosi, dan segmentasi peserta. Penelitian ini bertujuan untuk memprediksi pemilihan kategori marathon berdasarkan karakteristik demografis, yaitu usia dan jenis kelamin, dengan menerapkan algoritma machine learning Decision Tree dan Random Forest. Dataset yang digunakan merupakan data primer dari dua event, yaitu RSDK Berlari dengan jumlah data sebanyak 1091 data dan Skybridgefunrun dengan 1519 data. Hasil penelitian menunjukan bahwa algoritma Decision Tree mendapatkan akurasi sebesar 56,81%, dan algoritma Random Forest sebesar 57,38%. Dengan hasil tersebut menunjukan bahwa algoritma Random Forest memiliki akurasi lebih tinggi dibandingkan algoritma Decision Tree yaitu dengan akurasi mencapai 57,38%. Namun, model cenderung bias terhadap kategori 7.9K, dengan recall mencapai 94%, sementara kategori 6K dan 11K sangat rendah. Kemudian, analisis feature importance menunjukkan faktor paling berpengaruh terhadap pemilihan kategori adalah usia, sedangkan gender lebih kecil. Penelitian ini memberikan wawasan bagi penyelenggara event dalam merancang strategi promosi dan segmentasi peserta secara lebih tepat.

Kata kunci: Decision Tree , Machine Learning, Olahraga, Prediksi, Random Forest**.**

1. LATAR BELAKANG

Perkembangan teknologi informasi di era digital telah menjangkau berbagai aspek kehidupan, termasuk dalam penyelenggaraan kegiatan olahraga . Salah satu bentuk kegiatan olahraga yang banyak diminati oleh Masyarakat saat ini adalah event lari marathon. Seperti pada penilitian yang dilakukan oleh (Simon P Simanjuntak, 2025)

Received: April 12, 2025; Revised: Mei 23, 2025; Accepted: Juni 21, 2025; Published: Juni 26, 2025

menyatakan bahwa olahraga marathon pada event Toba Marathon mengalami peningkatan pada jumlah peserta tiap tahunnya. Kegiatan ini tidak hanya menjadi ajang kompetisi, tetapi juga sarana untuk mempromosikan gaya hidup sehat dan mempererat interaksi sosial antarpeserta. Umumnya, penyelenggara menyediakan beberapa kategori lomba berdasarkan jarak tempuh, seperti 6 kilometer, 7 kilometer dan 11 kilometer , yang dapat dipilih secara bebas oleh peserta , sesuai dengan preferensi dan kemampuan masing-masing.

Proses pemilihan kategori lomba oleh peserta umumnya dipengaruhi berbagai faktor, seperti usia, jenis kelamin, ukuran tubuh, dan tingkat kebugaran. Namun demikian, pemetaan preferensi ini belum banyak dianalisis secara sistematis berdasarkan karakteristik demografis peserta. Padahal, informasi semacam ini sangat penting, terutama bagi pihak penyelenggara, untuk mendukung pengambilan keputusan dalam hal perencanaan logistik, pengadaan perlengkapan, serta strategi promosi yang lebih tersegmentasi.

Machine learning merupakan pendekatan komputasional yang memungkinkan sistem untuk mempelajari pola dari data dan melakukan prediksi atau klasifikasi secara otomatis. Bidang keilmuan ini secara khusus mengkaji pengembangan berbagai model, algoritma, dan teknik pembelajaran yang memungkinkan mesin memperoleh pengetahuan dengan cara yang menyerupai proses belajar pada manusia. Dengan memanfaatkan teknik statistik dan matematis, machine learning mampu mengolah data untuk menghasilkan prediksi atau kesimpulan yang berguna, serta mengevaluasi seberapa efektif metode - metode pembelajaran yang digunakan dalam membuat prediksi dari data (Venkatesa Palanichamy Narasimma Bharathi, 2025).

Dalam konteks ini, pendekatan berbasis machine learning dapat dimanfaatkan untuk mengidentifikasi pola dari data peserta dan memprediksi pilihan kategori lomba yang kemungkinan besar akan dipilih oleh peserta baru. Metode yang dinilai efektif dalam melakukan klasifikasi data adalah algoritma Decision Tree dan Random Forest, yang mampu menangani variabel numerik maupun kategorikal dengan baik (Jaiswal, 2021), serta memiliki tingkat akurasi yang tinggi dalam berbagai studi sebelumnya.

Penelitian ini dilakukan dengan menggunakan data peserta dari event lari marathon RSDK25 dan Skybridgefunrun yang telah dikumpulkan secara digital. Tujuan utama dari penelitian ini adalah membangun model klasifikasi untuk memprediksi pilihan kategori lomba peserta berdasarkan atribut demografis, serta mengevaluasi performa model yang dibangun. Diharapkan hasil penelitian ini dapat memberikan kontribusi

dalam pengembangan sistem pendukung keputusan berbasis data untuk mendukung penyelenggaraan event olahraga yang lebih efektif, efisien dan terarah di masa mendatang.

2. KAJIAN TEORITIS

Dalam sebuah event marathon faktor demografis dapat memengaruhi keputusan peserta dalam memilih kategori lomba. Penelitian oleh (Anja Wittho, 2024) dalam *PLOS ONE* dengan judul Running trends in Switzerland from 1999 to 2019: Anexploratory observational study yang menganalisis peserta lomba lari di Switzerland menemukan bahwa usia dan jenis kelamin secara signifikan memengaruhi pemilihan jarak lomba. Peserta pria dan yang berusia lebih tua cenderung memilih kategori half-marathon atau marathon, sementara perempuan dan peserta muda lebih banyak memilih kategori 10K. Model regresi logistik dan pohon keputusan yang digunakan dalam studi tersebut mampu memprediksi pilihan kategori lomba berdasarkan data demografis peserta dengan akurasi yang cukup tinggi. Temuan serupa dikonfirmasi oleh studi dari (Leo Lundy, 2024) dalam artikel Demographics, culture and participatory nature of multi-marathoning—An observational study highlighting issues with recommendations yang menganalisis peserta Multi – Marathon dari seluruh dunia. Studi tersebut menemukan bahwa marathon dipengaruhi secara signifikan oleh faktor demografis seperti gender dan usia.

Berdasarkan dari temuan tersebut, penggunaan algoritma prediktif seperti Decision Tree dan Random Forest menjadi relevan untuk diterapkan dalam memodelkan hubungan antara karakteristik demografis peserta dan pilihan kategori lomba. Kedua algoritma ini telah banyak digunakan dalam berbagai bidang seperti kesehatan, sosial, bencana termasuk olahraga. Salah satunya pada penilitian yang dilakukan oleh (Pakawan Pugsee, 2020) , model Random Forest digunakan untuk memprediksi hasil dari pertandingan Liga Premier Inggris. Hasil penelitian menunjukkan bahwa model Random Forest mampu memberikan prediksi dengan akurasi yang tinggi yakni mencapai 80%. Pengujian tersebut didukung dengan memanfaatkan data historis performa tim dan pemain. Hasil tersebut menjadikan algoritma Random Forest dapat dikatakan sebagai model yang mampu digunakan dalam melakukan sebuah prediksi untuk bidang olahraga.

Model random forest juga digunakan dalam penelitian (Mengjie Jia, 2020) yang menghasilkan akurasi yang tinggi untuk memprediksi jumlah medali Olimpiade Musim Panas berdasarkan faktor makroekonomi dan demografi, seperti GDP, populasi, ukuran tim nasional, dan keuntungan tuan rumah sebesar 89.76 %. Di luar bidang olahraga,

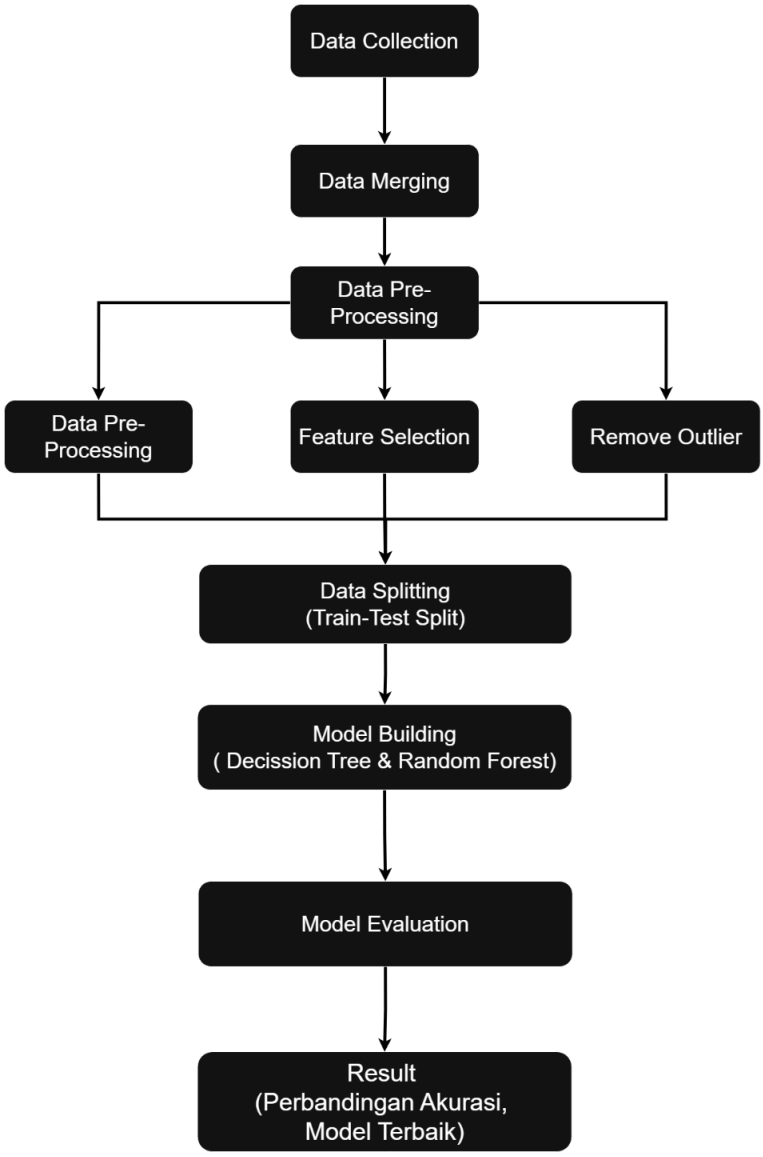
efektivitas algoritma ini juga dibuktikan dalam konteks prediksi bencana. Penelitian yang dilakukan oleh (Muhammad Bagas Arya Darmawan, 2023) dalam mengembangkan model prediksi banjir dengan membandingkan algoritma Decision Tree, Random Forest, dan Naïve Bayes. Hasilnya menunjukkan bahwa Random Forest mampu mencapai akurasi prediksi sebesar 99,05%, menjadikannya metode yang unggul untuk klasifikasi kejadian bencana berdasarkan parameter seperti curah hujan dan tinggi muka air sungai.

Selain itu, dalam konteks medis, algoritma ini juga terbukti efektif. Penelitian oleh (Massimo Giotta, 2022) menunjukkan bahwa Decision Tree dapat memprediksi prognosis pasien COVID-19 dengan skor F1 mencapai 75,93%, sehingga dapat membantu dokter memperoleh informasi yang berguna sebagai peringatan perkembangan penyakit pada pasien COVID-19.

Berbagai studi tersebut menunjukkan kapabilitas tinggi algoritma Decision Tree dan Random Forest dalam mengolah data multivariabel untuk tujuan prediktif. Oleh karena itu, kedua metode ini dipandang tepat untuk digunakan dalam penelitian ini guna memprediksi preferensi kategori lomba peserta event marathon berdasarkan karakteristik demografis, khususnya usia dan jenis kelamin.

3. METODE PENELITIAN

Penelitian ini bertujuan untuk mengembangkan model klasifikasi kategori lomba berdasarkan data peserta event marathon dari dua sumber berbeda. Tahapan metode penelitian ini terdiri atas beberapa langkah utama, yaitu pengumpulan data, preprocessing data, perancangan model, pelatihan dan evaluasi model, serta analisis perbandingan performa model. Berikut adalah tahapan secara sistematis pada Gambar 1.



Gambar 1. Alur Tahapan Penelitian

3.1. Data Collection

Data dikumpulkan dari dua event marathon yang berbeda, yaitu RSDK Berlari dan Skybridge Fun Run. Tujuannya adalah untuk memperoleh variasi data yang lebih beragam agar model yang dibangun mampu melakukan generalisasi terhadap karakteristik peserta yang heterogen. Dalam penelitian ini digunakan dua dataset, masing-masing berasal dari event Skybridge Fun Run yang terdiri dari 1.519 data peserta, serta event RSDK Berlari yang terdiri dari 1.091 data peserta. Kombinasi kedua dataset ini diharapkan dapat memberikan representasi yang lebih luas terhadap pola pemilihan kategori lomba oleh peserta dari berbagai latar belakang.

3.2. Data Integration atau Merging

Setelah data terkumpul, langkah integrasi diperlukan untuk menyatukan kedua dataset ke dalam satu kerangka kerja yang konsisten. Proses ini meliputi penyesuaian

skema kolom, normalisasi format tanggal, dan penanganan duplikasi. Tahapan ini penting untuk meminimalkan risiko bias akibat heterogenitas sumber data (I Made Putrama, 2024).

3.3.Data Preprocessing

Preprocessing merupakan tahap awal yang bertujuan untuk meningkatkan kualitas data sebelum digunakan dalam pelatihan model (Hakim, 2021). Beberapa tahapan dalam preprocessing antara lain **:**

3.3.1. Data Cleaning

Proses ini mengatasi nilai kosong (missing values) dengan metode imputasi seperti mean, median, atau modus, tergantung pada tipe data. Contoh pengisian nilai rata-rata:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x^i$$

dengan $\bar{x}$ adalah nilai rata-rata yang digunakan untuk mengganti data yang hilang.

Pembersihan menghindarkan model dari kesalahan akibat nilai yang tidak valid atau hilang (Anisa Widianti, 2024)

3.3.2. Removing Outlier

Outlier dapat mengacaukan proses pelatihan, terutama bagi Decision Tree yang sensitif terhadap nilai ekstrem. Metode deteksi seperti IQR (Interquartile Range) atau z-score dapat digunakan untuk menyingkirkan data ekstrem tersebut (Dazhi Fu, 2021).

a. Rumus Z-Score :

$$Z = \frac{(X-\mu)}{\sigma}$$

Dengan μ adalah rata-rata dan σ (sigma) adalah standar deviasi populasi.

b. Rumus IQR :

$$IQR = Q3 + Q1$$

Outlier ditentukan dengan:

$$X < Q1 - 1.5\ X\ IQR \ \ atau\ X > Q3 - 1.5\ X\ IQR$$

3.3.3. Feature Selection

Memilih atribut yang paling berkontribusi terhadap prediksi kategori lomba (misalnya usia, jenis kelamin, komunitas). Teknik seperti *Recursive Feature Elimination* (RFE) atau analisis korelasi membantu mengurangi dimensi data sekaligus mempercepat

proses training (SLN, 2023). Korelasi antar variabel dapat dihitung menggunakan koefisien Pearson:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})2 \sum(y_i - \bar{y})2}}$$

3.4. Data Splitting (Train-Test Split)

Membagi dataset menjadi subset training dan testing (misalnya 80:20) bertujuan untuk mengukur kinerja model pada data yang belum pernah dilihat. Strategi ini juga mencegah *overfitting* dengan memastikan evaluasi pada sampel yang independen (Wijiyanto, 2024).

3.5. Model Building (Decision Tree dan Random Forest)

3.5.1. Decisiom Tree

Algoritma ini membentuk struktur pohon berdasarkan pemilihan atribut terbaik melalui Information Gain atau Gini Indeks.

$$Gain\ (S.A) = Entropy(S) \sum_{i=0}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)$$

Dengan :

S : Himpunan kasus

A : Atribut

n : Jumlah partisi atribut A

$|S_i|$: Jumlah kasus pada partisi ke i

$|S|$: Jumlah kasus dalam S

$$Entropy(S) = \sum_{i=0}^{n} - pi * log_2\ pi$$

Dengan :

S : Himpunan kasus

A : Fitur

n : Jumlah partisi S

p_i : Proporsi dari S_i terhadap S

Memiliki keunggulan interpretabilitas karena struktur pohon yang mudah diikuti, namun rawan *overfitting* jika pohon terlalu dalam.

3.5.2. Random Forest

Ensemble dari beberapa pohon keputusan yang bekerja berdasarkan voting, sehingga mengurangi varian dan cenderung memberikan akurasi lebih tinggi dengan kestabilan yang lebih baik (Aufar Faiq Fadhlullah, 2024).

3.6. Model Evaluation

Evaluasi menggunakan metrik kuantitatif seperti akurasi, precision, recall, dan F1-score. Confusion matrix dikaji untuk memahami kesalahan prediksi per kelas. Penggunaan metrik-komprehensif memastikan model tidak hanya akurat secara umum, tetapi juga seimbang antar kelas (Marina Sokolova, 2020).

3.6.1. Akurasi

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy yang merupakan rasio prediksi yang benar dengan keseluruhan data. Hasil yang didapat menggambarkan seberapa akurat pengklasifikasian model dengan benar.

3.6.2. Precision

$$\text{Precission} = \frac{TP}{TP + FP}$$

Precision yaitu tingkat akurat data dari perbandingan prediksi yang benar (positif) dengan semua hasil prediksi yang benar (positif).

3.6.3. Recall

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

Recall adalah perbandingan antara prediksi benar (positif) dengan seluruh data yang benar (positif) tetapi prediksinya salah.

3.6.4. F1 – Score

$$\text{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

F1-score adalah hasil yang diperoleh untuk melihat apakah hasil precision dan recall baik atau tidak dengan membandingkan di antara keduanya. Parameter performa yang dilakukan dalam penelitian ini berupa accuracy, precision, recall, f1-score, serta waktu komputasi yaitu lama waktu proses machine learning bekerja.

a. TP (True Positive): Jumlah data yang diklasifikasikan positif oleh model dan memang benar positif. Contohnya Model memprediksi peserta memilih kategori 10K, dan ternyata benar memilih 10K.
b. TN (True Negative): Jumlah data yang diklasifikasikan negatif dan memang benar negatif. Contohnya Model memprediksi peserta tidak memilih kategori 10K, dan ternyata memang tidak.
c. FP (False Positive): Jumlah data yang diyakini positif oleh model tapi salah. Contohnya Model memprediksi peserta ikut 10K, padahal dia ikut 5K.
d. FN (False Negative): Jumlah data yang diyakini negatif oleh model tapi salah. Contohnya Model memprediksi peserta tidak ikut 10K, padahal sebenarnya ikut 10K

4. HASIL DAN PEMBAHASAN

4.1.Analisis Statistik Deskriptif Dataset

4.1.1. Dataset Sebelum dan Sesudah Preprocessing

Sebelum dilakukan proses preprocesing, dataset peserta lomba berasal dari dua sumber berbeda, yaitu RSDK dan SkyBridgeRun. Kedua dataset ini memiliki strutur kolom yang berbeda dan belum seragam. Proses preprocesing dilakukan untuk menyelaraskan nama kolom, menyamakan format penulisan gender dari L dan P menjadi M dan F, menghitung usia dari kolom Tanggal Lahir ,dan menggabungkan kedua dataset mejadi satu dataset. Gabungan dataset sebelum preprocessing seperti pada tabel .

Tabel 1. Contoh Data RSDK Sebelum Preprocessing

Tanggal Lahir	Jenis Kelamin	Kategori Lomba	Lomba
2001-05-20	P	Lari 8K	RSDK
1982-07-08	L	Lari 6K	RSDK
1983-03-18	L	Lari 6K	RSDK

Tabel 2 merupakan contoh data SkyBridgeRun sebelum dilakukan proses preprocessing.

Tabel 2. Contoh Data SkyBridgeRun Sebelum Preprocessing

Jenis Kelamin	Kategori	Tanggal Lahir
M	Lari 7.9K	1982-07-08
F	Lari 7.9K	1987-08-05
M	Lari 7.9K	1980-03-28

Pengolahan data setelah dilakukan proses preprocessing seperti pada tabel 3.

Tabel 3. Contoh Data Setelah Preprocessing

Tanggal Lahir	Gender	Kategori	Event	Usia
2001-05-20	F	Lari 8K	RSDK	24.0
1982-07-08	M	Lari 6K	RSDK	43.0
1983-08-05	F	Lari 7.9K	SkyBridgeRun	38.0
1960-04-26	M	Lari 7.9K	SkyBridgeRun	65.0
1963-08-24	F	Lari 7.9K	SkyBridgeRun	62.0

4.1.2. Statistik Deskriptif Usia dan Jenis Kelamin

Untuk mendapatkan pemahaman awal terhadap karakteristik peserta pada masing-masing kategori lomba, dilakukan analisis statistik deskriptif terhadap dua atribut utama, yaitu usia dan jenis kelamin. Tabel 4 menyajikan nilai rata-rata dan median usia peserta serta distribusi gender untuk setiap kategori lomba.

Tabel 4. Analisis Statistik Deskriptif

Kategori	Usia Mean	Usia Median	% Laki-laki	% Perempuan
Lari 11K	38.15	37.0	71.47%	28.53%
Lari 6K	34.61	35.0	59.19%	40.81%
Lari 7.9K	35.63	35.0	62.44%	37.56%

Hasil analisis menunjukkan bahwa peserta kategori Lari 11K memiliki usia rata-rata tertinggi (38,15 tahun) dan didominasi oleh peserta laki-laki (71,47%). Sebaliknya, peserta kategori 6K dan 7.9K memiliki usia rata-rata yang lebih rendah serta komposisi gender yang relatif lebih seimbang, meskipun laki-laki tetap mendominasi di semua kategori.

Perbedaan ini menunjukkan adanya kecenderungan bahwa peserta yang lebih tua dan laki-laki lebih memilih kategori dengan jarak tempuh yang lebih panjang. Temuan ini selaras dengan hasil *feature importance* dari model Random Forest, di mana variabel usia memberikan kontribusi paling dominan dalam klasifikasi kategori lomba, yakni sebesar 94%, sedangkan gender hanya menyumbang 6%.

Selain itu, ketimpangan distribusi gender pada kategori tertentu juga dapat memengaruhi performa model klasifikasi, khususnya dalam memprediksi kategori dengan representasi peserta minoritas. Ketidakseimbangan ini dapat menyebabkan model cenderung bias terhadap kelas mayoritas, sebagaimana akan dijelaskan pada bagian evaluasi model berikutnya.

4.2.Hasil Pelatihan Model

Setelah dilakukan tahapan preprocessing data, pengkodean fitur, dan pembagian dataset menjadi data latih dan uji, maka dilakukan pelatihan model menggunakan dua algoritma klasifikasi yaitu Decision Tree dan Random Forest.

Evaluasi dilakukan menggunakan metrik akurasi, precision, recall, dan f1-score. Hasil evaluasi model dapat dilihat pada Tabel 5

Tabel 5. Hasil Evaluasi Model Klasifikasi

Model	Akurasi
Decision Tree	56,81%
Random Forest	57,39%

Berdasarkan hasil pada Tabel 5, model Random Forest menunjukkan akurasi yang sedikit lebih tinggi dibandingkan Decision Tree. Namun, selisih akurasi antara keduanya relatif kecil, mengindikasikan bahwa performa model belum optimal secara keseluruhan. Untuk mendapatkan gambaran yang lebih mendalam mengenai kinerja per kategori lomba, dilakukan analisis *classification report* yang disajikan pada Tabel 6 dan Tabel 7.

Tabel 6. Classification Report Decision Tree

Kategori	Precision	Recall	F1-score	Jumlah Data
Lari 11K	0.50	0.01	0.03	67
Lari 6K	0.33	0.09	0.14	148
Lari 7.9K	0.59	0.92	0.72	306
Rata-rata makro	0.47	0.34	0.30	521
Rata-rata tertimbang	0.50	0.57	0.46	521

Dari hasil tersebut dapat dilihat bahwa model Decision Tree memiliki kinerja cukup baik pada kategori mayoritas, yaitu Lari 7.9K, dengan nilai F1-score sebesar 0.62 dan recall sebesar 0.92. Namun, pada kategori minoritas seperti Lari 6K dan Lari 11K, performa model tergolong rendah, bahkan nilai F1-score untuk Lari 11K hanya sebesar 0.03. Hal ini menunjukkan bahwa model Decision Tree belum mampu mengenali kelas-kelas minoritas dengan baik, yang kemungkinan disebabkan oleh ketidakseimbangan jumlah data antar kategori.

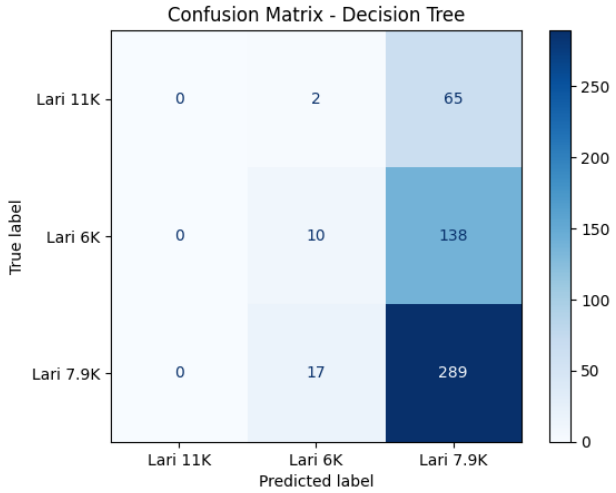
Tabel 7. Classification Report Random Forest

Kategori	Precision	Recall	F1-score	Jumlah Data
Lari 11K	0.00	0.00	0.00	67
Lari 6K	0.34	0.07	0.11	148
Lari 7.9K	0.59	0.94	0.72	306
Rata-rata makro	0.31	0.34	0.28	521
Rata-rata tertimbang	0.44	0.57	0.46	521

Dari hasil tersebut dapat disimpulkan bahwa model memiliki kinerja sangat baik pada kategori mayoritas, yaitu Lari 7.9K, dengan nilai F1-score mencapai 0,72 dan recall yang tinggi (0,94). Hal ini mengindikasikan bahwa sebagian besar prediksi benar diarahkan ke kelas ini. Sebaliknya, kinerja model terhadap kategori Lari 6K dan Lari 11K tergolong sangat rendah, bahkan model hampir tidak mampu mengenali kategori Lari 11K sama sekali (nilai precision, recall, dan F1-score = 0,00). Kondisi ini menunjukkan adanya masalah serius dalam keseimbangan kelas yang memengaruhi performa prediktif model secara keseluruhan.

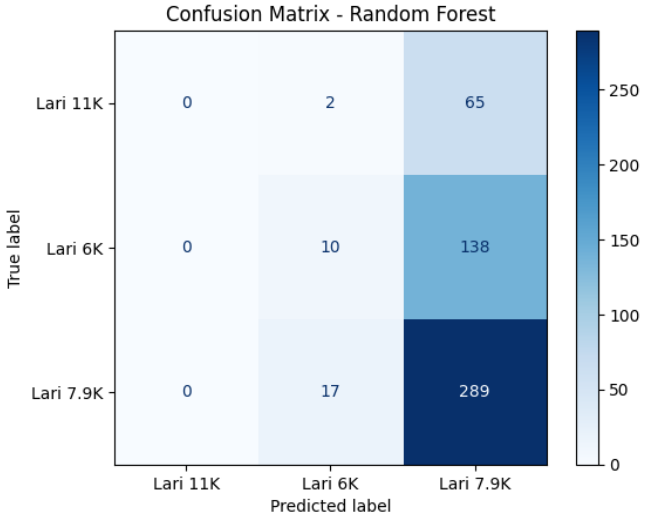
4.3.Confusion Matrix

Confusion matrix digunakan untuk melihat distribusi prediksi terhadap masing-masing kategori. Dari hasil yang ditampilkan, hampir seluruh peserta diprediksi ke kategori Lari 7.9K, meskipun secara aktual berasal dari kategori lain.



Gambar 4. Hasil Confusion Matrix Decision Tree

Berdasarkan hasil pada Gambar 4, terlihat bahwa mayoritas peserta diprediksi masuk ke dalam kategori Lari 7.9K, meskipun sejumlah besar dari mereka sebenarnya berasal dari kategori Lari 6K maupun 11K. Hal ini mencerminkan ketidakseimbangan kelas (*class imbalance*) yang cukup signifikan dalam dataset, di mana jumlah data peserta kategori 7.9K mendominasi secara proporsional.



Gambar 5. Hasil Confusion Matrix Random Forest

Hasil pada Gambar 5 juga menunjukkan pola distribusi prediksi yang sangat mirip dengan model Decision Tree. Hampir seluruh peserta diprediksi ke kategori Lari 7.9K, bahkan sebagian besar dari kategori Lari 6K dan Lari 11K pun dipetakan ke dalam kelas tersebut. Hal ini semakin memperkuat temuan bahwa ketidakseimbangan jumlah data antar kategori memberikan dampak besar terhadap performa model klasifikasi.

Ketidakseimbangan ini menyebabkan model cenderung mengarahkan prediksi ke kelas mayoritas, sehingga mengurangi akurasi dalam mengidentifikasi kelas minoritas. Konsekuensinya, kemampuan model untuk melakukan klasifikasi secara adil dan merata terhadap seluruh kelas menjadi menurun. Oleh karena itu, perlu dilakukan penanganan lebih lanjut terhadap isu ini, seperti menerapkan teknik *resampling* (oversampling atau undersampling) atau menggunakan algoritma yang lebih robust terhadap ketidakseimbangan data.

4.4.Faktor yang Paling Berpengaruh

Berdasarkan hasil analisis feature importance dari model Random Forest, diketahui bahwa fitur usia memiliki kontribusi terbesar terhadap hasil prediksi dengan nilai mencapai 94%, sedangkan gender memiliki pengaruh yang lebih rendah yakni 6%.

Studi ini menunjukkan bahwa usia merupakan faktor utama yang membedakan preferensi peserta dalam memilih kategori lomba marathon. Artinya, kecenderungan peserta dalam memilih jarak tempuh lomba lebih banyak dipengaruhi oleh rentang usia mereka, dibandingkan oleh jenis kelamin. Hal ini konsisten dengan hasil analisis statistik

deskriptif sebelumnya, di mana kategori lomba dengan jarak lebih jauh diikuti oleh peserta dengan usia rata-rata yang lebih tinggi.

5. KESIMPULAN DAN SARAN

Penelitian ini bertujuan untuk memprediksi preferensi peserta event marathon dalam memilih kategori lomba berdasarkan karakteristik demografis seperti usia dan jenis kelamin, dengan menerapkan algoritma machine learning Decision Tree dan Random Forest. Berdasarkan hasil evaluasi model, Random Forest menunjukkan performa yang lebih baik dibandingkan Decision Tree dengan akurasi sebesar 57,39%. Namun, performa tersebut belum optimal secara keseluruhan karena adanya ketidakseimbangan kelas dalam dataset, yang menyebabkan model cenderung memprediksi ke kategori mayoritas (Lari 7.9K).

Analisis feature importance menunjukkan bahwa usia merupakan faktor dominan yang memengaruhi keputusan peserta dalam memilih kategori lomba, dengan kontribusi sebesar 94%, sedangkan jenis kelamin hanya memberikan kontribusi 6%. Hal ini mengindikasikan bahwa preferensi peserta lebih banyak dipengaruhi oleh rentang usia dibandingkan gender.

Secara umum, penggunaan algoritma Random Forest cukup menjanjikan dalam membangun model prediktif preferensi peserta marathon, namun perlu dilakukan perbaikan lebih lanjut, seperti penanganan imbalance data dan eksplorasi fitur tambahan, untuk meningkatkan akurasi dan generalisasi model.

DAFTAR REFERENSI

Ajmi, N. A. A., & Subhi, M. (2021). A review of big data analytic in healthcare. *Turkish Journal of Computer and Mathematics Education*, *12*(3), 4542–4548.

Anisa Widianti, I. P. (2024). Penanganan missing values dan prediksi data timbunan sampah berbasis machine learning. *RABIT: Jurnal Teknologi dan Sistem Informasi Univra*, *3*(1), 242–251.

Aufar Faiq Fadhlullah, & Widodo, T. W. (2024). Comparative analysis of decision tree and random forest algorithms for diabetes prediction. *JTAM (Jurnal Teori dan Aplikasi Matematika)*, *8*(2), 1121–1132.

Bharathi, V. P. N., & Mohan, K. (2025). Using machine learning models to forecast methane emissions from agriculture in India. *Plant Science Today*, *12*(1), 3.

Darmawan, M. B. A., & Dewi, F. (2023). Analisis perbandingan algoritma decision tree, random forest, dan naïve Bayes untuk prediksi banjir di Desa Dayeuhkolot. *Proceedings of the 2023 International Conference on Machine Learning and Automation*, 52.

Fu, D., & Zhao, Z. (2021). Dense projection for anomaly detection. In *The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)* (pp. 670–678).

Hakim, B. (2021). Data text pre-processing sentiment analysis in data mining using machine learning. *JBASE Journal of Business and Audit Information Systems*, *9*(1), 16.

Jaiswal, J. K. (2021). Application of random forest algorithm on feature subset selection and classification and regression. In *World Congress on Computing and Communication Technologies (WCCCT)* (pp. 65–72).

Jia, M., & Zhang, Y. (2020). A random forest regression model predicting the winners of summer Olympic events. In *International Conference on Big Data Engineering* (pp. 62–69).

Lundy, L., & Borrie, R. (2024). Demographics, culture and participatory nature of multi-marathoning: An observational study highlighting issues with recommendations. *PLOS ONE*, *19*(3), 1–14.

Marina Sokolova, & Lapalme, G. (2020). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *56*(3), 427–437.

Massimo Giotta, & Testa, P. (2022). Application of a decision tree model to predict the outcome of non-intensive inpatients hospitalized for COVID-19. *International Journal of Environmental Research and Public Health*, *19*(2), 1–12.

Pakawan Pugsee, & Phetkaew, P. (2020). Football match result prediction using the random forest classifier. In *Proceedings of the 2020 2nd International Conference on Big Data Engineering* (pp. 62–69).

Putrama, I. M., & Muliarta, P. (2024). Heterogeneous data integration: Challenges and opportunities. *Data in Brief*, *49*, 1–21.

Simanjuntak, S. P., & Ambarita, R. (2025). Collaborative governance dalam pelaksanaan event Toba Marathon Festival di Kabupaten Toba tahun 2023. *Journal of Governance and Policy*, *10*(1), 121–131.

SLN, F. (2023). *Basic data mining from A to Z*. Bandung: ResearchGate.

Wijiyanto, A. I. (2024). Teknik K-fold cross validation untuk mengevaluasi kinerja mahasiswa. *Jurnal Algoritma*, *21*(2), 241–248.

Wittho, A., & Meier, T. M. (2024). Running trends in Switzerland from 1999 to 2019: An exploratory observational study. *PLOS ONE*, *19*(4), 1–19.