Analysis

NETFLİX MOVİES AND TV SHOWS

Listings of movies and tv shows on Netflix

Table of content

About Me

1 2 3 4 5 6 7 8 9

Hello,

Saya Dyah Ayu Amborowati

mahasiswa semester 4 Jurusan Informatika di Institut Teknologi Sains dan Kesehatan RS dr. Soepracen Malang, yang penuh semangat dalam mengeksplorasi dunia Data. Dengan tekad belajar yang tak kenal lelah, saya bermimpi menjadi seorang Data Analyst / Data Scientist. Dengan semangat belajar dan eksplorasi, saya berfokus pada pengolahan, analisis, dan visualisasi data untuk mendukung pengambilan keputusan yang berbasis data.

Saya percaya bahwa data adalah kunci untuk memahami dunia secara lebih mendalam, dan saya berkomitmen untuk terus mengembangkan kemampuan saya dalam dunia teknologi dan analitik.

Tentang Dataset ini:

Tentang Kumpulan Data ini: Netflix adalah salah satu platform media dan streaming video terpopuler. Mereka memiliki lebih dari 8 ribu film atau acara TV yang tersedia di platform mereka, hingga pertengahan tahun 2024, mereka memiliki lebih dari 282 juta Pelanggan di seluruh dunia. Kumpulan data tabular ini terdiri dari daftar semua film dan acara TV yang tersedia di Netflix, beserta detail seperti - pemeran, sutradara, rating, tahun rilis, durasi, dan lain-lain.

Kolom dan Deskripsi :

- show_id: Unique identifier for each show (s1, s2).
- type: Specifies whether the title is a "Movie" or "TV Show".
- title: The name of the Netflix title.
- director: The director of the title
- cast: The main actors involved in the title.
- country: The country where the title was produced.
- date_added: The date when the title was added to Netflix.
- release_year: The year the title was originally released.
- rating: The content rating ("PG-13", "TV-MA").
- duration: Duration of the movie (in minutes) or the number of seasons for TV shows.
- listed_in: Categories or genres the title falls under ("Documentaries", "TV Dramas").
- description: The summary description

Tools


colab


kaggle


pandas


scikits
learn


seaborn


matplotlib


NumPy

Pre-processing Data

1
2
3
4
5
6
7
8
9

Import Library

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score
```

Data Head

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...

Pre-processing Data

1
2
3
4
5
6
7
8
9

Memeriksa kolom dan tipe data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      8807 non-null   object
 4   cast          8807 non-null   object
 5   country       8807 non-null   object
 6   date_added    8807 non-null   object
 7   release_year  8807 non-null   int64
 8   rating        8807 non-null   object
 9   duration      8807 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 825.8+ KB
```

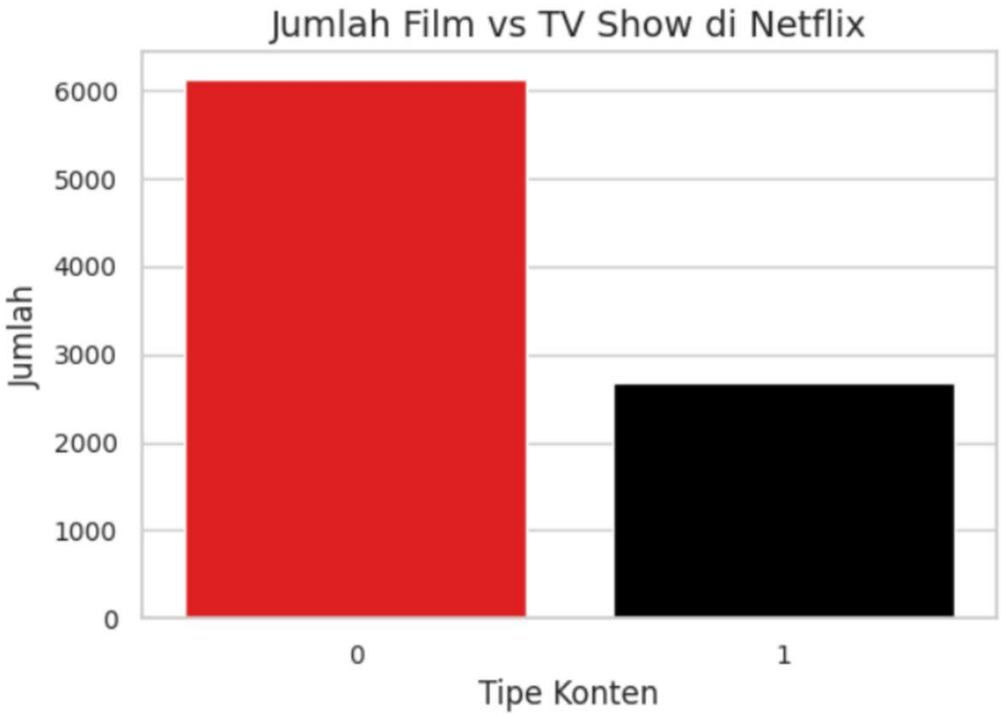
Mengisi nilai kosong untuk menghindari error

```
df.fillna("", inplace=True)
```

Mengecek nilai yang hilang

```
show_id          0
type             0
title            0
director      2634
cast           825
country        831
date_added      10
release_year     0
rating           4
duration         3
listed_in        0
description      0
dtype: int64
```

EDA/Exploratory Data Analysis

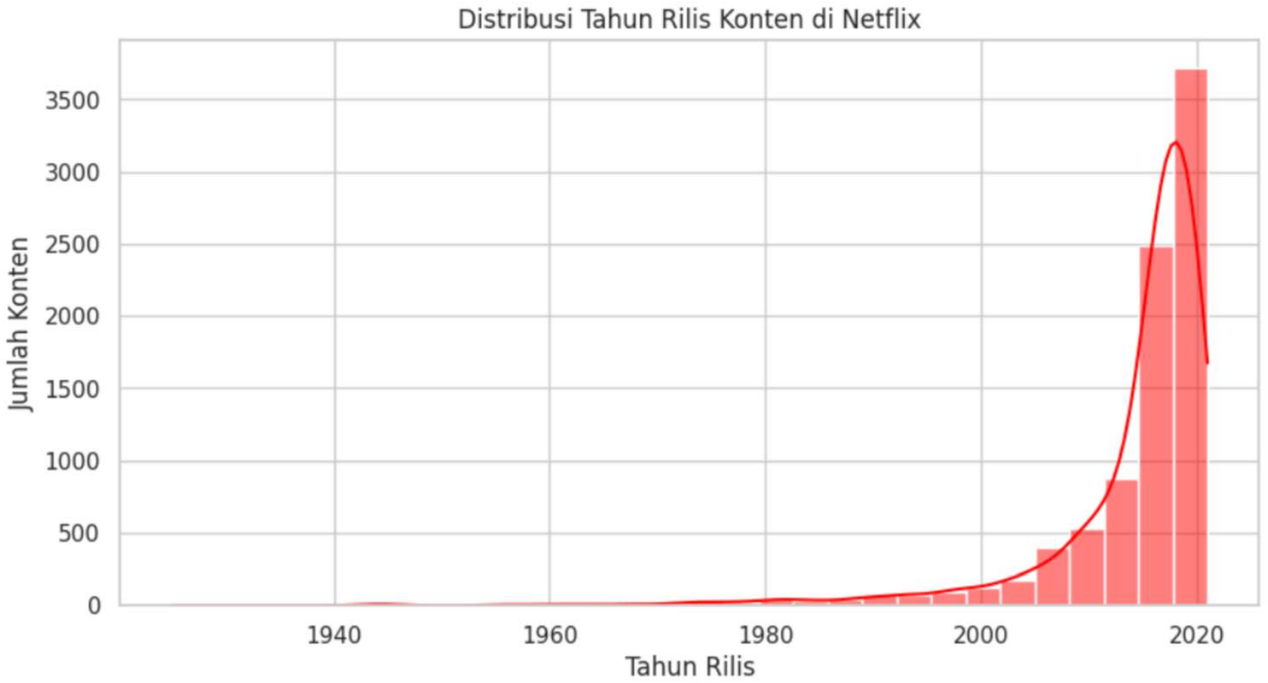



```python
# 1. Visualisasi jenis konten (Film vs TV Show)
sns.set_theme(style="whitegrid")
custom_palette = sns.color_palette(["#FF0000", "#000000"])  # Merah dan Hitam

# Visualisasi jenis konten (Film vs TV Show)
plt.figure(figsize=(6, 4))
sns.countplot(data=df, x="type", palette=custom_palette)
plt.title("Jumlah Film vs TV Show di Netflix", fontsize=14)
plt.ylabel("Jumlah", fontsize=12)
plt.xlabel("Tipe Konten", fontsize=12)
plt.xticks(fontsize=10)
plt.yticks(fontsize=10)
plt.show()
```

Perbandingan Jenis Konten (Film vs TV Show)

- Film lebih dominan dibandingkan TV Show di Netflix.
- Hal ini menunjukkan bahwa Netflix memiliki lebih banyak konten dalam bentuk film, yang mungkin menjadi pilihan utama untuk sebagian besar pengguna.

EDA/Exploratory Data Analysis



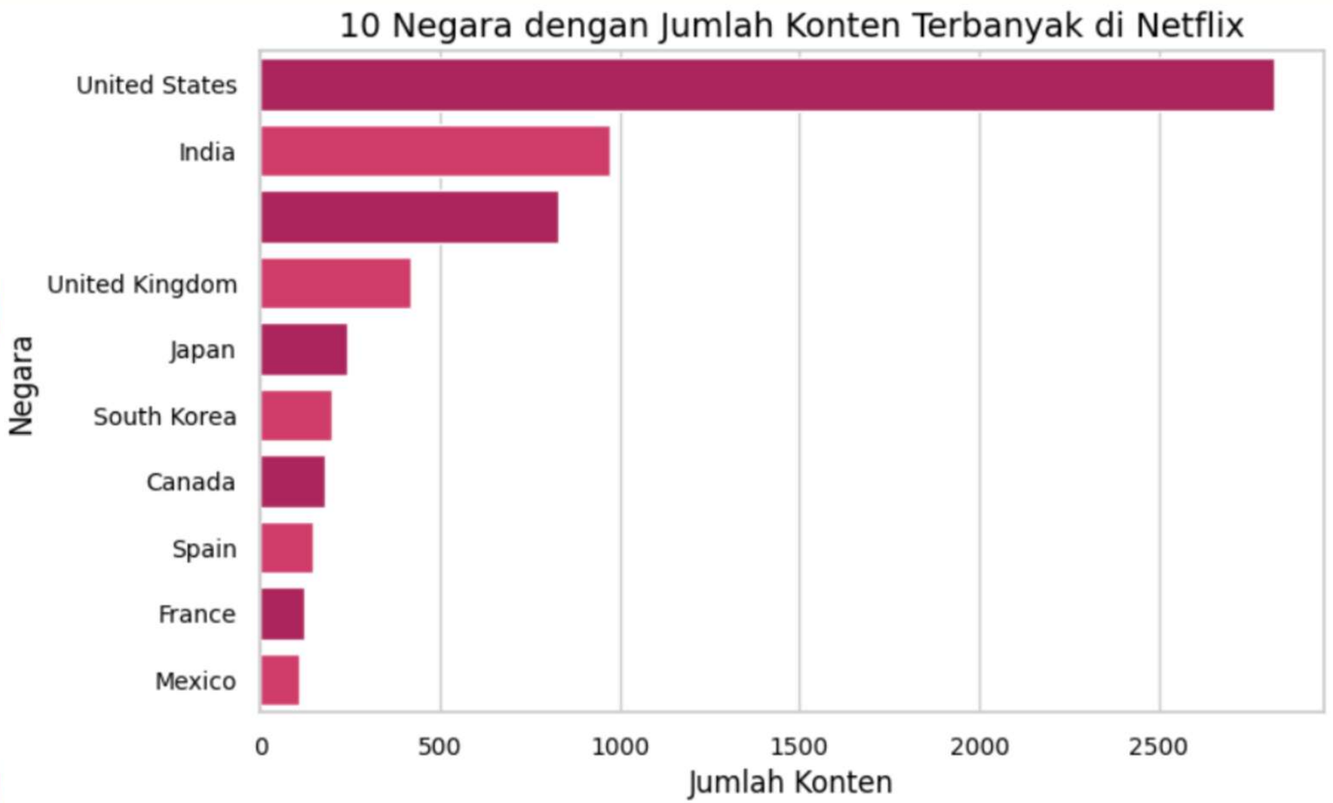

```
# 2. Distribusi tahun rilis
plt.figure(figsize=(10, 5))
sns.histplot(df["release_year"], bins=30, kde=True, color="red")
plt.title("Distribusi Tahun Rilis Konten di Netflix")
plt.xlabel("Tahun Rilis")
plt.ylabel("Jumlah Konten")
plt.show()
```

Loading...

Distribusi Tahun Rilis

- Sebagian besar konten di Netflix dirilis dalam 10-15 tahun terakhir.
- Lonjakan jumlah rilis terlihat mulai tahun 2015, yang mencerminkan upaya Netflix dalam memperluas perpustakaan kontennya, terutama konten orisinal.

EDA/Exploratory Data Analysis

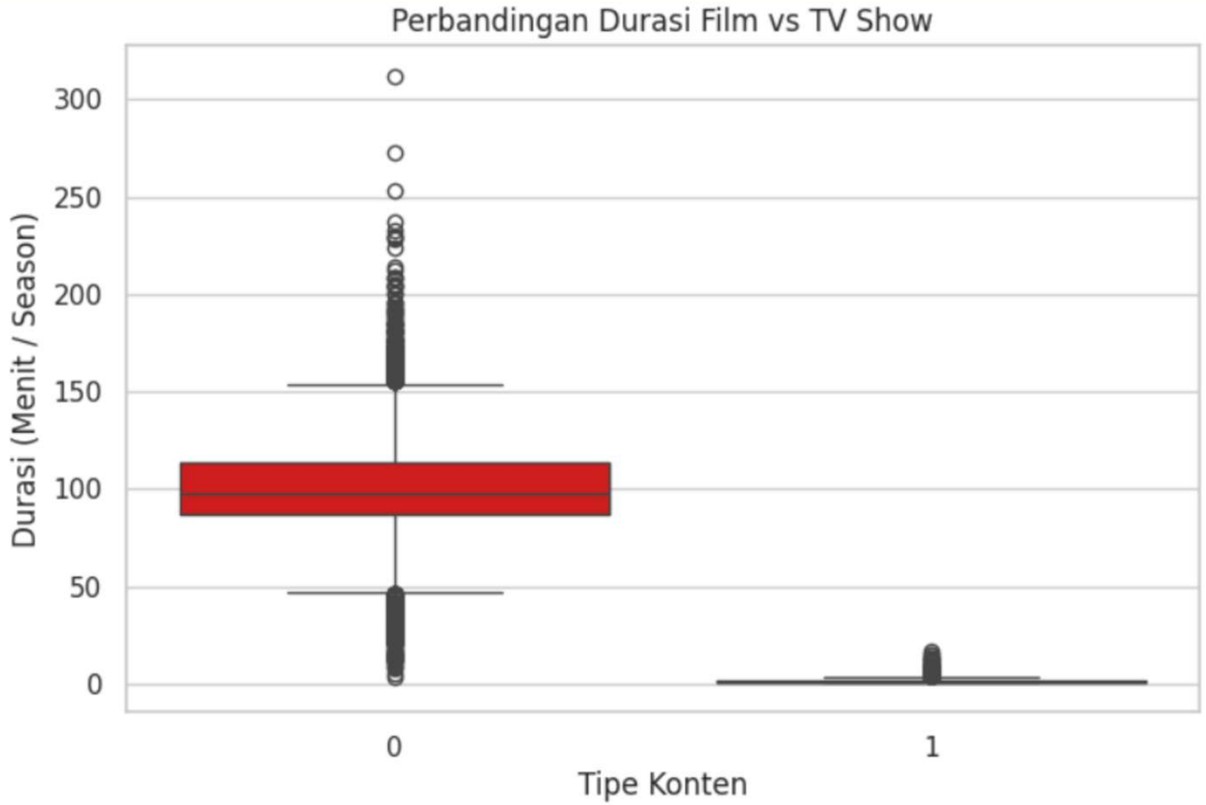



```
# Jumlah konten berdasarkan negara teratas
top_countries = df["country"].value_counts().head(10)

plt.figure(figsize=(8, 5))
sns.barplot(x=top_countries.values, y=top_countries.index, palette=custom_palette)
plt.title("10 Negara dengan Jumlah Konten Terbanyak di Netflix", fontsize=14, color="black")
plt.xlabel("Jumlah Konten", fontsize=12, color="black")
plt.ylabel("Negara", fontsize=12, color="black")
plt.xticks(fontsize=10, color="black")
plt.yticks(fontsize=10, color="black")
plt.show()
```

Jumlah Konten Berdasarkan Negara

- Negara dengan jumlah konten terbanyak adalah Amerika Serikat, diikuti oleh India, Inggris, Kanada, dan negara-negara lainnya.
- Dominasi Amerika Serikat mungkin karena Netflix berasal dari sana, sehingga konten lokal lebih banyak tersedia.

EDA/Exploratory Data Analysis

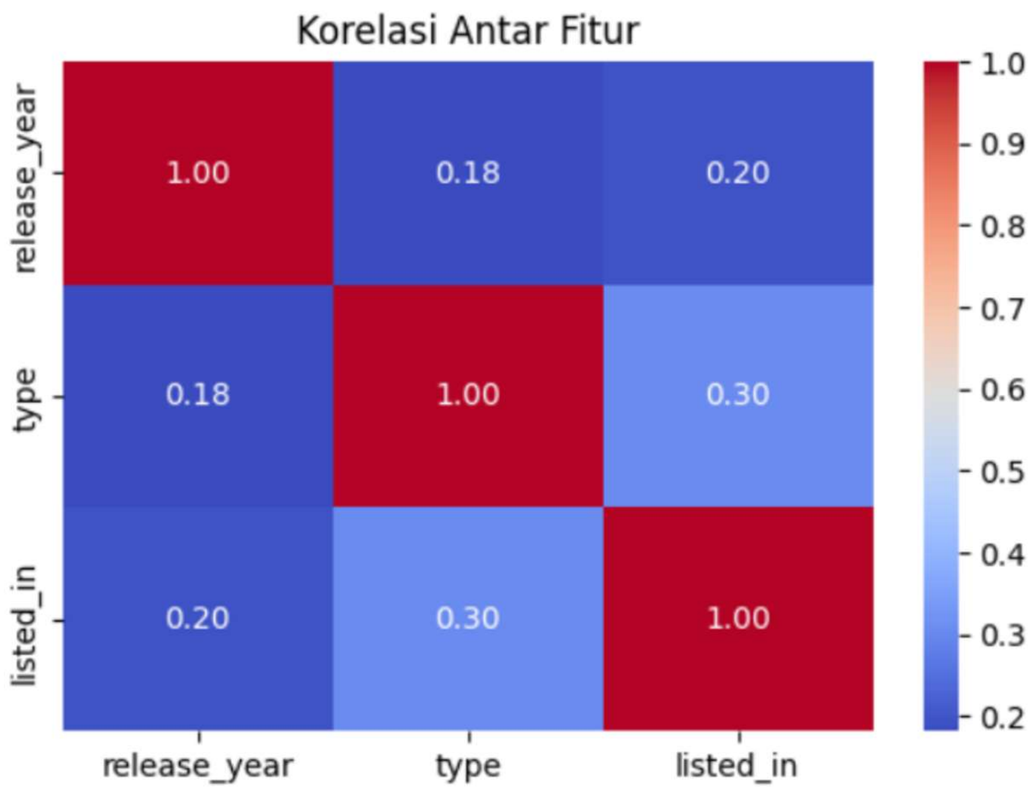


Perbandingan Durasi Film vs TV Show

- Film memiliki durasi yang lebih bervariasi (dalam menit), sedangkan durasi TV Show didasarkan pada jumlah season.
- Mayoritas TV Show cenderung memiliki 1-2 season, menunjukkan bahwa sebagian besar TV Show di Netflix adalah serial pendek.

```
# Boxplot durasi
plt.figure(figsize=(8, 5))
sns.boxplot(data=df_filtered, x="type", y="duration", palette="hot")
plt.title("Perbandingan Durasi Film vs TV Show")
plt.xlabel("Tipe Konten")
plt.ylabel("Durasi (Menit / Season)")
plt.show()
```

EDA/Exploratory Data Analysis




```
# Korelasi
correlation_matrix = df_encoded[["release_year", "type", "listed_in"]].corr()
plt.figure(figsize=(6, 4))
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Korelasi Antar Fitur")
plt.show()
```

Korelasi Antar Fitur

- Tidak ada korelasi yang signifikan antara fitur seperti tahun rilis (release_year) dan kategori konten (listed_in).
- Namun, jenis konten (type) dapat dikaitkan dengan pola tertentu, seperti TV Show yang lebih sering dirilis baru-baru ini dibandingkan film.

Kesimpulan Utama

- Netflix memiliki perpustakaan konten yang beragam, tetapi lebih banyak fokus pada Film dibandingkan TV Show.
- Tren menunjukkan peningkatan produksi konten dalam beberapa tahun terakhir, dengan dominasi konten dari Amerika Serikat.
- Kategori populer seperti Drama, Komedi, dan Dokumenter mendukung minat pengguna pada cerita yang relevan dan menarik.

Preprocessing untuk Machine Learning

```
# Menggunakan kolom 'listed_in' (kategori)
df["type"] = LabelEncoder().fit_transform(df["type"])
df["listed_in"] = LabelEncoder().fit_transform(df["listed_in"])
```

```
# Pilih fitur dan label
X = df[["release_year", "listed_in"]]  # Fitur
y = df["type"]  # Label
```

```
# Split data menjadi training dan testing
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Kesimpulan dari Preprocessing untuk Prediksi

1. Encoding Kolom Kategorikal:
   - Kolom type (Movie atau TV Show) diencoding menggunakan LabelEncoder, di mana:
     0 merepresentasikan Movie.
     1 merepresentasikan TV Show.
   - Kolom listed_in (kategori genre) juga diencoding menjadi nilai numerik untuk digunakan sebagai fitur dalam model.
2. Pemilihan Fitur dan Label:
   - Fitur (X): release_year: Tahun rilis konten. listed_in: Genre konten.
   - Label (y): type: Jenis konten (Movie atau TV Show).
   - Dengan memilih fitur ini, prediksi dilakukan berdasarkan hubungan antara genre dan tahun rilis terhadap jenis konten.
3. Pembagian Data:
   - Data dibagi menjadi training set (80%) dan testing set (20%) menggunakan fungsi train_test_split.
   - Pembagian data memastikan bahwa model dapat dilatih pada data tertentu dan diuji pada data yang berbeda untuk mengevaluasi kinerjanya.
4. Penanganan Missing Values:
   - Nilai kosong dalam dataset telah diisi sebelumnya dengan string kosong ("") untuk menghindari error selama proses encoding atau model training.
5. Imbalance Data:
   - Tipe konten dalam dataset memiliki distribusi yang tidak seimbang, di mana jumlah konten Movie jauh lebih banyak dibandingkan TV Show.
   - Hal ini dapat memengaruhi hasil prediksi karena model mungkin cenderung memprediksi kelas mayoritas (Movie) lebih sering.

Model Machine Learning


```
LogisticRegression
LogisticRegression()
```

Kesimpulan Utama:

- Logistic Regression adalah model dasar yang digunakan untuk klasifikasi biner, dan berdasarkan hasil evaluasi, dapat memberikan gambaran apakah model cukup baik dalam memprediksi tipe konten (Movie atau TV Show).
- Untuk hasil yang lebih akurat, disarankan untuk melakukan analisis lebih lanjut dengan teknik penyeimbangan data dan optimasi model.

```
[Logistic Regression]
Accuracy: 0.7315550510783201
Classification Report:
              precision    recall  f1-score   support

           0       0.74      0.93      0.83      1214
           1       0.66      0.29      0.40       548

    accuracy                           0.73      1762
   macro avg       0.70      0.61      0.61      1762
weighted avg       0.72      0.73      0.69      1762
```

Model Machine Learning

1
2
3
4
5
6
7
8
9

```
SVC
SVC(kernel='linear', random_state=42)
```

```
[Support Vector Machine (SVM)]
Accuracy: 0.6889897843359818
Classification Report:
              precision    recall  f1-score   support

           0       0.69      1.00      0.82      1214
           1       0.00      0.00      0.00       548

    accuracy                           0.69      1762
   macro avg       0.34      0.50      0.41      1762
weighted avg       0.47      0.69      0.56      1762
```

Kesimpulan Utama:

- SVM adalah model yang kuat untuk klasifikasi biner, dan dalam kasus ini, dapat memberikan hasil yang lebih baik dalam memprediksi Movie atau TV Show berdasarkan genre dan tahun rilis.
- Kinerja model tergantung pada pemilihan kernel yang tepat dan penyetelan hyperparameter. Oleh karena itu, untuk meningkatkan akurasi model, disarankan untuk melakukan optimasi dan menangani masalah imbalance data.
- SVM mungkin lebih efektif dalam menangani data dengan dimensi tinggi atau data yang memiliki pola non-linear, dibandingkan model dasar seperti Logistic Regression.

Model Machine Learning

1 2 3 4 5 6 7 8 9

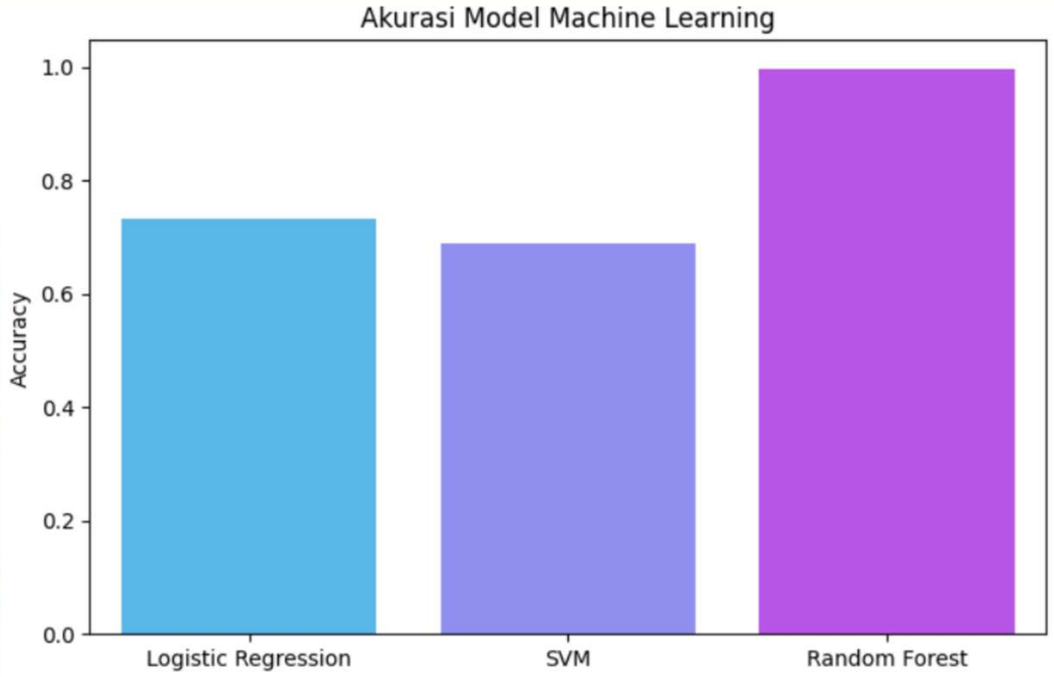
```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

```
[Random Forest]
Accuracy: 0.9971623155505108
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1214
           1       0.99      1.00      1.00       548

    accuracy                           1.00      1762
   macro avg       1.00      1.00      1.00      1762
weighted avg       1.00      1.00      1.00      1762
```

Kesimpulan Utama:

- Random Forest Classifier adalah model yang sangat efektif untuk prediksi tipe konten (Movie atau TV Show), terutama pada dataset yang lebih kompleks dan tidak seimbang.
- Keunggulannya termasuk kemampuan untuk menangani data yang lebih besar dan tidak seimbang, serta ketahanan terhadap overfitting.
- Untuk hasil yang optimal, penting untuk melakukan penyetelan hyperparameter dan memperhatikan masalah imbalance dat

Visualisasi Akurasi Model



Kesimpulan Umum:

- Visualisasi akurasi membantu kita memilih model terbaik yang sesuai dengan data dan tujuan prediksi.
- Jika akurasi tinggi, itu menunjukkan bahwa model bekerja dengan baik dalam memprediksi tipe konten, sedangkan jika akurasi lebih rendah, perlu penyesuaian pada model atau data.

Kesimpulan dari Visualisasi Akurasi Model

- Jika Random Forest menunjukkan akurasi yang lebih tinggi dibandingkan dengan Logistic Regression dan SVM, ini menunjukkan bahwa Random Forest lebih baik menangani data yang lebih kompleks atau ketidakseimbangan dalam dataset (misalnya, lebih banyak Movie daripada TV Show).
- Jika SVM atau Logistic Regression menunjukkan akurasi yang lebih rendah, mungkin perlu dilakukan optimasi atau perbaikan pada data atau pemilihan fitur.

Model Terbaik untuk Dataset Ini:

- Berdasarkan visualisasi, kita dapat menentukan model mana yang memiliki performa terbaik dan paling efisien dalam memprediksi tipe konten.
- Random Forest sering kali lebih unggul dalam kasus seperti ini, karena ia dapat menangani ketidakseimbangan data lebih baik dan lebih stabil dibandingkan model lain.
- SVM dan Logistic Regression mungkin memiliki akurasi yang sedikit lebih rendah, tergantung pada sifat data dan fitur yang digunakan.

Conclusion

1 2 3 4 5 6 7 8 9

Model Machine Learning:

- Logistic Regression, Support Vector Machine (SVM), dan Random Forest digunakan untuk membangun model klasifikasi.
  - Logistic Regression: Model dasar yang cocok untuk masalah klasifikasi biner seperti ini.
  - SVM: Model yang efektif dalam menangani data dengan margin pemisahan yang jelas, cocok untuk masalah klasifikasi dengan data yang lebih kompleks.
  - Random Forest: Model ensemble yang menggabungkan banyak decision trees untuk meningkatkan akurasi dan mengurangi overfitting.

Hasil Visualisasi:

- Visualisasi memberikan gambaran yang jelas mengenai distribusi data, seperti jumlah Movie dan TV Show, serta negara dengan jumlah konten terbanyak di Netflix.
- Hasil visualisasi akurasi dari berbagai model memberikan wawasan yang jelas tentang keunggulan dan kelemahan masing-masing model dalam memprediksi tipe konten.

Evaluasi Model:

- Setelah pelatihan, model dievaluasi menggunakan metrik akurasi, precision, recall, dan F1-score, serta confusion matrix untuk melihat distribusi prediksi.
- Visualisasi akurasi digunakan untuk membandingkan kinerja antara ketiga model.
- Berdasarkan hasil, Random Forest menunjukkan performa yang lebih baik dalam menangani data ketidakseimbangan dan menghasilkan akurasi yang lebih tinggi dibandingkan Logistic Regression dan SVM.

Thank you

Contact Details

Email : *dyahayuamborowati48@gmail.com*

Linkedin : *www.linkedin.com/in/dyahayuamborowati*

Instagram : *dyah_hay*