TUGAS BESAR

Menggali Pola Keselamatan Penumpang Titanic

IF5100 – PEMROGRAMAN DATA ANALITIK

Kelompok 16:

Dama Dhananjaya Daliman / 18222047

Fitra Rachma Saphira / 23525009

Diaz Abdul Matin Annabila / 23525013

PROGRAM STUDI MAGISTER INFORMATIKA

SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA

INSTITUT TEKNOLOGI BANDUNG

2025

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

1. Penjelasan Umum Dataset

Gambar 1.1 Kapal Titanic

Kejadian tenggelamnya Titanic merupakan salah satu kecelakaan kapal paling terkenal dalam sejarah. Pada 15 April 1912, selama pelayaran perdananya, RMS Titanic yang dianggap "tidak bisa tenggelam" tenggelam setelah menabrak gunung es. Sayangnya, tidak ada cukup perahu penyelamat untuk semua orang di kapal, mengakibatkan kematian 1.502 dari 2.224 penumpang dan awak kapal. Meskipun ada unsur keberuntungan dalam kemungkinan selamatnya para penumpang, sepertinya beberapa kelompok orang lebih mungkin selamat daripada yang lain.

Alasan dipilihnya dataset ini adalah karena mudahnya interpretasi terhadap fitur yang ada dan memiliki variasi tipe data yang cocok untuk analisa data dan pengembangan model *machine learning*.

2. Deskripsi Teknis Dataset

Dataset ini diambil dari *platform* perlombaan data yang terkenal, Kaggle, tepatnya pada tautan ini. Pada laman Kaggle tersebut, dataset diberikan sudah terpisah antara data untuk pelatihan (*training*) model dan untuk pengujian (*testing*), keduanya disediakan dalam format *comma separated values* (CSV). Dataset pelatihannya sendiri terdiri dari 891 baris dan 12 kolom, dengan daftar kolom beserta deskripsinya dijelaskan pada Tabel 2.1 di bawah ini.

Tabel 2.1 Kamus penjelasan fitur dalam dataset

Nama Fitur	Definisi	Tipe Data
PassengerId	ID dari penumpang	integer
Survival	Angka yang menunjukkan selamat atau tidaknya penumpang (0 = tidak selamat, 1 = selamat)	integer
Pclass	Kelas tiket dari penumpang (1 = kelas 1, 2 = kelas 2, 3 = kelas 3)	integer

Tabel 2.1 Kamus penjelasan fitur dalam dataset (lanjutan)

Nama Fitur	**Definisi**	**Tipe Data**
Name	Nama dari penumpang	string
Sex	Jenis kelamin dari penumpang (male, female)	string {male | female}
Age	Umur dari penumpang. Umur bisa pecahan (desimal) ketika kurang dari satu (1) tahun.	float
SibSp	Banyaknya saudara (*sibling*) atau pasangan (*spouse*) yang ikut menumpangi Titanic. Saudara termasuk saudara kandung dan saudara angkat laki-laki ataupun perempuan. Pasangan termasuk suami dan istri.	integer
Parch	Banyaknya orangtua atau anak yang menumpangi Titanic. Orangtua termasuk ibu dan ayah. Anak termasuk anak kandung maupun anak angkat. Beberapa anak ada yang menumpangi Titanic hanya dengan perawatnya (*nanny*) maka memiliki Parch = 0.	integer
Ticket	Nomor tiket	string
Fare	Biaya yang dibayarkan untuk menumpangi Titanic	float
Cabin	Nomor kabin	string
Embarked	Dermaga keberangkatan (S = Southampton, C = Cherbourg, Q = Queenstown)	string {S | C | Q}

3. Proses Persiapan Data

3.1 Eksplorasi

Tahap eksplorasi dilakukan dengan mengecek tipe data (df.info), jumlah baris dan kolom (df.shape) jumlah *missing value* per kolom (df.isna().sum()) dan menggunakan statistik deskriptif (df.describe) untuk mengidentifikasi kualitas data awal sebelum pemodelan. Dari hasil eksplorasi ditemukan adanya missing values pada kolom Age (177), Cabin (687), dan Embarked(2) serta adanya perbedaan skala yang cukup besar pada fitur Fare (0-512) dan Age (0-80).

3.2 Pembersihan Kolom Tidak Relevan

Penghapusan kolom yang irrelevan dilakukan dengan menduplikasi df ke data yang terpisah (new_df) agar versi asli tetap aman, lalu menggunakan .drop(axis=1) untuk menghapus kolom PassengerID, Name, Ticket, dan Cabin. Kolom-kolom tersebut dihapus karena tidak memberi informasi yang berguna bagi model. Fitur yang tidak berhubungan dengan target atau memiliki variasi teks sangat tinggi (seperti PassengerId, Name, dan Ticket) cenderung menjadi noise dan memicu *overfitting*. Kolom dengan missing value ekstrem seperti Cabin (memiliki nilai kosong 687 dari 891 baris) juga dihapus untuk menghindari bias imputasi.

3.3 Penanganan *Missing Values*

Missing value dataset berada di fitur age dengan total 177 baris.

```
1  new_df.isna().sum()
```

```
Survived      0
Pclass        0
Sex           0
Age         177
SibSp         0
Parch         0
Fare          0
Embarked      2
dtype: int64
```

Gambar 3.1. Missing Value

Untuk menangani data missing tersebut, kami memiliki asumsi jika Pclass yang lebih tinggi pastinya di huni oleh umur yang lebih tua. Karena asumsinya lebih tua = lebih kaya. Selain itu, asumsinya adalah akan ada perbedaan rata rata umur untuk masing masing gender nya karena jika pasangan naik kapal titanic otomatis kelas nya akan sama tapi umur kedua orang tersebut kemungkinan besar berbeda. Dan Ketika umurnya berbeda, kemungkinan besar umur wanita akan lebih muda

dari laki lakinya karena memang naturalnya seperti itu. Untuk pembuktian nilai rata rata nya bisa dilihat pada gambar 3.2.

```
median_age = df.groupby(['Sex', 'Pclass'])['Age'].median()
median_age
```

```
Sex     Pclass
female  1         35.0
        2         28.0
        3         21.5
male    1         40.0
        2         30.0
        3         25.0
Name: Age, dtype: float64
```

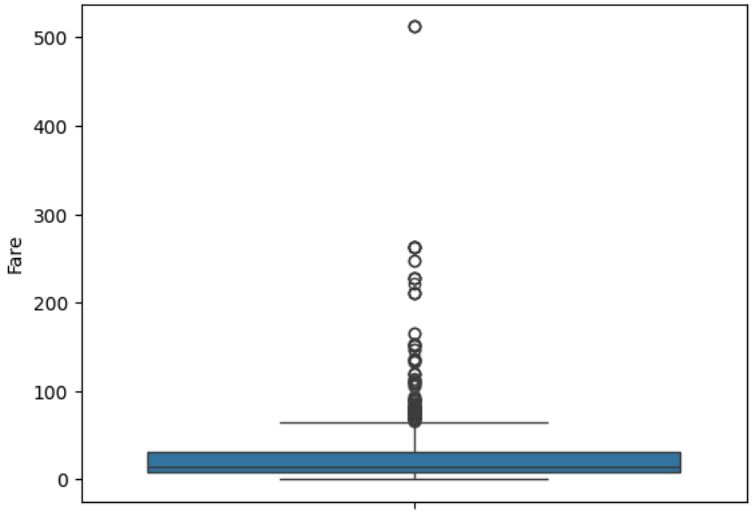
Gambar 3.2 Median age berdasarkan sex dan Pclass

3.4 Validasi Data

Tahap ini dilakukan dengan menggunakan assert Python dan logika boolean Pandas untuk mengecek validitas logika data sebelum masuk ke model. Kode memeriksa bahwa Age dan Fare tidak bernilai negatif serta memastikan Pclass hanya berisi 1, 2, atau 3 melalui .isin(). Validasi ini dilakukan untuk mencegah data tidak logis yang biasanya menandakan kesalahan preprocessing dan dapat merusak kinerja model.

3.5 Penanganan *Outliers*

Salah satu fitur yang ditemukan memiliki *outliers* (pencilan) ekstrem adalah fitur "Fare", yang distribusi nilainya terlihat pada Gambar 3.1 berikut ini.



Gambar 3.3. *Boxplot* dari kolom "Fare" yang menunjukkan adanya pencilan

Pencilan harganya terlihat ada yang sangat jauh, tetapi ini masih masuk akal karena di Titanic mungkin ada sebagian kecil orang yang membayar sangat mahal untuk kemewahan-kemewahan tertentu. Berlandaskan hal tersebut, kurang bijak kalau kita langsung membuang pencilan yang masih *feasible*. Cara lain untuk menangani pencilan semacam ini adalah dengan teknik *capping* atau *transformation*. Pada penanganan kali ini, diterapkan keduanya, tentu ini akan menghasilkan multikolinearitas, tetapi ini akan ditangani di tahap berikutnya.

3.6 Feature Engineering

Feature engineering yang kami lakukan adalah menambah fitur FamilySize dan IsAlone sebagai komponen informasi baru yang tidak tersedia secara eksplisit pada dataset awal. Kedua fitur ini berasal dari penggabungan informasi pada variabel SibSp dan Parch, yang merepresentasikan jumlah anggota keluarga yang pergi bersama penumpang.

Berdasarkan analisis pada dataset Titanic, peluang selamat penumpang dipengaruhi oleh apakah mereka bepergian sendiri atau bersama keluarga dengan ukuran kelompok tertentu. Penumpang dengan ukuran keluarga kecil cenderung memiliki tingkat keselamatan yang lebih tinggi, sementara mereka yang bepergian sendirian atau dalam keluarga besar memiliki tingkat keselamatan lebih rendah. Oleh karena itu, penambahan fitur FamilySize dan IsAlone dapat meningkatkan kemampuan model machine learning dalam mempelajari pola keselamatan penumpang.

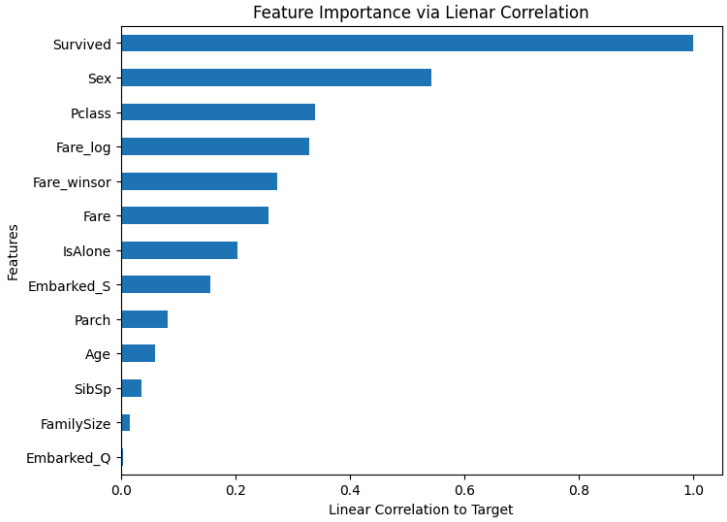
3.7 *Encoding* (Pengkodean)

Pada tahap ini, kolom bertipe non-numerik yang tersisa hanyalah kolom “Sex” dan “Embarked”. Kolom “Sex” hanya memiliki dua nilai valid yaitu “male” dan “female” yang dijelaskan pada kamus fitur, berlandaskan hal tersebut maka kolom ini bisa kita *encoding* dengan cara mapping sederhana menggunakan *dictionary*. Sedangkan kolom “Embarked” memiliki 3 nilai unik, ketiga nilai ini tidak bersifat ordinal sehingga tidak bisa menggunakan *label encoding*. Salah satu teknik pengkodean yang bisa dilakukan adalah *one-hot encoding* (OHE), pada *notebook* yang dibuat, OHE diterapkan dengan menggunakan metode “get_dummies()” milik pustaka pandas.

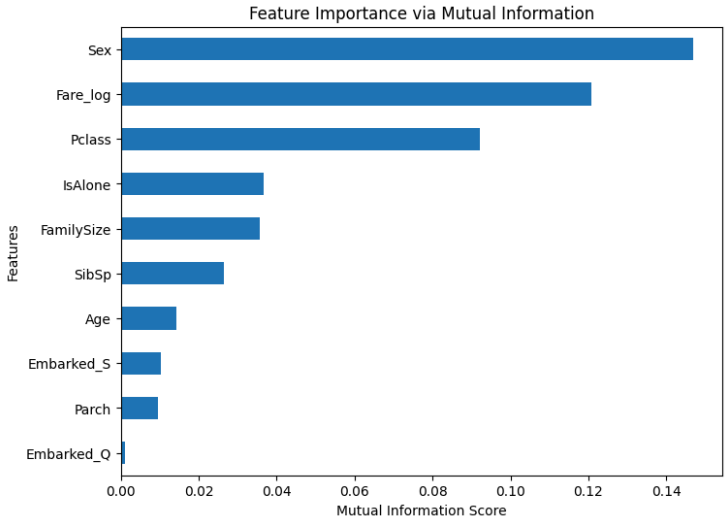
3.8 Pemilihan Fitur

Tahap pemilihan fitur dilakukan berlandaskan dua statistik dasar tentang keterhubungan kolom numerik, yaitu korelasi linear dan informasi mutual. Pada

Gambar 3.3 dan Gambar 3.4 berturut-turut menunjukkan grafik batang korelasi dan informasi mutual dari setiap fitur terhadap target.



Gambar 3.4 Grafik batang korelasi linear setiap fitur dengan target



Gambar 3.5 Grafik batang informasi mutual setiap fitur dengan target

Dari kedua nilai statistik di atas, akhirnya dipilih masing-masing 5 tertinggi. Kemudian disatukan menjadi sekumpulan fitur yang disimpan dalam variabel “final_feats” yang isinya adalah ['Survived', 'Sex', 'Pclass', 'Fare_log', 'IsAlone', 'Embarked_S'].

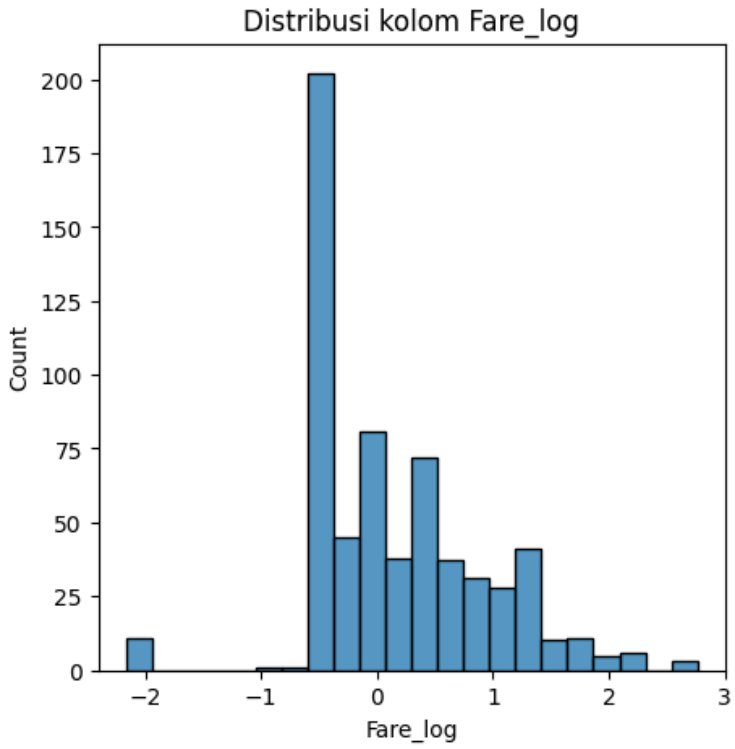
3.9 Penskalaan

Tahap pra-pemrosesan yang terakhir adalah penskalaan, penskalaan dilakukan pada semua kolom numerik (yang pada tahap ini seharusnya semua sudah numerik) menggunakan RobustScaler dari Scikit-Learn. Pemilihan RobustScaler

dilandaskan pengetahuan bahwa salah satu fitur (“Fare”) memiliki pencilan, sehingga penskalaan dengan RobustScaler akan mengurangi efek dari pencilan ini pada proses pelatihan model.

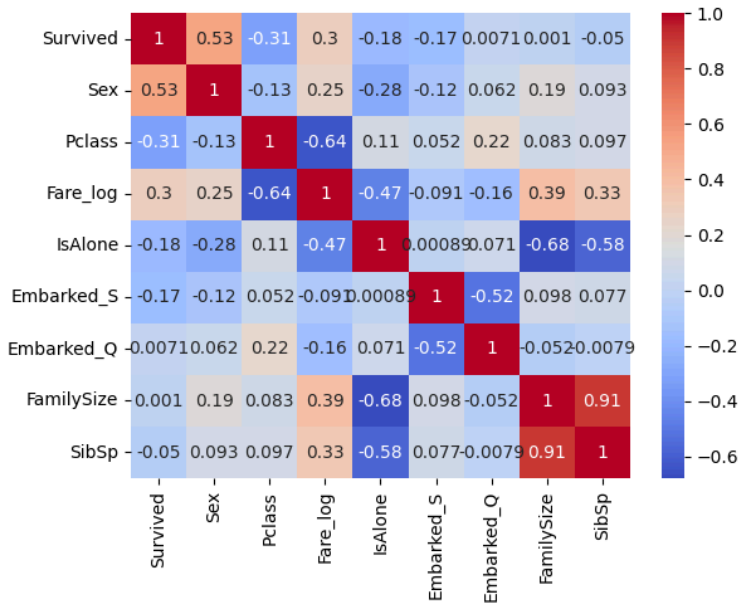
4. Proses EDA

Bagian ini akan melakukan eksplorasi lebih lanjut terkait dataset yang telah melalui pra-pemrosesan yang dijelaskan sebelumnya. Setelah pra-pemrosesan, dataset yang digunakan untuk pelatihan jadinya berdimensi 623 baris x 9 kolom (termasuk kolom target). Kolom yang tersisa adalah “Survived”, “Sex”, “Pclass”, “Fare_log”, “IsAlone”, “Embarked_S”, “Embarked_Q”, “FamilySize”, “SibSp”. Distribusi dari kolom-kolom ini kurang lebih mirip dengan distribusi awalnya, karena mayoritas adalah fitur diskrit yang sebenarnya menunjukan kategori. Distribusi yang signifikan untuk ditunjukkan adalah distribusi kolom “Fare_log”, distribusinya ditunjukkan pada Gambar 4.1 di bawah ini.



Gambar 4.1 Histogram distribusi “Fare_log”

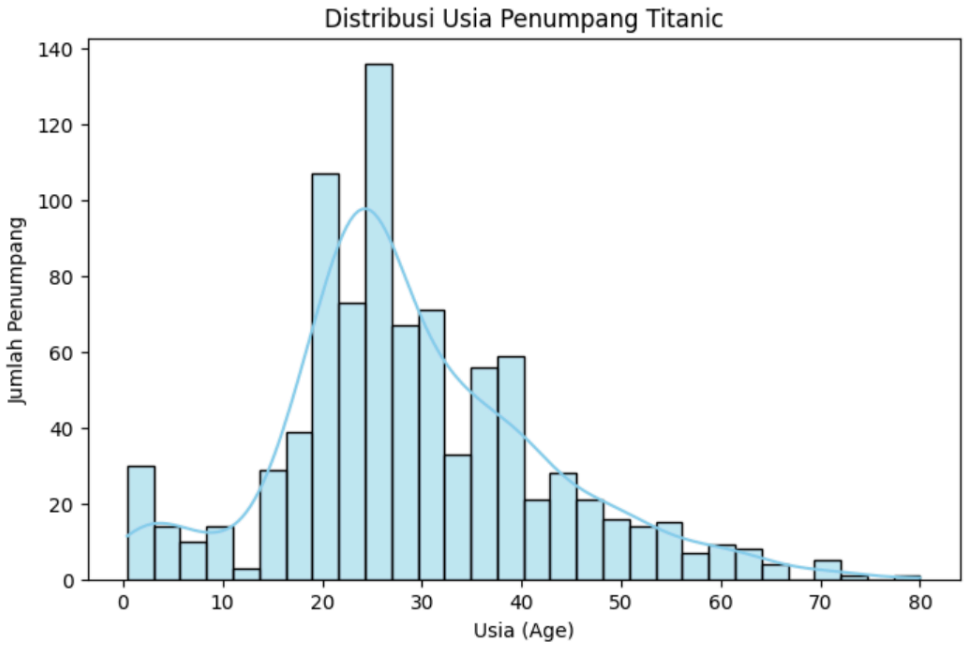
Selain distribusi, pada tahap EDA ini juga dilihat korelasi dari semua fitur yang sekarang ada. Korelasinya divisualisasikan dalam suatu *heatmap* seperti pada Gambar 4.2.



Gambar 4.2 *Heatmap* korelasi linear semua fitur

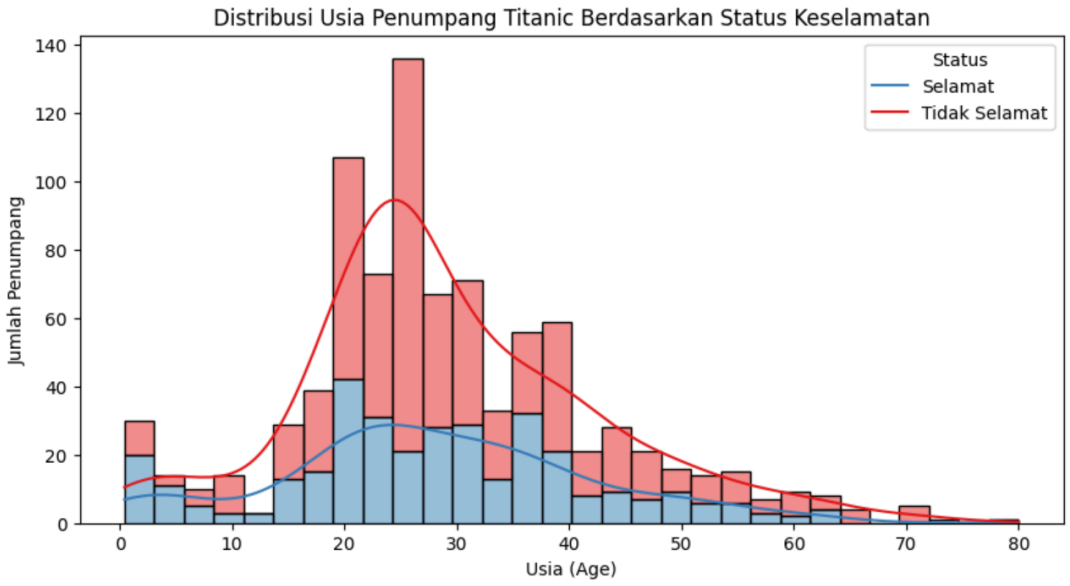
5. Visualisasi dan Penjelasannya

5.1 Visualisasi 1



Gambar 5.1 Visualisasi Distribusi Usia penumpang titanic

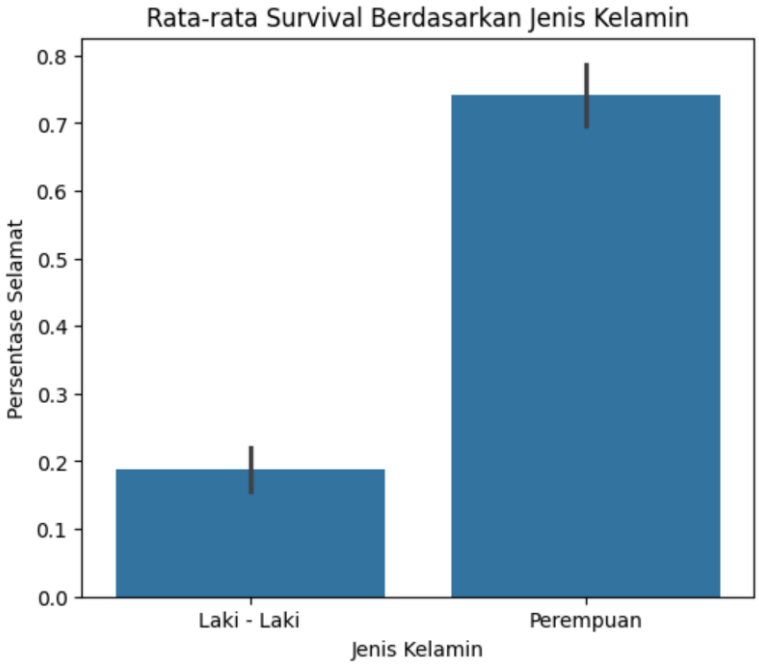
Visualisasi diatas digunakan untuk mengetahui distribusi usia penumpang yang menaiki kapal titanic. Selanjutnya dari visualisasi di atas, ditambahkan juga fitur *survive* untuk mengetahui distribusi penumpang selamat berdasarkan umur.



Gambar 5.2 Visualisasi distribusi penumpang selamat berdasarkan usia

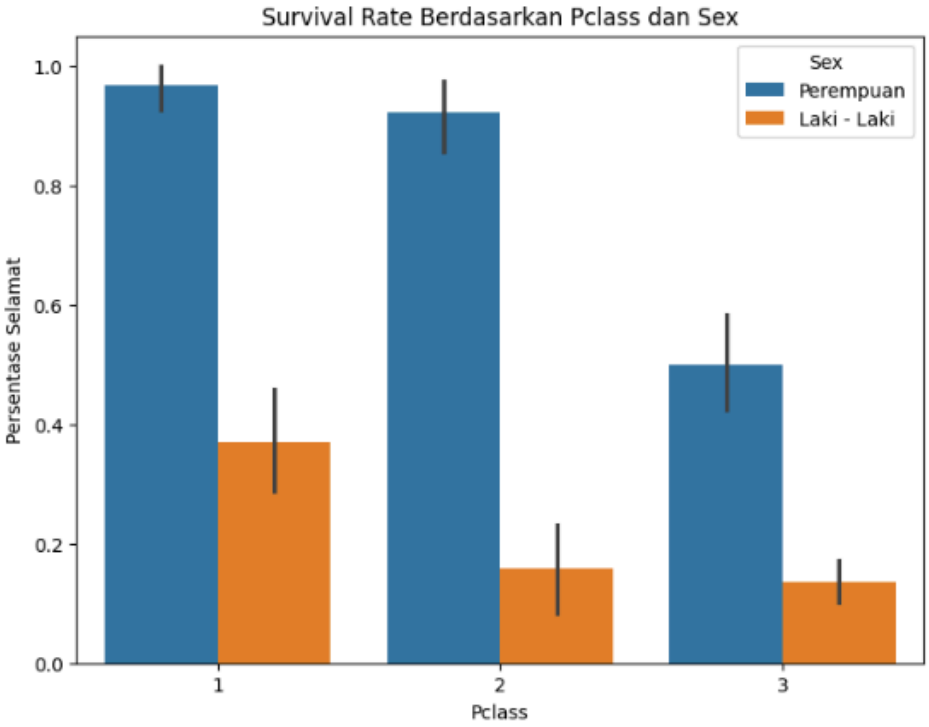
5.2 Visualisasi 2

Setelah diketahui distribusi umur dan korelasi antara umur dan survival rate, visualisasi selanjutnya adalah menampilkan survival rate berdasarkan jenis kelamin. Untuk hasilnya bisa dilihat pada Gambar 5.3.



Gambar 5.3 Visualisasi 2 Survival rate berdasarkan sex

5.3 Visualisasi 3



Gambar 5.4 Visualisasi 3 Survival Rate Berdasarkan Pclass dan Sex

Visualisasi 3 ini menggunakan bar chart dengan dua atribut (Pclass dan Sex) untuk membandingkan rata-rata tingkat keselamatan penumpang Titanic. Bar chart dipilih karena efektif menunjukkan perbedaan kategori secara jelas. Dari grafik terlihat bahwa perempuan memiliki tingkat keselamatan jauh lebih tinggi di semua kelas, terutama pada kelas 1 dan 2, sedangkan laki-laki memiliki tingkat survival rendah di seluruh kelas, khususnya kelas 3. Grafik ini menunjukkan bahwa Pclass dan Sex merupakan fitur yang sangat informatif dalam memprediksi kemungkinan seseorang selamat.

5.4 Visualisasi Tambahan

Gambar 5.5 Visualisasi Tambahan : Hubungan Age, Gender, Psclass, Family Size, dan Fare terhadap Survived

Visualisasi scatter plot animasi ini menampilkan hubungan antara Age, Fare dan Survived dengan pemisahan berdasarkan Sex dan Pclass serta ukuran bubble merepresentasikan ukuran Family Size. Grafik ini efektif menampilkan multidimensi dalam satu tampilan. Pada grafik terlihat penumpang perempuan (terutama di Pclass 1) memiliki peluang selamat lebih besar. Sebaliknya Pclass 3 didominasi dengan titik merah (tidak selamat). Family size besar umumnya muncul di Pclass 3 dan lebih banyak tidak selamat. Insight ini menguatkan fitur-fitur ini penting untuk model prediksi keselamatan.

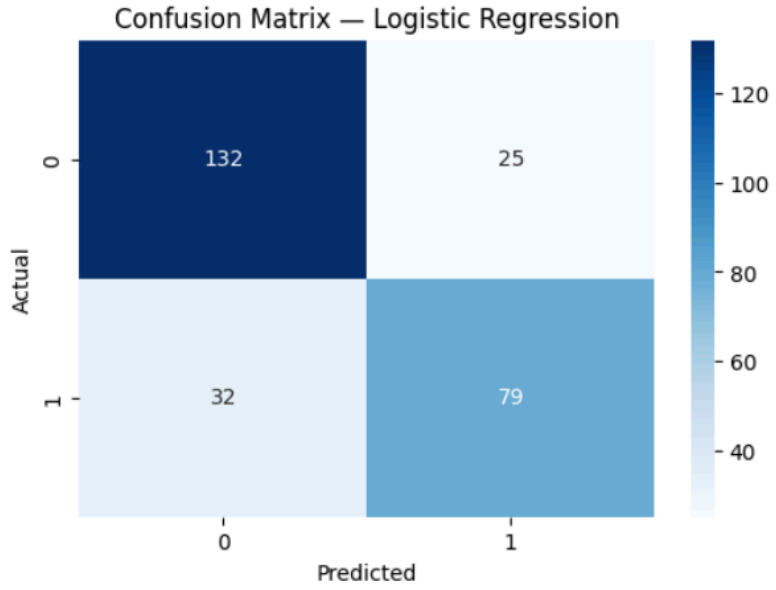
6. Model Pembelajaran Mesin dan Penjelasannya

Dalam pembuatan semua model pembelajaran mesin, fitur latih yang digunakan sama, yaitu: “Sex”, “Pclass”, “Fare_log”, “IsAlone”, “Embarked_S”, “Embarked_Q”, 'FamilySize', dan 'SibSp'. Tahap ini mengembangkan tiga model pembelajaran mesin yaitu Logistic Regression, Random Forest, dan Gradient Boosting.

6.1 Model 1: Logistic Regression

Classification Report:

	precision	recall	f1-score	support
0.0	0.80	0.84	0.82	157
1.0	0.76	0.71	0.73	111
accuracy			0.79	268
macro avg	0.78	0.78	0.78	268
weighted avg	0.79	0.79	0.79	268



Gambar 6.1 Classification Report & Confusion Matrix Model 1

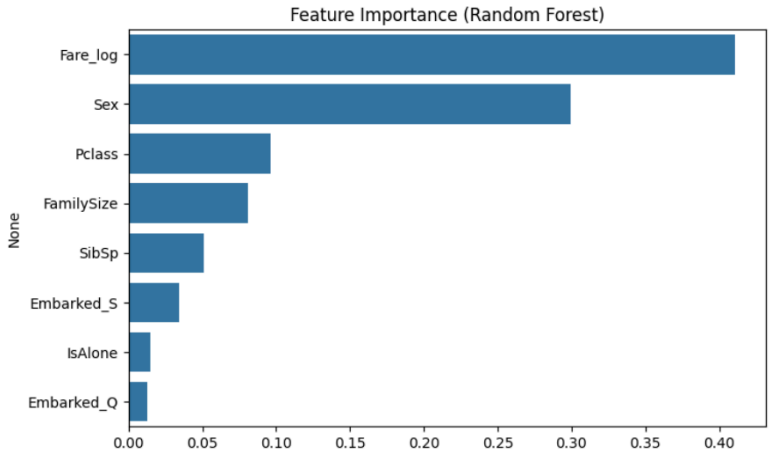
Model pertama menggunakan Logistic Regression menghasilkan akurasi 79%. Logistic Regression dipilih karena model ini sederhana, cepat, dan mudah diinterpretasikan, sehingga cocok dijadikan baseline. Dari visualisasi *Confusion Matrix*, model berhasil memprediksi dengan benar 132 penumpang yang tidak selamat (*True Negative)* dan 79 penumpang yang selamat (*True Positive*). Namun, terdapat 32 penumpang yang sebenarnya selamat namun diprediksi meninggal (*False Negative*) serta 25 penumpang yang sebenarnya meninggal tetapi diprediksi selamat (*False Positive*), sehingga kedua jenis kesalahan ini menjadi pertimbangan penting dalam evaluasi performa model.

6.2 Model 2: Random Forest Classifier

Classification Report:

	precision	recall	f1-score	support
0.0	0.79	0.82	0.81	157
1.0	0.73	0.69	0.71	111
accuracy			0.77	268
macro avg	0.76	0.76	0.76	268
weighted avg	0.77	0.77	0.77	268

Gambar 6.2.1 Classification Report Model 2



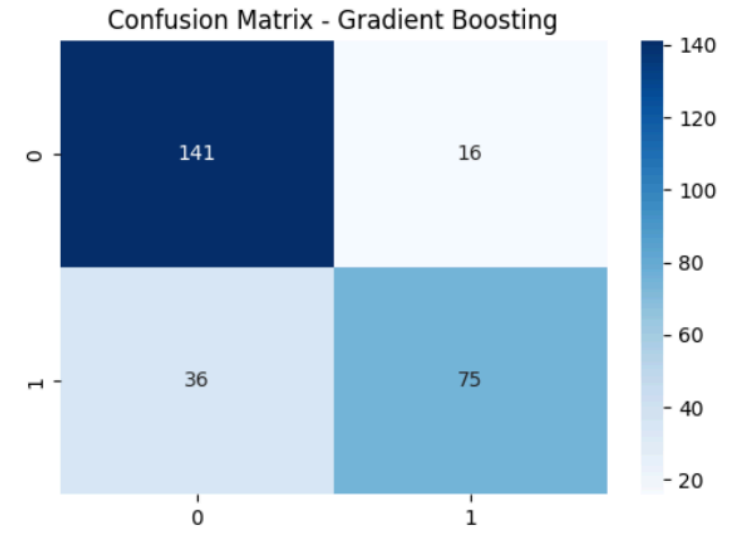
Gambar 6.2.2 *Bar Chart* Fitur Penting dari Model 2

Model kedua menggunakan Random Forest memiliki akurasi 0.77, sedikit lebih rendah, kemungkinan akibat *overfitting* atau parameter yang belum optimal. Random Forest dipilih karena mampu menangani pola non-linear, lebih tahan terhadap noise, dan memberikan informasi feature importance. Melalui feature importance, model ini menunjukkan bahwa Fare_log, Sex, dan Pclass merupakan prediktor paling dominan. Temuan ini konsisten dengan hasil EDA bahwa tarif tiket, jenis kelamin, dan kelas kabin berpengaruh besar pada keselamatan. Namun performanya pada kelas selamat masih kurang kuat (recall 0.69), sehingga model ini lebih unggul dalam interpretasi dibanding performa prediksi.

6.3 Model 3: Gradient Boosting Classifier

Classification Report:

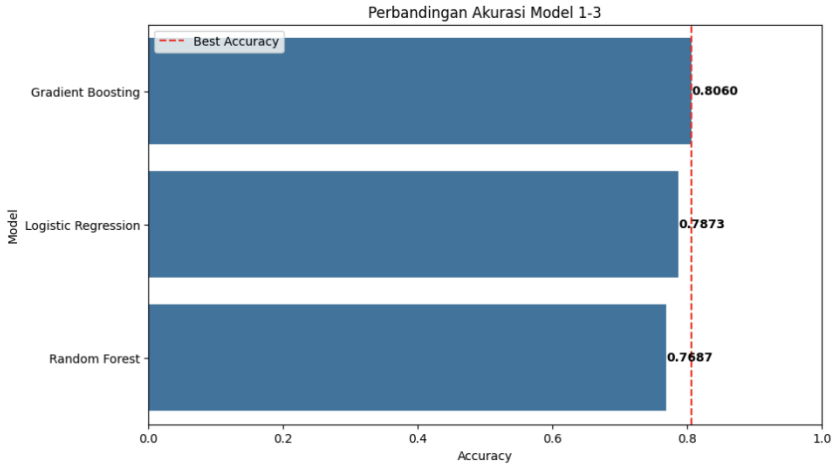
	precision	recall	f1-score	support
0.0	0.80	0.90	0.84	157
1.0	0.82	0.68	0.74	111
accuracy			0.81	268
macro avg	0.81	0.79	0.79	268
weighted avg	0.81	0.81	0.80	268



Gambar 6.3 Classification Report & Cofusion Matrix Model 3

Model ketiga adalah Gradient Boosting Classifier dengan akurasi 0.81. Gradient Boosting dipilih karena mampu mempelajari pola yang lebih kompleks melalui proses boosting bertahap, sehingga sering memberikan performa prediksi yang lebih kuat. Hal ini terlihat dari recall kelas tidak selamat yang sangat tinggi (0.90) serta peningkatan f1-score pada kelas selamat dibandingkan Random Forest. Confusion Matrix juga menunjukkan bahwa jumlah kesalahan prediksi pada kelas tidak selamat menurun cukup signifikan. Secara keseluruhan, model ini menjadi yang paling unggul karena memiliki akurasi tertinggi di antara ketiga model yang diuji.

6.4 Kesimpulan



Gambar 6.4 Perbandingan Akurasi Model 1-3

Ketiga model menunjukkan performa yang berbeda dalam memprediksi keselamatan penumpang Titanic. Logistic Regression memperoleh akurasi 0.79 dan cukup baik dalam memprediksi penumpang yang tidak selamat, meskipun masih menghasilkan false negative yang cukup tinggi. Random Forest sedikit lebih rendah dengan akurasi 0.77. Model 2 ini memberikan insight penting mengenai fitur paling berpengaruh seperti Fare_log, Sex, dan Pclass, namun performanya pada kelas selamat kurang optimal. Model terbaik adalah Gradient Boosting dengan akurasi tertinggi, yaitu 0.81. Mekanisme boosting membuatnya lebih efektif dalam menangkap pola kompleks dan menurunkan kesalahan prediksi pada kelas tidak selamat. Secara keseluruhan, Gradient Boosting adalah model paling baik dan menjadi pilihan paling tepat untuk prediksi ini.

7. Lampiran

a. Lampiran A. Pembagian Tugas

Tabel 8.1 Rincian pembagian tugas

Anggota	Pekerjaan
Dama Dhananjaya Daliman / 18222047	- Notebook: Feature selection, EDA - Laporan: Penjelasan umum dataset, deskripsi teknis dataset, pemilihan fitur, Proses EDA - PPT & Video
Fitra Rachma Saphira / 23525009	- Notebook : Eksplorasi, Validasi, Visualisasi Tambahan, Inference - Laporan: Persiapan data, Inference - PPT & Video
Diaz Abdul Matin Annabila / 23525013	- Notebook: Visualisasi, Encoding, Handle Missing Value - Laporan: Visualisasi - PPT & Ngedit Video - Berdoa semoga kerjaan lancar

b. Lampiran B. Pranala Luar

i. Tautan Github: https://github.com/RunningPie/tubesPDA_Titanic

ii. Tautan Video Penjelasan: https://www.youtube.com/watch?v=xVLVs9T-3DY