

Klasifikasi Biner untuk Prediksi Keselamatan Penumpang Titanic: Perbandingan Algoritma Logistic Regression, Decision Tree, dan Random Forest

Dian Ayu Kusri¹, Ay Kumalaka², Nur Adila³

^{1,2,3}Program Studi : Sistem Informasi, Universitas Darwan Ali

Email : dianayukusri@gmail.com¹, aykumalaka26@gmail.com², nura08052@gmail.com³

ABSTRACT— Binary classification is one of the most widely used supervised machine learning techniques for predicting outcomes with two possible classes. The Titanic passenger dataset has become a benchmark dataset for evaluating classification algorithms because it contains demographic and socioeconomic information related to passenger survival. This study aims to compare the performance of three machine learning algorithms, namely Logistic Regression, Decision Tree, and Random Forest, in binary classification of Titanic passenger survival, while also investigating the possibility of data leakage that may affect model evaluation. The dataset consists of 418 passenger records with 12 attributes, including passenger class, age, sex, fare, family relationships, and embarkation information. Data preprocessing included missing value imputation, removal of irrelevant features, categorical feature encoding, and an 80:20 stratified train-test split. Standardization was applied only to the Logistic Regression model, while Decision Tree and Random Forest were trained using the original feature scales. Model performance was evaluated using accuracy, precision, recall, F1-score, confusion matrix, correlation analysis, and Random Forest feature importance. Experimental results showed that all three algorithms achieved perfect performance with accuracy, precision, recall, and F1-score values of 1.00 on the testing dataset. Further investigation revealed that these exceptional results were caused by data leakage, where the Sex feature exhibited a feature importance score of 0.83 and showed a perfect correlation with the target variable Survived. This finding indicates that the dataset used is likely a derived dataset generated using a gender-based labeling rule rather than the original historical survival records. Therefore, the study concludes that perfect classification performance should not be interpreted as evidence of superior model capability without first verifying data quality, label provenance, and the presence of potential data leakage. These findings emphasize the importance of comprehensive dataset validation before drawing conclusions from machine learning experiments.

Keywords— Binary Classification, Titanic Dataset, Logistic Regression, Random Forest, Data Leakage, Machine Learning.

ABSTRAK— Klasifikasi biner merupakan salah satu teknik *supervised machine learning* yang banyak digunakan untuk memprediksi keluaran ke dalam dua kelas berdasarkan pola pada data. Dataset Titanic menjadi salah satu dataset yang paling sering digunakan sebagai media pembelajaran dan evaluasi algoritma klasifikasi karena memuat informasi demografis dan sosial ekonomi penumpang yang berkaitan dengan status keselamatan. Penelitian ini bertujuan untuk membandingkan kinerja tiga algoritma klasifikasi, yaitu Logistic Regression, Decision Tree, dan Random Forest, dalam memprediksi keselamatan penumpang Titanic serta menganalisis kemungkinan terjadinya data leakage yang dapat memengaruhi hasil evaluasi model. Dataset yang digunakan terdiri atas 418 data penumpang dengan 12 atribut, meliputi kelas penumpang, usia, jenis kelamin, tarif, jumlah anggota keluarga, dan pelabuhan keberangkatan. Tahap pra-proses data meliputi penanganan nilai hilang, penghapusan atribut yang tidak relevan, pengkodean data kategorikal, serta pembagian data latih dan data uji menggunakan metode stratified sampling dengan rasio 80:20. Standarisasi data diterapkan khusus pada model Logistic Regression, sedangkan Decision Tree dan Random Forest dilatih menggunakan data tanpa standarisasi. Kinerja model dievaluasi menggunakan metrik akurasi, presisi, recall, F1-score, *confusion matrix*, analisis korelasi, serta *feature importance* pada Random Forest. Hasil penelitian menunjukkan bahwa ketiga algoritma menghasilkan nilai akurasi, presisi, recall, dan F1-score sebesar 1,00 pada data uji. Namun, analisis lebih lanjut menunjukkan bahwa hasil tersebut disebabkan oleh data leakage, di mana atribut Sex memiliki nilai *feature importance* sebesar 0,83 dan berkorelasi sempurna dengan label Survived. Temuan ini mengindikasikan bahwa dataset yang digunakan kemungkinan merupakan dataset turunan yang label targetnya dibentuk berdasarkan aturan sederhana berbasis jenis kelamin, sehingga akurasi sempurna tidak mencerminkan kemampuan model yang sesungguhnya. Oleh karena itu, penelitian ini menegaskan pentingnya validasi kualitas data, pemeriksaan hubungan antara fitur dan target, serta deteksi data leakage sebelum menarik kesimpulan mengenai performa algoritma *machine learning*.

Kata kunci— Klasifikasi Biner, Titanic, Logistic Regression, Random Forest, Data Leakage, Machine Learning.

I. PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi dalam beberapa dekade terakhir telah mendorong peningkatan pemanfaatan Artificial Intelligence (AI) dan Machine Learning (ML) dalam berbagai bidang kehidupan. Machine learning merupakan cabang dari kecerdasan buatan yang memungkinkan komputer mempelajari pola dari data untuk menghasilkan prediksi atau keputusan tanpa harus diprogram secara eksplisit. Kemampuan machine learning dalam mengolah data berukuran besar dan menemukan pola yang kompleks menjadikannya sebagai salah satu teknologi utama dalam mendukung pengambilan keputusan berbasis data (*data-driven decision making*). Saat ini, machine learning telah diterapkan secara luas pada berbagai sektor, seperti kesehatan untuk diagnosis penyakit, keuangan untuk deteksi penipuan transaksi, pendidikan untuk analisis tingkat kelulusan mahasiswa, pemasaran untuk rekomendasi produk, hingga transportasi untuk prediksi risiko kecelakaan.

Salah satu permasalahan yang paling sering diselesaikan menggunakan machine learning adalah klasifikasi biner, yaitu proses mengelompokkan suatu objek ke dalam dua kategori berdasarkan karakteristik yang dimilikinya. Pada klasifikasi biner, setiap data hanya memiliki dua kemungkinan kelas, misalnya "ya" atau "tidak", "positif" atau "negatif", serta "selamat" atau "tidak selamat". Keberhasilan suatu model klasifikasi sangat dipengaruhi oleh kualitas data, pemilihan fitur, proses pra-proses data, serta algoritma yang digunakan. Oleh karena itu, pemilihan algoritma klasifikasi yang sesuai menjadi salah satu faktor penting untuk memperoleh model yang memiliki kemampuan prediksi yang baik.

Dataset Titanic merupakan salah satu dataset yang paling populer dalam pembelajaran machine learning maupun penelitian klasifikasi. Dataset ini berisi informasi mengenai penumpang kapal RMS Titanic yang tenggelam pada tahun 1912, meliputi kelas penumpang (*Pclass*), jenis kelamin (*Sex*), usia (*Age*), tarif (*Fare*), jumlah anggota keluarga (*SibSp* dan *Parch*), pelabuhan keberangkatan (*Embarked*), serta status keselamatan penumpang (*Survived*). Karena memiliki kombinasi data numerik dan kategorikal, dataset Titanic sering digunakan sebagai media pembelajaran untuk membandingkan performa berbagai algoritma klasifikasi.

Berbagai algoritma machine learning telah diterapkan untuk menyelesaikan permasalahan klasifikasi pada dataset Titanic. Logistic Regression merupakan algoritma yang sederhana, mudah diinterpretasikan, dan sering digunakan sebagai model dasar (*baseline*) pada klasifikasi biner. Algoritma ini bekerja dengan menghitung probabilitas suatu data termasuk ke dalam salah satu kelas menggunakan fungsi logistik. Meskipun memiliki struktur yang relatif sederhana, Logistic Regression mampu memberikan performa yang baik pada data yang memiliki hubungan linier antara variabel prediktor dan variabel target.

Selain Logistic Regression, algoritma Decision Tree juga banyak digunakan karena mampu menghasilkan model

yang mudah dipahami. Decision Tree membangun struktur pohon keputusan berdasarkan proses pemilihan atribut terbaik pada setiap percabangan sehingga menghasilkan aturan (*rule*) yang dapat dijelaskan secara visual. Namun demikian, Decision Tree memiliki kelemahan berupa kecenderungan mengalami overfitting, terutama ketika pohon dibangun terlalu kompleks sehingga mampu menghafal data latih tetapi kurang baik dalam melakukan prediksi terhadap data baru.

Untuk mengatasi kelemahan tersebut, dikembangkan algoritma Random Forest yang merupakan metode *ensemble learning*. Random Forest membangun sejumlah pohon keputusan secara acak (*bootstrap aggregating*) kemudian menggabungkan hasil prediksi seluruh pohon melalui mekanisme *majority voting*. Pendekatan ini mampu meningkatkan stabilitas model, mengurangi risiko overfitting, serta menghasilkan performa yang lebih baik dibandingkan Decision Tree tunggal pada berbagai kasus klasifikasi.

Beberapa penelitian sebelumnya menunjukkan bahwa Random Forest umumnya menghasilkan tingkat akurasi yang lebih tinggi dibandingkan Logistic Regression dan Decision Tree pada berbagai domain aplikasi. Namun demikian, performa suatu algoritma tidak hanya dipengaruhi oleh karakteristik algoritma itu sendiri, tetapi juga oleh kualitas dataset, distribusi kelas, teknik pra-proses data, serta pemilihan fitur. Oleh karena itu, hasil evaluasi model perlu dianalisis secara menyeluruh sebelum digunakan sebagai dasar penarikan kesimpulan. Salah satu permasalahan yang sering diabaikan dalam proses pengembangan model machine learning adalah data leakage. Data leakage merupakan kondisi ketika informasi yang seharusnya hanya tersedia pada variabel target secara tidak sengaja ikut digunakan sebagai fitur prediktor selama proses pelatihan model. Akibatnya, model memperoleh informasi yang tidak semestinya sehingga menghasilkan nilai evaluasi yang sangat tinggi tetapi tidak mampu melakukan generalisasi terhadap data baru. Fenomena ini dapat menyebabkan peneliti menarik kesimpulan yang keliru mengenai kemampuan algoritma yang digunakan.

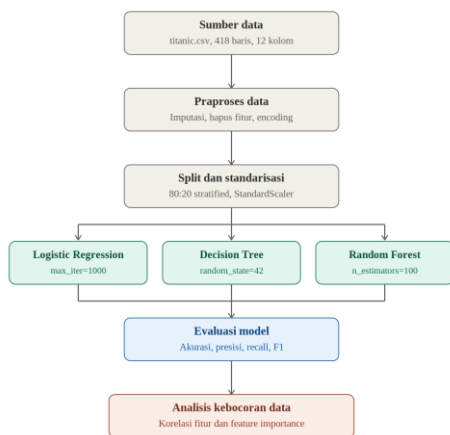
Pada penelitian ini, dilakukan perbandingan tiga algoritma klasifikasi, yaitu Logistic Regression, Decision Tree, dan Random Forest, menggunakan dataset penumpang Titanic yang terdiri atas 418 data. Proses penelitian meliputi tahapan pra-proses data, pembagian data latih dan data uji menggunakan stratified sampling, pelatihan model, serta evaluasi menggunakan metrik akurasi, presisi, recall, F1-score, dan *confusion matrix*. Selain itu, dilakukan pula analisis korelasi antarfitur dan analisis *feature importance* untuk mengidentifikasi kemungkinan terjadinya data leakage yang dapat memengaruhi hasil klasifikasi.

Hasil penelitian menunjukkan bahwa ketiga algoritma menghasilkan nilai akurasi, presisi, *recall*, dan *F1-score* sebesar 1,00 pada data uji. Meskipun secara sekilas hasil tersebut menunjukkan performa yang sangat baik, analisis lebih lanjut memperlihatkan adanya hubungan yang sangat kuat antara atribut Sex dengan label Survived, sehingga mengindikasikan adanya data leakage pada dataset yang digunakan. Oleh karena itu, penelitian ini

tidak hanya bertujuan membandingkan performa tiga algoritma klasifikasi, tetapi juga memberikan pemahaman mengenai pentingnya validasi kualitas data sebelum melakukan interpretasi terhadap hasil evaluasi model. Temuan penelitian ini diharapkan dapat menjadi referensi bagi mahasiswa, peneliti, maupun praktisi data science dalam membangun model klasifikasi yang lebih andal, valid, dan memiliki kemampuan generalisasi yang lebih baik.

II. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen (*experimental research*) untuk membandingkan kinerja tiga algoritma klasifikasi biner, yaitu Logistic Regression, Decision Tree, dan Random Forest. Seluruh tahapan penelitian dilakukan secara sistematis mulai dari pengumpulan data, praproses data, pembagian data latih dan data uji, pembangunan model, evaluasi model, hingga analisis hasil klasifikasi. Alur penelitian secara keseluruhan ditunjukkan pada Gambar 1.



Gambar 1. Alur Metodologi Penelitian

A. Sumber Data

Dataset yang digunakan dalam penelitian ini adalah Titanic Dataset (titanic.csv) yang terdiri atas 418 data penumpang dengan 12 atribut, yaitu PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, dan Embarked. Dataset ini dipilih karena merupakan salah satu dataset klasifikasi yang paling banyak digunakan dalam penelitian machine learning sehingga memudahkan proses evaluasi dan perbandingan performa algoritma.

Variabel target (target variable) pada penelitian ini adalah Survived, yang terdiri atas dua kelas, yaitu 0 (tidak selamat) dan 1 (selamat). Sementara itu, atribut lainnya digunakan sebagai variabel prediktor yang berfungsi untuk membangun model klasifikasi. Dataset tersebut memiliki kombinasi atribut numerik dan kategorikal sehingga memerlukan tahapan praproses sebelum digunakan dalam proses pelatihan model.

B. Praproses Data

Tahap praproses dilakukan untuk meningkatkan kualitas data sehingga dapat digunakan secara optimal oleh

algoritma machine learning. Tahapan praproses yang dilakukan meliputi beberapa langkah sebagai berikut.

1. Penanganan Missing Value

Beberapa atribut pada dataset memiliki nilai yang hilang (missing value), terutama pada atribut Age dan Fare. Nilai yang hilang pada kedua atribut tersebut diisi menggunakan median karena metode ini lebih tahan terhadap keberadaan outlier dibandingkan rata-rata (mean). Sementara itu, atribut Cabin memiliki proporsi nilai kosong yang sangat tinggi sehingga dihapus dari proses analisis karena dianggap tidak mampu memberikan informasi yang cukup bagi model.

2. Penghapusan Atribut Tidak Relevan

Tidak seluruh atribut pada dataset memiliki kontribusi terhadap proses klasifikasi. Oleh karena itu, atribut PassengerId, Name, dan Ticket dihapus karena hanya berfungsi sebagai identitas penumpang dan tidak memiliki hubungan langsung dengan status keselamatan.

3. Transformasi Data Kategorikal

Algoritma machine learning pada umumnya hanya dapat memproses data numerik. Oleh karena itu, atribut kategorikal seperti Sex dan Embarked dikonversi menjadi bentuk numerik menggunakan teknik Label Encoding. Teknik ini memberikan representasi angka pada setiap kategori sehingga dapat diproses oleh algoritma klasifikasi tanpa mengubah jumlah data.

C. Pembagian Data

Setelah proses praproses selesai, dataset dibagi menjadi data latih (training data) dan data uji (testing data) menggunakan metode Stratified Train-Test Split dengan rasio 80:20. Teknik stratifikasi dipilih untuk memastikan bahwa proporsi kelas pada data latih dan data uji tetap seimbang sehingga hasil evaluasi model menjadi lebih representatif.

Hasil pembagian menghasilkan 334 data latih dan 84 data uji. Data latih digunakan untuk membangun model, sedangkan data uji digunakan untuk mengevaluasi kemampuan model dalam melakukan prediksi terhadap data yang belum pernah dilihat sebelumnya.

D. Standardisasi Data

Tahap standardisasi dilakukan khusus pada algoritma Logistic Regression menggunakan metode StandardScaler. Standardisasi bertujuan mengubah skala setiap atribut numerik sehingga memiliki rata-rata mendekati nol dan simpangan baku mendekati satu. Langkah ini diperlukan karena Logistic Regression merupakan algoritma yang sensitif terhadap perbedaan skala antarfitur.

Sebaliknya, algoritma Decision Tree dan Random Forest tidak memerlukan proses standardisasi karena kedua algoritma tersebut membangun model berdasarkan proses pemisahan (splitting) atribut, bukan berdasarkan perhitungan jarak antar data.

E. Pembangunan Model

Penelitian ini membandingkan tiga algoritma klasifikasi yang diimplementasikan menggunakan pustaka Scikit-learn pada bahasa pemrograman Python.

Logistic Regression digunakan sebagai model dasar (baseline model) dengan parameter max_iter = 1000 agar proses optimasi dapat mencapai konvergensi.

Decision Tree Classifier digunakan dengan parameter `random_state = 42` untuk memastikan hasil eksperimen dapat direproduksi.

Random Forest Classifier dibangun menggunakan 100 pohon keputusan (`n_estimators = 100`) dan `random_state = 42`. Algoritma ini bekerja dengan menggabungkan hasil prediksi dari banyak pohon keputusan sehingga menghasilkan model yang lebih stabil dan memiliki risiko *overfitting* yang lebih rendah dibandingkan Decision Tree tunggal.

Seluruh model dilatih menggunakan data latih yang sama sehingga perbandingan performa dilakukan secara adil dan konsisten.

F. Evaluasi Model

Evaluasi performa model dilakukan menggunakan beberapa metrik klasifikasi yang umum digunakan pada penelitian machine learning, yaitu:

1. Accuracy, untuk mengukur proporsi prediksi yang benar terhadap seluruh data uji.
2. Precision, untuk mengukur ketepatan model dalam memprediksi kelas positif.
3. Recall, untuk mengukur kemampuan model menemukan seluruh data positif.
4. F1-Score, yaitu rata-rata harmonis antara precision dan recall sehingga memberikan ukuran performa yang lebih seimbang.

Selain keempat metrik tersebut, penelitian ini juga menggunakan Confusion Matrix untuk melihat distribusi prediksi benar dan salah pada masing-masing kelas. Analisis ini memberikan gambaran yang lebih rinci mengenai kemampuan model dibandingkan hanya menggunakan nilai akurasi.

G. Analisis Diagnostik

Tahap akhir penelitian adalah melakukan analisis diagnostik terhadap hasil klasifikasi. Analisis ini dilakukan karena seluruh algoritma menghasilkan nilai akurasi yang sangat tinggi. Untuk mengetahui penyebabnya, dilakukan dua analisis tambahan, yaitu analisis korelasi antarfitur dan analisis feature importance menggunakan Random Forest.

Analisis korelasi digunakan untuk mengidentifikasi hubungan antara setiap atribut dengan variabel target, sedangkan feature importance digunakan untuk mengetahui atribut yang paling berpengaruh dalam proses pengambilan keputusan model. Hasil kedua analisis tersebut digunakan untuk mendeteksi kemungkinan terjadinya data leakage yang dapat menyebabkan nilai evaluasi model menjadi terlalu tinggi dan tidak mencerminkan kemampuan generalisasi yang sebenarnya.

III. DESAIN, HASIL DAN PEMBAHASAN

Jika penelitian dilakukan dengan menghasilkan produk, maka desain produk dapat dibahas dalam subbab ini. Dengan memperlihatkan rancangan sistem yang akan dibangun. Kemudian penjelasan sekilas mengenai rancangan tersebut. Misal rancangan berupa gambar Konteks Diagram. Perhatikan gambar 2 yang menunjukkan rancangan konteks diagram dibawah ini.

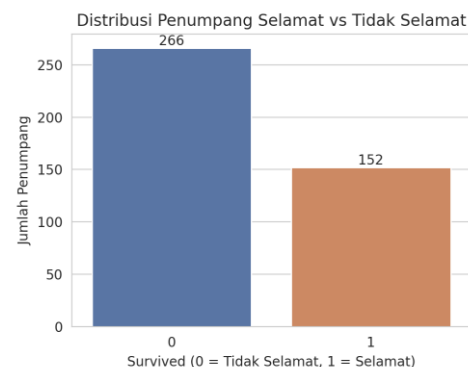
A. Eksplorasi Data

Tahap eksplorasi data (*Exploratory Data Analysis/EDA*) dilakukan untuk memahami karakteristik dataset sebelum proses pemodelan. Dataset yang digunakan terdiri atas 418 data penumpang dengan 12 atribut, yang meliputi *PassengerId*, *Survived*, *Pclass*, *Name*, *Sex*, *Age*, *SibSp*, *Parch*, *Ticket*, *Fare*, *Cabin*, dan *Embarked*. Setelah proses *praproses*, atribut yang tidak relevan dihapus sehingga hanya atribut yang memiliki potensi sebagai prediktor yang digunakan dalam proses pelatihan model.

Distribusi kelas target menunjukkan bahwa terdapat 266 penumpang (63,6%) dengan status tidak selamat (*Survived* = 0) dan 152 penumpang (36,4%) dengan status selamat (*Survived* = 1). Distribusi ini menunjukkan adanya ketidakseimbangan kelas (*class imbalance*) dalam tingkat yang masih dapat diterima untuk proses klasifikasi biner.

Meskipun demikian, distribusi tersebut tetap perlu diperhatikan karena model klasifikasi cenderung menghasilkan bias terhadap kelas mayoritas apabila ketidakseimbangan kelas terlalu tinggi. Oleh karena itu, penelitian ini menggunakan metode *Stratified Train-Test Split* sehingga proporsi kedua kelas tetap terjaga pada data latih maupun data uji. Pendekatan ini membantu menghasilkan evaluasi model yang lebih representatif dibandingkan pembagian data secara acak tanpa stratifikasi.

Selain distribusi kelas, eksplorasi data juga menunjukkan bahwa atribut *Age* dan *Fare* memiliki beberapa nilai yang hilang (*missing values*). Nilai tersebut telah ditangani menggunakan metode imputasi median sehingga tidak mengurangi jumlah data yang digunakan dalam proses pelatihan model.



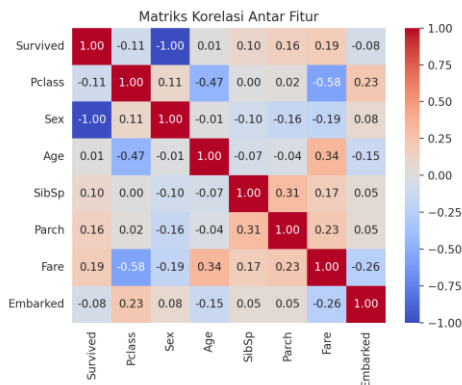
Gambar 2. Distribusi Penumpang Selamat vs Tidak Selamat

Berdasarkan Gambar 2 terlihat bahwa jumlah penumpang dengan status tidak selamat lebih banyak dibandingkan penumpang yang selamat. Distribusi tersebut menunjukkan bahwa dataset memiliki ketidakseimbangan kelas (*class imbalance*) dalam tingkat yang masih dapat diterima. Oleh karena itu, penelitian ini menggunakan metode *Stratified Train-Test Split* agar proporsi masing-masing kelas tetap terjaga pada data latih maupun data uji.

B. Analisis Korelasi Antar Fitur

Analisis korelasi dilakukan untuk mengetahui hubungan antara setiap atribut dengan variabel target (*Survived*). Korelasi dihitung menggunakan matriks korelasi

(*correlation matrix*) yang divisualisasikan dalam bentuk *heatmap*.



Gambar 3. Matriks Korelasi Antar Fitur

Hasil analisis menunjukkan bahwa atribut Sex memiliki hubungan yang paling kuat terhadap variabel target dibandingkan atribut lainnya. Sebaliknya, atribut Age, Fare, Pclass, SibSp, dan Parch hanya menunjukkan hubungan yang relatif lemah terhadap status keselamatan penumpang.

Secara umum, kondisi tersebut berbeda dengan karakteristik dataset Titanic asli yang biasanya memperlihatkan pengaruh beberapa atribut secara bersamaan terhadap peluang keselamatan penumpang. Oleh karena itu, hubungan yang sangat dominan antara atribut Sex dan Survived menjadi indikasi awal bahwa terdapat karakteristik khusus pada dataset yang digunakan sehingga memerlukan analisis lebih lanjut.

Temuan ini menunjukkan bahwa analisis korelasi merupakan tahap penting sebelum membangun model machine learning karena dapat membantu peneliti mengidentifikasi atribut yang berpotensi menyebabkan bias ataupun kebocoran informasi (*data leakage*).

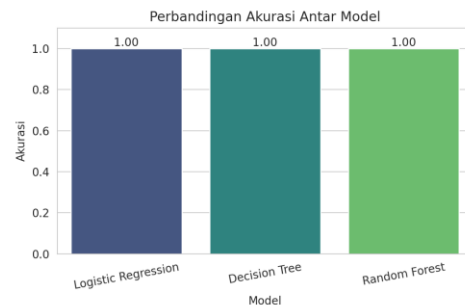
C. Hasil Klasifikasi

Setelah seluruh tahapan praproses selesai dilakukan, tiga algoritma klasifikasi, yaitu Logistic Regression, Decision Tree, dan Random Forest, dilatih menggunakan data latih dan dievaluasi pada data uji.

Hasil evaluasi menunjukkan bahwa ketiga algoritma menghasilkan nilai akurasi, presisi, recall, dan F1-score sebesar 1,00. Nilai tersebut menunjukkan bahwa seluruh data uji berhasil diprediksi dengan benar tanpa menghasilkan kesalahan klasifikasi.

TABEL I
PERBANDINGAN KINERJA MODEL PADA DATA UJI

Model	Akurasi	Presisi	Recall	F1-Score
Logistic Regression	1,00	1,00	1,00	1,00
Decision Tree	1,00	1,00	1,00	1,00
Random Forest	1,00	1,00	1,00	1,00



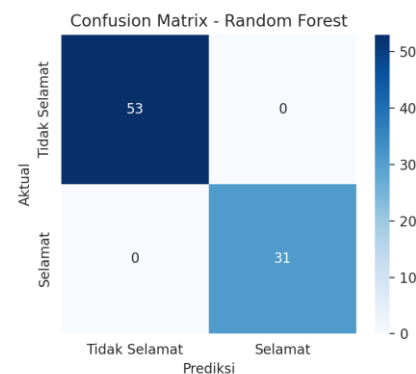
Gambar 4. Perbandingan Akurasi Antar Model

Secara teoritis, hasil tersebut mengindikasikan bahwa model memiliki performa yang sangat baik. Akan tetapi, dalam penelitian machine learning, nilai evaluasi yang terlalu sempurna justru perlu dianalisis lebih lanjut. Pada dataset dunia nyata, hampir tidak pernah ditemukan model yang mampu menghasilkan akurasi 100% karena selalu terdapat variasi karakteristik data, *noise*, maupun faktor lain yang memengaruhi proses prediksi.

Oleh karena itu, hasil evaluasi pada penelitian ini tidak langsung diinterpretasikan sebagai bukti bahwa ketiga algoritma memiliki performa terbaik, melainkan dijadikan dasar untuk melakukan analisis lebih mendalam terhadap kualitas dataset yang digunakan.

D. Analisis Confusion Matrix

Analisis menggunakan Confusion Matrix menunjukkan bahwa seluruh data pada kelas Survived = 0 maupun Survived = 1 berhasil diprediksi secara benar oleh model Random Forest. Tidak ditemukan False Positive (FP) maupun False Negative (FN) pada data uji.



Gambar 5. Confusion Matrix Model Random Forest

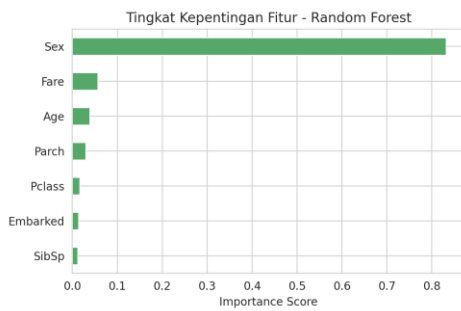
Kondisi tersebut menyebabkan nilai Precision, Recall, dan F1-score mencapai nilai maksimum, yaitu 1,00. Meskipun hasil tersebut tampak ideal, Confusion Matrix saja belum cukup untuk menyimpulkan bahwa model memiliki kemampuan generalisasi yang tinggi.

Dalam praktik machine learning, Confusion Matrix perlu diinterpretasikan bersama karakteristik dataset. Oleh karena itu, penelitian ini melanjutkan analisis menggunakan Feature Importance untuk mengetahui atribut yang paling dominan dalam proses pengambilan keputusan model.

E. Analisis Feature Importance dan Data Leakage

Hasil analisis menggunakan algoritma Random Forest menunjukkan bahwa atribut Sex memiliki nilai Feature

Importance sebesar 0,83, jauh lebih tinggi dibandingkan atribut lainnya seperti Fare, Age, maupun Pclass.



Gambar 6. Tingkat Kepentingan Fitur pada Model Random Forest

Dominasi atribut tersebut menunjukkan bahwa sebagian besar keputusan model ditentukan oleh informasi mengenai jenis kelamin penumpang. Untuk memastikan penyebab kondisi tersebut, dilakukan tabulasi silang antara atribut Sex dengan variabel target Survived.

Hasil tabulasi memperlihatkan bahwa seluruh penumpang perempuan memiliki label selamat, sedangkan seluruh penumpang laki-laki memiliki label tidak selamat. Dengan demikian, atribut Sex memiliki hubungan sempurna terhadap label target.

Kondisi tersebut merupakan indikasi kuat adanya data leakage, yaitu ketika informasi target secara langsung terkandung dalam salah satu atribut prediktor. Akibatnya, model tidak benar-benar mempelajari pola dari berbagai atribut, melainkan hanya mengandalkan satu atribut yang telah memiliki hubungan sempurna dengan label.

Temuan ini menjelaskan mengapa seluruh algoritma menghasilkan akurasi 100%, meskipun karakteristik ketiga algoritma berbeda.

F. Pembahasan

Hasil penelitian menunjukkan bahwa Logistic Regression, Decision Tree, dan Random Forest sama-sama menghasilkan performa yang sempurna pada dataset yang digunakan. Namun, hasil tersebut tidak dapat dijadikan dasar untuk menyatakan bahwa ketiga algoritma memiliki kemampuan klasifikasi yang sama baiknya.

Analisis lebih lanjut membuktikan bahwa tingginya nilai evaluasi disebabkan oleh karakteristik dataset, bukan oleh keunggulan algoritma. Dengan adanya hubungan sempurna antara atribut Sex dan Survived, model tidak perlu memanfaatkan atribut lain untuk menghasilkan prediksi yang benar.

Temuan ini memperkuat pentingnya proses validasi dataset sebelum melakukan pemodelan. Dalam penelitian machine learning, kualitas data memiliki pengaruh yang sama pentingnya dengan pemilihan algoritma. Dataset yang mengandung data leakage dapat menghasilkan nilai evaluasi yang sangat tinggi tetapi tidak mampu merepresentasikan kemampuan model pada data baru.

Oleh karena itu, penelitian ini menegaskan bahwa evaluasi model tidak boleh hanya didasarkan pada nilai akurasi. Analisis terhadap korelasi fitur, *feature importance*, *confusion matrix*, serta asal-usul label target perlu dilakukan untuk memastikan bahwa model benar-benar mempelajari pola dari data dan memiliki kemampuan generalisasi yang baik. Dengan demikian,

hasil penelitian ini tidak hanya memberikan perbandingan tiga algoritma klasifikasi, tetapi juga menjadi studi kasus mengenai pentingnya deteksi data leakage dalam penelitian machine learning.

G. Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan dalam menginterpretasikan hasil yang diperoleh. Pertama, dataset yang digunakan hanya terdiri atas 418 data penumpang, sehingga jumlah data relatif terbatas dibandingkan dataset Titanic asli yang tersedia pada kompetisi Kaggle. Jumlah data yang lebih sedikit dapat memengaruhi kemampuan model dalam mempelajari variasi karakteristik penumpang.

Kedua, hasil evaluasi menunjukkan bahwa seluruh algoritma memperoleh nilai akurasi, presisi, *recall*, dan *F1-score* sebesar 1,00. Setelah dilakukan analisis lebih lanjut, diketahui bahwa kondisi tersebut disebabkan oleh adanya data leakage, yaitu hubungan yang sangat kuat antara atribut Sex dan label Survived. Akibatnya, model tidak benar-benar mempelajari hubungan dari seluruh atribut, melainkan lebih banyak bergantung pada satu atribut yang telah memiliki korelasi sempurna dengan variabel target.

Ketiga, penelitian ini hanya membandingkan tiga algoritma klasifikasi, yaitu Logistic Regression, Decision Tree, dan Random Forest. Masih terdapat berbagai algoritma lain seperti Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbor (KNN), Extreme Gradient Boosting (XGBoost), maupun Artificial Neural Network (ANN) yang belum dievaluasi. Perbandingan dengan algoritma tersebut berpotensi memberikan gambaran yang lebih luas mengenai performa metode klasifikasi pada dataset Titanic.

Keempat, proses optimasi parameter (*hyperparameter tuning*) belum dilakukan secara mendalam. Penelitian ini menggunakan parameter standar dari pustaka Scikit-learn sehingga performa model yang diperoleh belum tentu merupakan performa terbaik yang dapat dicapai. Penerapan teknik optimasi seperti Grid Search atau Random Search dapat digunakan pada penelitian selanjutnya untuk memperoleh kombinasi parameter yang lebih optimal.

Dengan mempertimbangkan berbagai keterbatasan tersebut, hasil penelitian ini lebih tepat dipandang sebagai studi komparatif dan studi kasus mengenai data leakage, bukan sebagai penentu algoritma terbaik untuk klasifikasi keselamatan penumpang Titanic.

H. Implikasi Penelitian

Penelitian ini memberikan beberapa implikasi penting, baik dari sisi akademik maupun praktik penerapan machine learning. Dari sisi akademik, hasil penelitian menunjukkan bahwa nilai evaluasi model yang sangat tinggi tidak selalu mencerminkan kemampuan prediksi yang sebenarnya. Oleh karena itu, peneliti perlu melakukan pemeriksaan terhadap kualitas data, distribusi kelas, hubungan antarfitur, serta asal-usul label target sebelum menarik kesimpulan mengenai performa algoritma.

Dari sisi metodologi, penelitian ini menegaskan bahwa tahapan praproses data memiliki peran yang sangat

penting dalam pembangunan model machine learning. Penanganan *missing value*, pemilihan atribut, transformasi data kategorikal, serta pemeriksaan kemungkinan data leakage harus menjadi bagian yang tidak terpisahkan dari proses pengembangan model. Tanpa tahapan tersebut, model dapat menghasilkan nilai evaluasi yang tinggi tetapi gagal melakukan prediksi secara akurat pada data baru.

Selain itu, penelitian ini juga memberikan kontribusi sebagai media pembelajaran bagi mahasiswa dan peneliti yang mempelajari klasifikasi biner menggunakan dataset Titanic. Kasus yang ditemukan menunjukkan bahwa evaluasi model tidak cukup dilakukan hanya berdasarkan nilai akurasi, tetapi perlu didukung oleh analisis Confusion Matrix, Feature Importance, dan korelasi antarfitur. Pendekatan ini dapat membantu mengidentifikasi potensi bias maupun kebocoran data yang sering tidak disadari pada tahap awal penelitian.

Secara praktis, hasil penelitian ini diharapkan dapat menjadi referensi dalam pengembangan sistem prediksi berbasis machine learning pada berbagai bidang, seperti kesehatan, pendidikan, dan keuangan. Dengan menerapkan proses validasi data yang lebih ketat, model yang dibangun akan memiliki kemampuan generalisasi yang lebih baik sehingga dapat digunakan sebagai dasar pengambilan keputusan secara lebih akurat dan dapat dipercaya.

IV. KESIMPULAN

Penelitian ini bertujuan untuk membandingkan kinerja tiga algoritma klasifikasi biner, yaitu Logistic Regression, Decision Tree, dan Random Forest, dalam memprediksi status keselamatan penumpang menggunakan dataset Titanic. Proses penelitian dilakukan melalui tahapan praproses data, pembagian data latih dan data uji menggunakan metode Stratified Train-Test Split, pelatihan model, serta evaluasi menggunakan metrik accuracy, precision, recall, F1-score, confusion matrix, analisis korelasi, dan feature importance.

Hasil penelitian menunjukkan bahwa ketiga algoritma menghasilkan nilai akurasi, presisi, recall, dan F1-score sebesar 1,00 pada data uji. Meskipun hasil tersebut menunjukkan performa yang sangat tinggi, analisis lebih lanjut membuktikan bahwa nilai evaluasi yang sempurna tidak sepenuhnya mencerminkan kemampuan model dalam melakukan generalisasi terhadap data baru. Analisis korelasi dan feature importance menunjukkan bahwa atribut Sex memiliki pengaruh yang sangat dominan terhadap variabel target Survived, dengan nilai feature importance sebesar 0,83. Hubungan yang sangat kuat antara atribut tersebut dengan label target mengindikasikan adanya data leakage, sehingga model memperoleh informasi yang secara langsung berkaitan dengan variabel yang diprediksi.

Berdasarkan temuan tersebut, penelitian ini menunjukkan bahwa hasil evaluasi model machine learning tidak dapat dinilai hanya berdasarkan tingginya nilai akurasi, tetapi juga harus disertai dengan analisis terhadap kualitas dataset, hubungan antara fitur dan target, serta

kemungkinan terjadinya data leakage. Dengan demikian, penelitian ini tidak hanya memberikan perbandingan performa tiga algoritma klasifikasi, tetapi juga menegaskan pentingnya proses validasi data sebelum membangun dan mengevaluasi model machine learning. Temuan ini diharapkan dapat menjadi referensi bagi penelitian selanjutnya dalam menerapkan proses evaluasi model yang lebih komprehensif sehingga hasil yang diperoleh benar-benar mencerminkan kemampuan model dalam melakukan prediksi pada data baru.

V. REFERENSI

- [1] Penulis Jutif, "Preventing Data Leakage in Classification via Integrated Machine Learning Pipeline: Studi Kasus Dataset Titanic," J. Inform. dan Tek. Inf. (JUTIF), vol. 7, no. 1, pp. 391–410, Feb. 2026.
- [2] Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Data Imbalanced (Studi Kasus: Klasifikasi Rumah Tangga Miskin di Kabupaten Karangasem, Bali Tahun 2017), ResearchGate, 2019.
- [3] Perbandingan Kinerja Algoritma Decision Tree dan Random Forest untuk Klasifikasi Nutrisi pada Makanan Cepat Saji, ResearchGate, 2025.
- [4] "Random Forest vs Decision Tree," School of Information Systems, Universitas Bina Nusantara, 2024. [Online]. Tersedia: <https://sis.binus.ac.id/2024/04/02/random-forest-vs-decision-tree/>
- [5] Perbandingan Algoritma Logistic Regression dan Random Forest, JUARA (Jurnal Aplikasi dan Riset Informatika), vol. 2, no. 1, Agu. 2023.
- [6] Analisis Perbandingan Decision Tree dan Random Forest pada Klasifikasi Teks Data Kesehatan, Bina Insani ICT Journal, 2023.
- [7] T. Z. Jasman et al., "Perbandingan Logistic Regression, Random Forest, dan XGBoost dalam Prediksi Penyakit Gagal Jantung," CSRID Journal, pp. 275–284, 2024.
- [8] J. Brownlee, "Data Leakage in Machine Learning," Machine Learning Mastery, 2020. [Online]. Tersedia: <https://machinelearningmastery.com/data-leakage-machine-learning/>
- [9] Riansyahputra et al., "Membangun Model Prediksi yang Robust: Penanganan Feature Leakage dalam Analisis Dampak Media Sosial terhadap Prestasi," EJECTS Journal of Computer Technology and Information Systems, 2026.
- [10] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
- [11] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
- [12] Kaggle, "Titanic - Machine Learning from Disaster: Data," Kaggle Competition Page. [Online]. Tersedia: <https://www.kaggle.com/c/titanic/data>

- [13] B. W. Nugroho dan Supriyanto, "Analisis Prediktif Berbasis Metaheuristik untuk Keselamatan Penumpang dalam Insiden Maritim: Studi Kasus Titanic," E-Link Jurnal Teknik Elektro dan Informatika, 2025.
- [14] A. Ambarwari, Q. J. Adrian, dan Y. Herdiyeni, "Analisis Pengaruh Data Scaling terhadap Performa Algoritme Machine Learning untuk Identifikasi Tanaman," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 4, no. 1, pp. 117–122, 2020.
- [15] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011.