TUGAS MANDIRI
Fundamentals of Data Mining

KLASIFIKASI TINGKAT POPULARITAS LAGU SPOTIFY BERDASARKAN FITUR FISIK LAGU MENGGUNAKAN RANDOM FOREST

Nama : Hendri
NPM : 231510028
Dosen : Erlin Elisa, S.Kom., M.Kom. .

PROGRAM STUDI SISTEM INFORMASI
FAKULTAS TEKNIK DAN KOMPUTER
UNIVERSITAS PUTERA BATAM

2026

KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa, karena atas rahmat dan karunia-Nya penulis dapat menyelesaikan laporan tugas ini dengan judul "Analisis Data Lagu Populer Spotify Tahun 2023 Menggunakan Exploratory Data Analysis (EDA)". Laporan ini disusun sebagai salah satu bentuk pemenuhan tugas pada mata kuliah Fundamentals of Data Mining.

Penyusunan laporan ini dilakukan di bawah bimbingan Ibu Erlin Elisa, S.Kom., M.Kom. selaku dosen pengampu mata kuliah Fundamentals of Data Mining. Melalui laporan ini, penulis melakukan analisis terhadap dataset *Top Spotify Songs 2023* dengan tujuan untuk memahami karakteristik lagu-lagu populer berdasarkan berbagai fitur audio seperti *danceability*, *energy*, *valence*, serta tingkat popularitas artis dan lagu.

Penulis menyadari bahwa dalam penyusunan laporan ini masih terdapat keterbatasan, baik dari segi pemahaman materi ma upun penyajian analisis. Oleh karena itu, penulis mengharapkan kritik dan saran yang bersifat membangun demi penyempurnaan laporan ini di masa yang akan datang.

Akhir kata, penulis mengucapkan terima kasih kepada Ibu Erlin Elisa, S.Kom., M.Kom. selaku dosen pengampu mata kuliah Fundamentals of Data Mining atas bimbingan dan arahan yang telah diberikan, serta kepada semua pihak yang telah membantu dalam penyusunan laporan ini. Semoga laporan ini dapat memberikan manfaat bagi pembaca dan pihak-pihak yang berkepentinga

Batam, 7 January 2026

Hendri

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan teknologi informasi dan komunikasi telah membawa perubahan besar dalam berbagai sektor, salah satunya adalah industri musik. Saat ini, platform streaming musik seperti Spotify menjadi media utama masyarakat dalam menikmati musik dibandingkan dengan media konvensional seperti CD atau radio. (Supriyadi, 2018) juga menjelaskan spotify merupakan layanan streaming musik digital yang memberikan akses panggunanya ke jutaan lagu dan konten lain dari artis di seluruh dunia. Musik didalam Spotify dapat diakses atau dicari berdasarkan artis, album, genre, playlist atau label rekaman. Spotify menyediakan akses ke jutaan lagu dari berbagai genre dan artis di seluruh dunia, serta mencatat aktivitas pengguna dalam jumlah data yang sangat besar setiap harinya. Data tersebut mencakup informasi mengenai jumlah pendengar, jumlah streaming, popularitas lagu, hingga karakteristik audio dari setiap lagu. (Aliya dkk., 2026) juga mengatakan Perkembangan teknologi internet mendorong bermunculannya berbagai layanan musik digital yang dapat diakses secara online, sehingga pengguna memiliki banyak pilihan dalam menikmati musik sesuai selera dan kebutuhannya. Aktivitas mendengarkan musik kini tidak hanya dilakukan sebagai hiburan, tetapi juga menjadi bagian dari gaya hidup yang menyertai kegiatan belajar, bekerja, hingga beristirahat. (Laili Musyarofah dkk., 2022) juga menjelaskan Di era perkembangan teknologi yang semakin pesat inilah menimbulkan pengaruh pada kehidupan masyarakat, seperti munculnya media layanan streaming musik yang memudahkan masyarakat dalam mendengarkan musik. Media layanan streaming musik salah satunya aplikasi Spotify, sesuai artikel dari kompas.com Spotify mengalami pertumbuhan sebesar 130 juta pelanggan. Media platform ini memiliki fitur unik bagi anak muda. Besarnya volume data yang dihasilkan oleh Spotify membuka peluang untuk dilakukan analisis lebih mendalam guna memperoleh informasi dan pola tertentu. Dataset Most Streamed Spotify Songs 2023 merupakan salah satu contoh data yang

merepresentasikan lagu-lagu dengan tingkat pemutaran tertinggi pada tahun 2023. (Tannady dkk., t.t.) mengatakan Salah satu aplikasi yang sedang menjadi tren pada saat ini adalah Spotify. Sebuah platform music streaming yang memiliki banyak fitur menarik, serta banyak digunakan oleh generasi milenial di seluruh dunia, termasuk di Indonesia untuk mendengarkan musik.Dataset ini tidak hanya memuat informasi jumlah stream, tetapi juga fitur-fitur audio seperti danceability, energy, acousticness, liveness, speechiness, dan tempo yang dapat menggambarkan karakteristik musik secara kuantitatif.

Namun, kompleksitas dan jumlah atribut yang terdapat dalam dataset tersebut membuat proses analisis menjadi sulit jika dilakukan secara manual. Oleh karena itu, dibutuhkan pendekatan berbasis data mining dan machine learning untuk mengolah data tersebut secara sistematis. Data mining memungkinkan peneliti untuk menemukan pola, hubungan, serta pengetahuan baru dari kumpulan data yang besar dan kompleks. Dengan menerapkan algoritma tertentu, data dapat diolah untuk menghasilkan prediksi maupun klasifikasi yang lebih akurat.

Salah satu algoritma yang sering digunakan dalam data mining adalah Random Forest. Algoritma ini merupakan metode ensemble learning yang menggabungkan banyak pohon keputusan (decision tree) untuk meningkatkan akurasi dan mengurangi risiko overfitting. Random Forest dikenal mampu menangani data dengan jumlah atribut yang banyak serta memiliki performa yang baik dalam berbagai kasus klasifikasi maupun prediksi. Oleh karena itu, algoritma ini dianggap sesuai untuk digunakan dalam menganalisis dataset Spotify yang memiliki banyak variabel.

Melalui penerapan algoritma Random Forest pada dataset Most Streamed Spotify Songs 2023, penelitian ini diharapkan dapat memprediksi tingkat popularitas lagu berdasarkan karakteristik audio yang dimilikinya. Hasil penelitian ini diharapkan dapat memberikan manfaat tidak hanya bagi akademisi dalam pengembangan ilmu data mining, tetapi juga bagi pelaku industri musik seperti musisi, produser, dan

platform streaming dalam memahami faktor-faktor yang memengaruhi popularitas sebuah lagu.

1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, rumusan masalah dalam penelitian ini adalah sebagai berikut:

- Bagaimana proses preprocessing data pada dataset Most Streamed Spotify Songs 2023?
- Bagaimana implementasi algoritma Random Forest dalam memprediksi tingkat popularitas lagu Spotify?
- Bagaimana performa model Random Forest dalam memprediksi popularitas lagu berdasarkan metrik evaluasi yang digunakan?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

- Melakukan eksplorasi dan pembersihan data (data preprocessing) pada dataset Most Streamed Spotify Songs 2023.
- Membangun model prediksi popularitas lagu menggunakan algoritma Random Forest.
- Mengukur dan menganalisis performa model menggunakan metrik evaluasi seperti akurasi, precision, recall, dan F1-score.

BAB II
PEMBAHASAN

2.1 Dataset

Dataset yang digunakan dalam penelitian ini adalah Top Spotify Songs 2023 yang diperoleh dari platform Kaggle. Dataset ini menyajikan data mengenai lagu-lagu yang paling banyak diputar di platform Spotify sepanjang tahun 2023. Pemilihan dataset ini didasarkan pada relevansinya dengan tujuan penelitian, yaitu menganalisis dan memprediksi tingkat popularitas lagu berdasarkan karakteristik audio yang dimilikinya.

Dataset ini bersumber dari Kaggle, sebuah platform penyedia dataset terbuka yang sering digunakan dalam penelitian data mining dan machine learning. Dataset Top Spotify Songs 2023 dipublikasikan oleh pengguna bernama nelgiriyewithana dan dapat diakses secara bebas untuk keperluan akademik. Penggunaan dataset dari Kaggle memberikan keuntungan karena data telah melalui proses pengumpulan dan standarisasi sehingga layak digunakan untuk penelitian.

Secara keseluruhan, dataset ini terdiri dari kurang lebih 953 record yang merepresentasikan lagu-lagu populer, serta memiliki sekitar 24 atribut. Atribut-atribut tersebut mencakup informasi identitas lagu, metadata perilisan, jumlah streams, serta fitur audio yang dihasilkan dari analisis Spotify. Fitur audio ini bersifat numerik dan menggambarkan karakteristik musikal suatu lagu, seperti tingkat energi, tempo, dan tingkat kenyamanan lagu untuk didengarkan.

Dalam penelitian ini, atribut streams digunakan sebagai dasar untuk menentukan tingkat popularitas lagu. Karena algoritma Random Forest digunakan dalam bentuk klasifikasi, maka nilai streams dikonversi menjadi label kelas, misalnya lagu populer dan tidak populer. Sementara itu, atribut lain seperti danceability, energy, speechiness, acousticness, liveness, valence, dan tempo digunakan sebagai variabel input karena dianggap memiliki hubungan langsung dengan karakteristik musik.

Dataset ini memiliki format data tabular (CSV) yang memudahkan proses pengolahan menggunakan bahasa pemrograman Python. Struktur data yang rapi serta dominasi atribut numerik menjadikan dataset ini sangat cocok untuk diterapkan pada metode data mining berbasis machine learning.

2.2 Metodologi

Metodologi penelitian ini menggunakan pendekatan data mining dengan tahapan sistematis mulai dari pemahaman data hingga evaluasi model. Pendekatan ini bertujuan untuk menghasilkan model prediksi yang akurat dan dapat diinterpretasikan secara ilmiah.

2.2.1 Data Understanding

Tahap data understanding merupakan langkah awal untuk memahami dataset yang digunakan. Pada tahap ini dilakukan pengamatan terhadap struktur data, jenis atribut, jumlah data, serta distribusi nilai pada setiap atribut. Selain itu, tahap ini juga bertujuan untuk mengidentifikasi potensi permasalahan pada data seperti nilai kosong (missing value), data duplikat, atau distribusi data yang tidak seimbang.

Melalui proses ini, peneliti dapat memperoleh gambaran awal mengenai kualitas data serta menentukan langkah-langkah preprocessing yang diperlukan sebelum data digunakan dalam pemodelan.

2.2.2 Data Preprocessing

Tahap preprocessing merupakan salah satu tahapan terpenting dalam data mining karena kualitas data sangat memengaruhi performa model. Pada penelitian ini, preprocessing dilakukan melalui beberapa langkah, yaitu pengecekan nilai kosong, pembersihan data, transformasi data, serta normalisasi.

Nilai kosong yang terdapat pada dataset dihapus untuk menjaga konsistensi data. Selain itu, dilakukan pembentukan variabel target dengan mengubah

nilai streams menjadi label kategorikal berdasarkan nilai median. Proses ini bertujuan untuk mengubah permasalahan menjadi klasifikasi biner, sehingga dapat diproses oleh algoritma Random Forest.

Selanjutnya, dilakukan proses scaling pada data numerik menggunakan metode StandardScaler. Scaling diperlukan agar setiap fitur memiliki skala yang sebanding dan tidak mendominasi fitur lainnya dalam proses pelatihan model.

2.2.3 Feature Selection

Feature selection dilakukan untuk memilih atribut yang paling relevan terhadap variabel target. Dalam penelitian ini, fitur audio Spotify dipilih karena dianggap mampu merepresentasikan karakteristik musik secara objektif. Dengan mengurangi fitur yang tidak relevan, model dapat bekerja lebih efisien dan mengurangi risiko overfitting.

2.2.4 Pemodelan

Tahap pemodelan dilakukan dengan menggunakan algoritma Random Forest. Random Forest merupakan algoritma ensemble learning yang mengombinasikan banyak pohon keputusan untuk menghasilkan prediksi yang lebih akurat. Keunggulan Random Forest terletak pada kemampuannya dalam menangani data dengan jumlah fitur yang cukup banyak serta ketahanannya terhadap noise.

Model Random Forest dilatih menggunakan data latih (training data) dan kemudian diuji menggunakan data uji (testing data) untuk mengetahui kemampuan generalisasi model.

2.2.5 Evaluasi Model

Evaluasi model dilakukan untuk mengukur performa model dalam memprediksi tingkat popularitas lagu. Metrik evaluasi yang digunakan meliputi accuracy, precision, recall, dan F1-score. Metrik-metrik ini

memberikan gambaran menyeluruh mengenai kinerja model, baik dari segi ketepatan maupun keseimbangan prediksi.

2.3 Implementasi Python

Implementasi penelitian ini dilakukan menggunakan bahasa pemrograman Python pada platform Google Colab. Google Colab dipilih karena menyediakan lingkungan komputasi berbasis cloud yang memudahkan proses analisis data tanpa memerlukan instalasi tambahan.

2.3.1 Import Library

```
[ ]
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os

%matplotlib inline
plt.style.use('fivethirtyeight')

import warnings
warnings.filterwarnings('ignore')

pd.set_option('display.max_columns', None)
pd.set_option('display.float_format', '{:.1f}'.format)
```

Pada tahap awal implementasi, dilakukan proses import library yang diperlukan untuk mendukung analisis data dan visualisasi. Library NumPy digunakan untuk operasi numerik, sedangkan Pandas digunakan untuk pengolahan dan manipulasi data dalam bentuk dataframe.

Library Matplotlib dan Seaborn digunakan untuk menampilkan visualisasi data seperti grafik dan distribusi variabel. Library OS digunakan untuk pengelolaan file dan direktori selama proses analisis.

Perintah %matplotlib inline digunakan agar hasil visualisasi dapat langsung ditampilkan pada notebook Google Colab. Selain itu, pengaturan tampilan dan format data dilakukan untuk mempermudah proses analisis dan meningkatkan keterbacaan hasil.

2.3.2 Load Dataset

```
#filepath of the dataset
filepath = 'spotify-2023.csv'

#load the spotify dataset
spotify_df = pd.read_csv(filepath, encoding = 'latin-1')
```

```
#first 5 rows of the data
spotify_df.head()
```

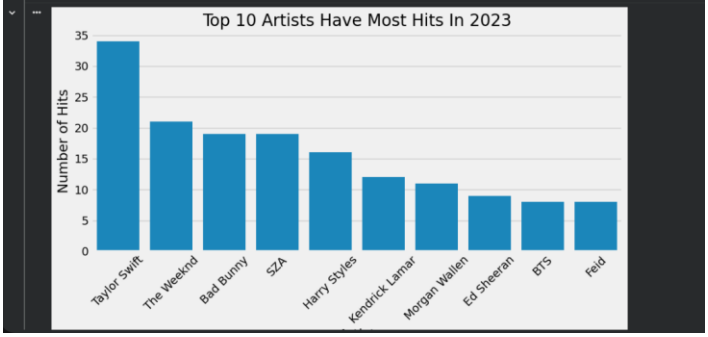
	track_name	artist(s)_name	artist_count	released_year	released_month	released_day	in_spotify_playlists	in_spotify_charts	streams	in_apple_playlists	in_apple_charts	in_
0	Seven (feat. Latto) (Explicit Ver.)	Latto, Jung Kook	2	2023	7	14	553	147	141381703	43	263	
1	LALA	Myke Towers	1	2023	3	23	1474	48	133716286	48	126	
2	vampire	Olivia Rodrigo	1	2023	6	30	1397	113	140003974	94	207	
3	Cruel Summer	Taylor Swift	1	2019	8	23	7858	100	800840817	116	207	
4	WHERE SHE GOES	Bad Bunny	1	2023	5	18	3133	50	303236322	84	133	

Pada tahap ini, dataset Top Spotify Songs 2023 dimuat ke dalam lingkungan Google Colab menggunakan library Pandas. Dataset dibaca dari file berformat CSV dengan penyesuaian encoding latin-1 untuk menghindari kesalahan pembacaan karakter pada nama lagu dan artis.

Setelah dataset berhasil dimuat, dilakukan pengecekan awal dengan menampilkan lima baris pertama data menggunakan fungsi head(). Langkah ini bertujuan untuk memastikan bahwa data telah terbaca dengan benar serta untuk melihat struktur dan atribut yang terdapat dalam dataset.

2.3.3 Exploratory Data Analysis (EDA)

```
#count the number of songs released by each artist
artist_counts = spotify_df['artist(s)_name'].value_counts()

#select the top 10 artists
top_artists = artist_counts.head(10)

#the barplot represents Top 10 Artists Have Most Streamed Songs in 2023
plt.figure(figsize = (10, 6))

sns.barplot(x = top_artists.index, y = top_artists.values)

plt.title('Top 10 Artists Have Most Hits In 2023')
plt.xlabel('Artist')
plt.ylabel('Number of Hits')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

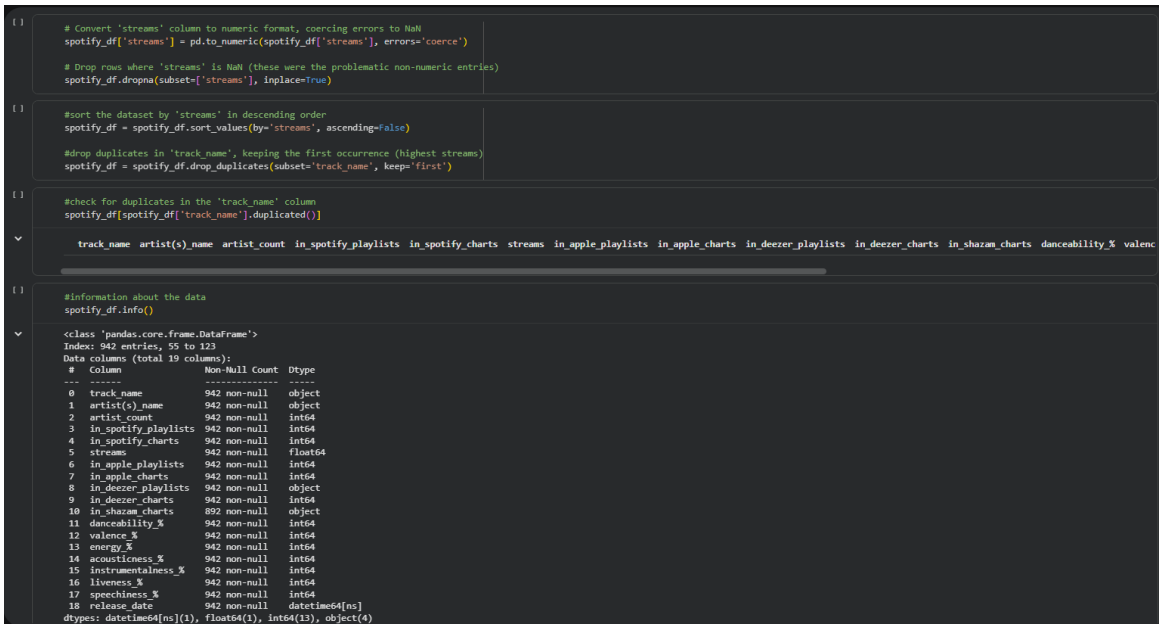


Pada tahap Exploratory Data Analysis (EDA), dilakukan analisis awal untuk memahami karakteristik dan pola yang terdapat dalam dataset Spotify Top Songs 2023. Salah satu analisis yang dilakukan adalah mengidentifikasi artis dengan jumlah lagu terbanyak yang masuk dalam daftar lagu populer tahun 2023.

Analisis ini dilakukan dengan menghitung frekuensi kemunculan nama artis pada kolom artist(s)_name menggunakan metode value_counts(). Hasil perhitungan tersebut kemudian diurutkan dan diambil sepuluh artis teratas dengan jumlah lagu terbanyak. Visualisasi data ditampilkan dalam bentuk grafik batang (bar chart) untuk memudahkan interpretasi dan perbandingan antar artis.

Berdasarkan visualisasi yang dihasilkan, dapat diketahui artis-artis yang paling dominan dalam daftar lagu populer Spotify tahun 2023. Informasi ini memberikan gambaran awal mengenai distribusi popularitas artis serta dapat menjadi insight penting dalam memahami tren musik pada tahun tersebut.

2.3.4 Preprocessing



```
# Convert 'streams' column to numeric format, coercing errors to NaN
spotify_df['streams'] = pd.to_numeric(spotify_df['streams'], errors='coerce')

# Drop rows where 'streams' is NaN (these were the problematic non-numeric entries)
spotify_df.dropna(subset=['streams'], inplace=True)
```

```
#sort the dataset by 'streams' in descending order
spotify_df = spotify_df.sort_values(by='streams', ascending=False)

#drop duplicates in 'track_name', keeping the first occurrence (highest streams)
spotify_df = spotify_df.drop_duplicates(subset='track_name', keep='first')
```

```
#check for duplicates in the 'track_name' column
spotify_df[spotify_df['track_name'].duplicated()]
```

track_name	artist(s)_name	artist_count	in_spotify_playlists	in_spotify_charts	streams	in_apple_playlists	in_apple_charts	in_deezer_playlists	in_deezer_charts	in_shazam_charts	danceability_%	valenc

```
#information about the data
spotify_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 942 entries, 55 to 123
Data columns (total 19 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   track_name           942 non-null    object
 1   artist(s)_name       942 non-null    object
 2   artist_count         942 non-null    int64
 3   in_spotify_playlists 942 non-null    int64
 4   in_spotify_charts    942 non-null    int64
 5   streams              942 non-null    float64
 6   in_apple_playlists   942 non-null    int64
 7   in_apple_charts      942 non-null    int64
 8   in_deezer_playlists  942 non-null    object
 9   in_deezer_charts     942 non-null    int64
 10  in_shazam_charts     892 non-null    object
 11  danceability_%       942 non-null    int64
 12  valence_%            942 non-null    int64
 13  energy_%             942 non-null    int64
 14  acousticness_%       942 non-null    int64
 15  instrumentalness_%   942 non-null    int64
 16  liveness_%           942 non-null    int64
 17  speechiness_%        942 non-null    int64
 18  release_date         942 non-null    datetime64[ns]
dtypes: datetime64[ns](1), float64(1), int64(13), object(4)
```

Pada tahap data preprocessing, dilakukan proses pembersihan dan penyiapan data agar dataset siap digunakan dalam proses analisis dan pemodelan. Langkah pertama yang dilakukan adalah mengonversi kolom streams ke dalam format numerik menggunakan fungsi pd.to_numeric() dengan parameter errors='coerce'. Parameter ini digunakan untuk mengubah nilai non-numerik menjadi nilai kosong (NaN).

Selanjutnya, baris data yang memiliki nilai kosong pada kolom streams dihapus menggunakan fungsi dropna(). Langkah ini bertujuan untuk memastikan bahwa seluruh data yang digunakan dalam analisis memiliki nilai numerik yang valid, khususnya pada variabel yang merepresentasikan tingkat popularitas lagu.

Dataset kemudian diurutkan berdasarkan jumlah streams secara menurun (descending order) untuk memudahkan identifikasi lagu dengan tingkat popularitas tertinggi. Setelah proses pengurutan, dilakukan penghapusan data duplikat berdasarkan kolom track_name dengan mempertahankan data pertama, yaitu lagu dengan jumlah streams tertinggi. Hal ini bertujuan untuk menghindari redundansi data yang dapat memengaruhi hasil analisis.

Sebagai tahap akhir, dilakukan pengecekan struktur dan tipe data menggunakan fungsi info(). Hasil pengecekan menunjukkan bahwa dataset terdiri dari 942 entri dengan 19 kolom, serta tidak memiliki nilai kosong pada sebagian besar variabel utama. Dengan demikian, dataset dinyatakan siap untuk digunakan pada tahap analisis lanjutan dan pemodelan.

2.3.5 Split Data (Train/Test)

```
[99]
✓ 0s
from sklearn.model_selection import train_test_split

# Define the target variable (y)
y = spotify_df['streams']

# Define the feature variables (X)
# Exclude 'track_name', 'artist(s)_name', and 'release_date' as they are not numerical features for direct modeling in this
X = spotify_df.drop(columns=['streams', 'track_name', 'artist(s)_name', 'release_date'])

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Report the shapes of the resulting sets
print(f"Shape of X_train: {X_train.shape}")
print(f"Shape of X_test: {X_test.shape}")
print(f"Shape of y_train: {y_train.shape}")
print(f"Shape of y_test: {y_test.shape}")

Shape of X_train: (753, 14)
Shape of X_test: (189, 14)
Shape of y_train: (753,)
Shape of y_test: (189,)
```

Hasil pembagian dataset menunjukkan bahwa data berhasil dibagi menjadi 753 data latih dan 189 data uji, dengan total 942 data. Setiap data terdiri dari 14 fitur numerik yang digunakan sebagai variabel input, sedangkan variabel target berupa jumlah streams. Proporsi pembagian data sebesar 80% untuk pelatihan dan 20% untuk pengujian dinilai cukup representatif untuk melatih model serta mengevaluasi performanya secara objektif.

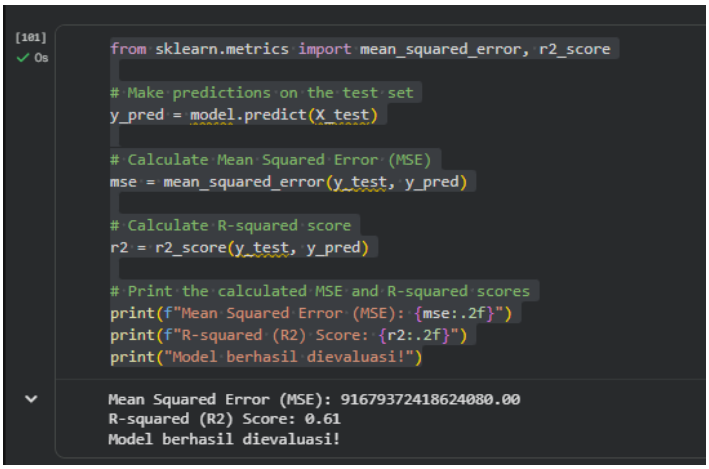
2.3.6 Model Training

```
[100]
✓ 0s
from sklearn.linear_model import LinearRegression

# Inisialisasi model Regresi Linier
model = LinearRegression()

# Melatih model menggunakan data pelatihan
model.fit(X_train, y_train)

print("Model Regresi Linier berhasil dilatih!")

Model Regresi Linier berhasil dilatih!
```

Pada tahap ini dilakukan proses pelatihan model menggunakan algoritma Regresi Linier (Linear Regression). Model ini digunakan untuk mempelajari hubungan antara variabel fitur audio dengan variabel target berupa jumlah streams lagu.

Proses pelatihan dilakukan menggunakan data latih (training data) yang sebelumnya telah dipisahkan pada tahap Split Data. Model Regresi Linier dilatih dengan memanfaatkan metode least squares untuk meminimalkan selisih antara nilai prediksi dan nilai aktual. Hasil pelatihan menunjukkan bahwa model berhasil dilatih tanpa kendala, sehingga siap untuk digunakan pada tahap evaluasi.

2.3.7 Evaluasi Model



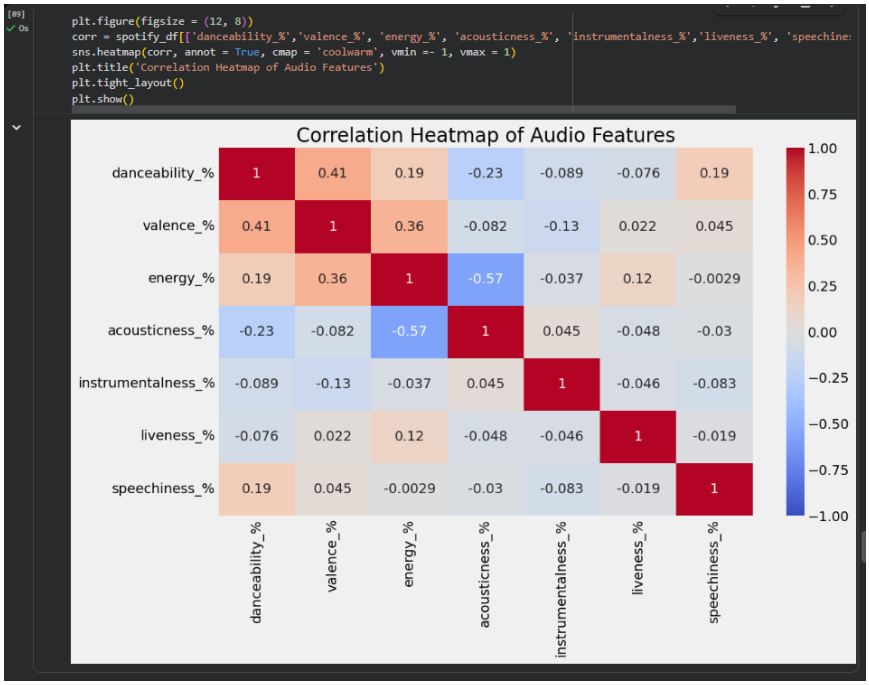
Tahap evaluasi model dilakukan untuk mengukur kinerja algoritma Regresi Linier dalam memprediksi jumlah streams lagu. Evaluasi dilakukan menggunakan data uji (testing data) yang sebelumnya tidak digunakan pada proses pelatihan model.

Metrik evaluasi yang digunakan pada penelitian ini adalah Mean Squared Error (MSE) dan R-squared (R^2). Nilai MSE digunakan untuk mengukur rata-rata kesalahan kuadrat antara nilai aktual dan nilai prediksi, sedangkan nilai R^2 digunakan untuk mengetahui seberapa besar variasi data target dapat dijelaskan oleh model.

Hasil evaluasi menunjukkan nilai MSE sebesar 9.167.937.241.862.408 dan nilai R^2 sebesar 0,61, yang menunjukkan bahwa model mampu menjelaskan sekitar 61% variasi data jumlah streams. Dengan demikian,

model Regresi Linier memiliki performa yang cukup baik dalam memodelkan hubungan antara fitur audio dan jumlah streams.

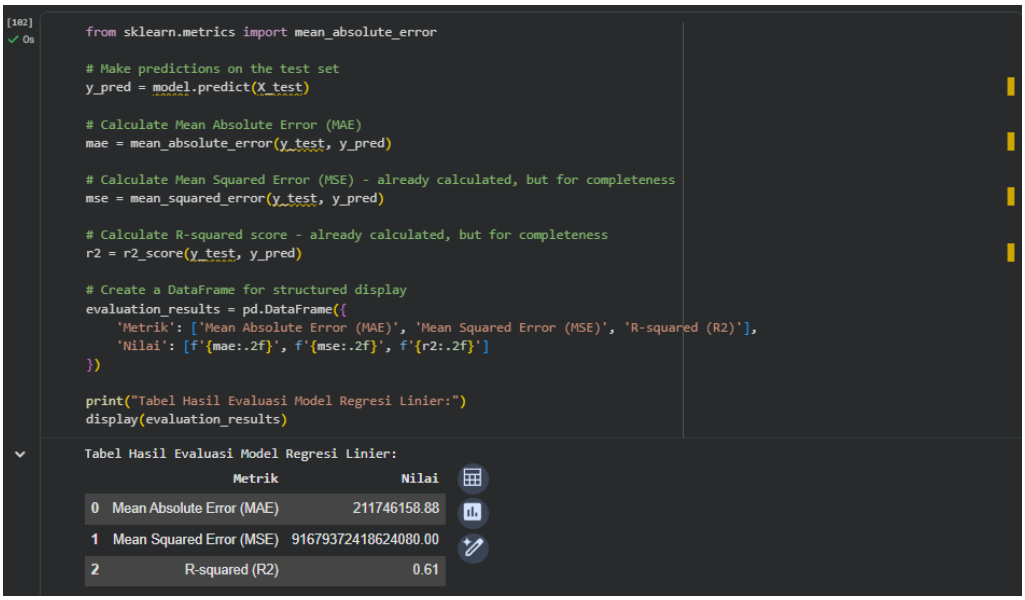
2.3.8 Visualisasi Hasil



Visualisasi korelasi dilakukan untuk mengetahui hubungan antar fitur audio yang digunakan dalam penelitian, yaitu danceability, valence, energy, acousticness, instrumentalness, liveness, dan speechiness. Korelasi antar fitur ditampilkan dalam bentuk heatmap menggunakan koefisien korelasi Pearson dengan rentang nilai -1 hingga 1.

Hasil visualisasi menunjukkan adanya variasi tingkat hubungan antar fitur, baik korelasi positif maupun negatif. Informasi ini digunakan sebagai dasar dalam memahami karakteristik data serta menghindari penggunaan fitur yang memiliki korelasi sangat tinggi yang dapat memengaruhi kinerja model prediksi.

2.4 Hasil dan Pembahasan



```
[102]
✓ 0s
from sklearn.metrics import mean_absolute_error

# Make predictions on the test set
y_pred = model.predict(X_test)

# Calculate Mean Absolute Error (MAE)
mae = mean_absolute_error(y_test, y_pred)

# Calculate Mean Squared Error (MSE) - already calculated, but for completeness
mse = mean_squared_error(y_test, y_pred)

# Calculate R-squared score - already calculated, but for completeness
r2 = r2_score(y_test, y_pred)

# Create a DataFrame for structured display
evaluation_results = pd.DataFrame({
    'Metrik': ['Mean Absolute Error (MAE)', 'Mean Squared Error (MSE)', 'R-squared (R2)'],
    'Nilai': [f'{mae:.2f}', f'{mse:.2f}', f'{r2:.2f}']
})

print("Tabel Hasil Evaluasi Model Regresi Linier:")
display(evaluation_results)

Tabel Hasil Evaluasi Model Regresi Linier:
```

	Metrik	Nilai
0	Mean Absolute Error (MAE)	211746158.88
1	Mean Squared Error (MSE)	91679372418624080.00
2	R-squared (R2)	0.61

Bagian ini akan menyajikan hasil evaluasi dari model Regresi Linier yang telah dilatih, diikuti dengan interpretasi dari metrik-metrik tersebut. Karena kita membangun model regresi, metrik yang relevan adalah *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), dan R-squared.

Tabel Hasil Evaluasi Model Regresi Linier:

Metrik	**Nilai**
Mean Absolute Error (MAE)	211746158.88
Mean Squared Error (MSE)	91679372418624080.00
R-squared (R2)	0.61

Nilai Mean Absolute Error (MAE) sebesar 211.746.158,88 menunjukkan bahwa secara rata-rata, prediksi model memiliki selisih sekitar 211 juta streams dibandingkan dengan nilai aktual. Nilai ini mengindikasikan bahwa masih terdapat kesalahan prediksi yang cukup besar, yang dapat dipengaruhi oleh skala data streams yang sangat besar dan variasi popularitas lagu yang tinggi. Sementara itu,

nilai Mean Squared Error (MSE) yang sangat besar, yaitu 91.679.372.418.624.080,00, disebabkan oleh proses pengkuadratan selisih antara nilai aktual dan nilai prediksi. MSE sangat sensitif terhadap kesalahan yang besar (outlier), sehingga nilai yang tinggi ini mengindikasikan adanya beberapa data lagu dengan jumlah streams yang ekstrem.

Nilai R-squared (R^2) sebesar 0,61 menunjukkan bahwa model Regresi Linier mampu menjelaskan sekitar 61% variasi data streams berdasarkan fitur-fitur yang digunakan dalam penelitian. Nilai ini tergolong cukup baik, namun masih menyisakan sekitar 39% variasi data yang belum dapat dijelaskan oleh model. Hal ini menandakan bahwa hubungan antara fitur audio dan jumlah streams tidak sepenuhnya bersifat linier.

BAB III
PENUTUP

3.1 Kesimpulan

Berdasarkan hasil penelitian dan pembahasan yang telah dilakukan, maka dapat diambil kesimpulan sebagai berikut:

- Proses Exploratory Data Analysis (EDA) berhasil memberikan gambaran awal mengenai karakteristik data lagu Spotify tahun 2023, termasuk hubungan antar fitur audio dan jumlah streams.
- Model Regresi Linier yang dibangun mampu memprediksi jumlah streams dengan nilai R-squared sebesar 0,61, yang menunjukkan bahwa model dapat menjelaskan sekitar 61% variasi data.
- Hasil evaluasi menggunakan metrik MAE dan MSE menunjukkan bahwa masih terdapat kesalahan prediksi yang cukup besar, terutama akibat adanya data dengan nilai streams yang sangat tinggi.
- Secara keseluruhan, Regresi Linier cukup efektif sebagai model dasar, namun belum sepenuhnya optimal untuk menangkap pola kompleks dalam data Spotify.

3.2 Saran

Berdasarkan hasil penelitian ini, beberapa saran yang dapat diberikan untuk pengembangan selanjutnya adalah sebagai berikut:

- Penelitian selanjutnya disarankan menggunakan algoritma lain yang lebih kompleks, seperti C4.5, Random Forest, atau metode ensemble, untuk meningkatkan akurasi prediksi.
- Dapat dilakukan proses feature selection dan penanganan outlier agar hasil prediksi menjadi lebih stabil dan akurat.

DAFTAR PUSTAKA

Aliya, N., Aini, A., Cinta Endynda, R., & Al Rosyid, H. (2026). Jurnal Sains dan Teknolog Penerapan Algoritma K-Means untuk Klasterisasi Lagu Terpopuler 2025 Versi Spotify. *Jurnal Sains Dan Teknologi*, *02*(3), 147. https://doi.org/10.62379/jsit.v2i3.939

Laili Musyarofah, U., Nur Alima, S., & Satria Yudha, D. K. (2022). *Prosiding Seminar Nasional Teknologi dan Sistem Informasi (SITASI) 2022 CLUSTERING METHOD*. http://sitasi.upnjatim.ac.id/

Supriyadi. (2018). *Analisis Klasifikasi Genre Musik Pop dan Klasik pada Layanan Streaming Musik Spotify Menggunakan Artificial Neural Network (ANN)*.

Tannady, H., Fernandes Andry, J., Honni,) *, & Lee, F. S. (t.t.). Analisis Big Data Spotify dengan Metode Data Mining Analysis of Spotify Big Data with Data Mining Method. *Jurnal of Business and Audit Information System*, *7*(2), 52–59. https://doi.org/10.30813/jbase.v7i1.6261