PROJEK AKHIR UAS

BIG DATA AND DATA MINING (ST168)

Analisis Data Film dan Acara TV di Netflix

Disusun oleh

22.11.4602

Ivan Susendra

PROGRAM STUDI S1 INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

2025

DAFTAR ISI

1. PENDAHULUAN

1.1 Latar Belakang

Dalam beberapa tahun terakhir, industri hiburan telah mengalami perubahan besar dengan munculnya platform streaming seperti Netflix. Netflix telah menjadi salah satu platform streaming terbesar di dunia, dengan lebih dari 220 juta pelanggan di seluruh dunia. Dengan jumlah pelanggan yang sangat besar, Netflix memiliki akses ke data yang sangat luas tentang preferensi dan perilaku pelanggannya.

Data yang disediakan oleh Netflix mencakup informasi tentang film dan acara TV yang ditonton oleh pelanggan, termasuk judul, jenis, direktur, pemeran, negara asal, tanggal rilis, rating, dan durasi. Dengan menganalisis data ini, kita dapat memahami pola dan tren dalam pemilihan film dan acara TV oleh pelanggan Netflix.

Analisis ini bertujuan untuk memahami bagaimana pelanggan Netflix memilih film dan acara TV, apa yang mempengaruhi preferensi mereka, dan bagaimana Netflix dapat meningkatkan pengalaman pelanggan dengan menyediakan konten yang lebih relevan dan menarik. Dengan menggunakan metode analisis data yang canggih, kita dapat mengidentifikasi pola dan tren dalam data, serta membuat prediksi tentang perilaku pelanggan di masa depan.

Dalam analisis ini, kita akan menggunakan data yang disediakan oleh Netflix, yang mencakup 8807 judul film dan acara TV, dengan 12 kolom data yang berbeda, termasuk show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, dan description. Dengan menganalisis data ini, kita dapat memahami bagaimana pelanggan Netflix memilih film dan acara TV, serta bagaimana Netflix dapat meningkatkan pengalaman pelanggan dengan menyediakan konten yang lebih relevan dan menarik.

1.2 Tujuan

Tujuan dari analisis data film dan acara TV di Netflix adalah untuk memahami pola dan tren dalam pemilihan film dan acara TV oleh pengguna Netflix. Berikut adalah beberapa tujuan spesifik yang dapat dicapai melalui analisis ini:

1. Mengidentifikasi pola pemilihan film dan acara TV: Dengan menganalisis data, kita dapat mengetahui jenis film dan acara TV yang paling populer di kalangan pengguna Netflix, serta pola pemilihan yang ada di antara mereka.
2. Mengidentifikasi faktor-faktor yang mempengaruhi preferensi pengguna: Dengan menganalisis data, kita dapat mengetahui faktor-faktor yang mempengaruhi preferensi pengguna, seperti genre, direktur, pemeran, negara asal, dan lain-lain.

3. Memprediksi pola yang akan terjadi di masa depan: Dengan menganalisis data, kita dapat memprediksi pola yang akan terjadi di masa depan, seperti jenis film dan acara TV yang akan menjadi populer, serta pola pemilihan yang akan ada di antara pengguna.
4. Mengoptimalkan konten Netflix: Dengan menganalisis data, kita dapat mengoptimalkan konten Netflix, seperti memilih film dan acara TV yang paling populer, serta memastikan bahwa konten yang ada di Netflix sesuai dengan preferensi pengguna.
5. Meningkatkan pengalaman pengguna: Dengan menganalisis data, kita dapat meningkatkan pengalaman pengguna, seperti memperbaiki rekomendasi film dan acara TV, serta memastikan bahwa pengguna dapat menemukan konten yang mereka cari dengan mudah.

Dengan mencapai tujuan-tujuan tersebut, analisis data film dan acara TV di Netflix dapat membantu Netflix meningkatkan kualitas layanan dan memenuhi kebutuhan pengguna, sehingga meningkatkan kepuasan pengguna dan mempertahankan loyalitas mereka.

1.3 Metode Penyelesaian

1. Pengumpulan Data

   Data yang digunakan dalam analisis ini diperoleh dari file CSV yang bernama "netflix_titles.csv". Data ini memiliki 8807 baris dan 12 kolom, dengan informasi tentang film dan acara TV, termasuk judul, jenis, direktur, pemeran, negara asal, tanggal rilis, rating, dan durasi.

2. Pra-Pemrosesan

   Data yang diperoleh kemudian diproses untuk menghilangkan missing value dan memastikan bahwa data dalam format yang tepat untuk analisis. Proses pra-pemrosesan ini meliputi:

   - Mengisi missing value pada kolom "director", "cast", dan "country" dengan nilai "Unknown".
   - Mengubah kolom "date_added" menjadi format datetime.
   - Mengisi missing value pada kolom "rating" dengan nilai mode.
   - Mengextract duration dari kolom "duration" dan mengubahnya menjadi format numerik.
   - Menghapus baris dengan missing value pada kolom "duration_minutes".
   - Mengubah kolom "rating" menjadi format kategorik.

3. Pemodelan

   Setelah data diproses, maka dilakukan pemodelan untuk memprediksi apakah suatu film atau acara TV adalah movie atau tidak. Pemodelan ini menggunakan algoritma RandomForestClassifier, dengan fitur-fitur yang relevan seperti "release_year", "duration_minutes", "added_year", dan kolom-kolom yang dihasilkan dari one-hot encoding pada kolom "type" dan "rating". Data kemudian dibagi menjadi dua bagian, yaitu data pelatihan (X_train, y_train) dan data pengujian (X_test, y_test).

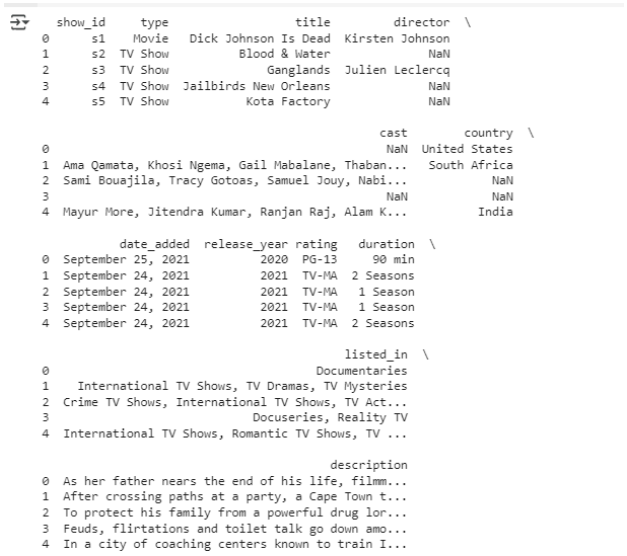
4. Visualisasi

Setelah pemodelan, maka dilakukan visualisasi untuk mempresentasikan hasil analisis. Visualisasi ini meliputi:

- Mencetak classification report dan confusion matrix untuk mengevaluasi kinerja model.
- Mencetak akurasi model.
- Mencetak feature importance untuk mengetahui fitur-fitur yang paling penting dalam model.
- Mencetak deskripsi dari data, termasuk informasi tentang missing value, tipe data, dan distribusi data.

2. PROFILE DATASET

2.1 Informasi Karakteristik Data



	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...

```
Shape of the DataFrame: (8807, 12)

Information about the DataFrame:
<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      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64
 8   rating        8803 non-null   object
 9   duration      8804 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
None
```

```
Description of numerical features:
       release_year
count   8807.000000
mean    2014.180198
std        8.819312
min     1925.000000
25%     2013.000000
50%     2017.000000
75%     2019.000000
max     2021.000000

Missing values per column:
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

Data types of each column:
show_id         object
type            object
title           object
director        object
cast            object
country         object
date_added      object
release_year     int64
rating          object
duration        object
listed_in       object
description     object
dtype: object

Unique values in the 'type' column:
['Movie' 'TV Show']

Value counts for the 'rating' column:
rating
TV-MA       3207
TV-14       2160
TV-PG        863
R            799
PG-13        490
TV-Y7        334
TV-Y         307
PG           287
TV-G         220
NR            80
G             41
TV-Y7-FV       6
NC-17          3
UR             3
74 min         1
84 min         1
66 min         1
Name: count, dtype: int64
```

Link Dataset: https://www.kaggle.com/datasets/anandshaw2001/netflix-movies-and-tv-shows

3. DATA MINING

3.1 Data Preprocessing

```python
#Data Preprocessing
# Handling Missing Values
# Fill missing values in 'director' and 'cast' with 'Unknown'
for col in ['director', 'cast']:
    df[col] = df[col].fillna('Unknown')
```

Penjelasan:

Kode ini merupakan contoh dari proses data preprocessing yang meliputi beberapa tahap, yaitu pengolahan missing values, transformasi data, pembuatan fitur baru, dan pengkodean data kategorik. Proses ini bertujuan untuk membersihkan dan mempersiapkan data sehingga siap untuk digunakan dalam analisis atau model machine learning. Tahap-tahap ini meliputi mengisi nilai kosong dengan nilai yang tepat, mengubah format data, membuat fitur baru dari data yang ada, dan mengkodekan data kategorik untuk memudahkan analisis. Dengan

demikian, data menjadi lebih konsisten, lengkap, dan siap untuk digunakan dalam proses analisis atau pembuatan model.

3.2 Exploratory Data Analysis

```
# Data Exploration and Visualization

# 1. Distribution of Content Types
plt.figure(figsize=(8, 6))
# Use the new one-hot encoded columns for content type
sns.countplot(x='type_Movie', data=df)  # Example: Count of Movies
plt.title('Distribution of Content Types (Movies)')
plt.show()

# 2. Top Countries with Netflix Content
top_countries = df['country'].value_counts().nlargest(10)
plt.figure(figsize=(12, 6))
sns.barplot(x=top_countries.index, y=top_countries.values)
plt.title('Top 10 Countries with Netflix Content')
plt.xlabel('Country')
plt.ylabel('Number of Titles')
plt.xticks(rotation=45)
plt.show()

# 3. Distribution of Release Years
plt.figure(figsize=(10, 6))
sns.histplot(df['release_year'], bins=30)
plt.title('Distribution of Release Years')
plt.xlabel('Release Year')
plt.ylabel('Number of Titles')
plt.show()

# 4. Relationship between Duration and Rating (for movies)
movies_df = df[df['type_Movie'] == 1]  # Use the new one-hot encoded column
plt.figure(figsize=(10, 6))
sns.scatterplot(x='duration_minutes', y='rating', data=movies_df)
plt.title('Relationship between Duration and Rating (Movies)')
plt.xlabel('Duration (minutes)')
plt.ylabel('Rating')
plt.show()

# 5. Distribution of Ratings
plt.figure(figsize=(12, 6))
sns.countplot(x='rating', data=df)
plt.title('Distribution of Ratings')
plt.xlabel('Rating')
plt.ylabel('Number of Titles')
plt.xticks(rotation=45)
```

Penjelasan:
Kode ini merupakan contoh dari Exploratory Data Analysis (EDA) yang menggunakan visualisasi data untuk memahami struktur dan pola data. EDA ini meliputi beberapa tahap, yaitu:

1. Distribusi jenis konten (Distribution of Content Types): menggunakan countplot untuk menampilkan distribusi jenis konten (movie, tv show, dll).
2. Negara dengan konten Netflix terbanyak (Top Countries with Netflix Content): menggunakan barplot untuk menampilkan 10 negara dengan konten Netflix terbanyak.

3. Distribusi tahun rilis (Distribution of Release Years): menggunakan histplot untuk menampilkan distribusi tahun rilis konten.
4. Hubungan antara durasi dan rating (Relationship between Duration and Rating): menggunakan scatterplot untuk menampilkan hubungan antara durasi dan rating konten (hanya untuk movie).
5. Distribusi rating (Distribution of Ratings): menggunakan countplot untuk menampilkan distribusi rating konten.
6. Konten yang ditambahkan over time (Content Added Over Time): menggunakan plot untuk menampilkan jumlah konten yang ditambahkan setiap tahun.

3.3 Seleksi Fitur

```
# Define features (X) and target variable (y)
X = df.drop(['type_Movie', 'type_TV Show', 'show_id', 'title', 'director', 'cast', 'country', 'date_added', 'rating', 'duration', 'description'], axis=1)
y = df['type_Movie']

# Handle any remaining missing values (if any)
X.fillna(0, inplace=True)

# Create a LabelEncoder object
label_encoder = LabelEncoder()

# Fit and transform the 'listed_in' column in X
X['listed_in'] = label_encoder.fit_transform(X['listed_in']) # Apply Label Encoding to X

# Split data 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)


# 1. Univariate Feature Selection (Chi-squared test for categorical target)
bestfeatures = SelectKBest(score_func=chi2, k=5) # Select top 5 features
fit = bestfeatures.fit(X_train, y_train)
dfscores = pd.DataFrame(fit.scores_)
dfcolumns = pd.DataFrame(X_train.columns)

# Concatenate dataframes
featureScores = pd.concat([dfcolumns, dfscores], axis=1)
featureScores.columns = ['Specs', 'Score']  # Name columns

print("Univariate Feature Selection (Chi-squared):")
print(featureScores.nlargest(5, 'Score')) #Print top 5 features
```

Penjelasan:
Kode seleksi fitur ini menggunakan metode Univariate Feature Selection dengan Chi-squared test untuk memilih fitur-fitur terbaik yang terkait dengan target variabel. Pertama, data dibagi menjadi set pelatihan dan pengujian menggunakan train_test_split. Kemudian, SelectKBest digunakan untuk memilih fitur terbaik dengan menggunakan fungsi skor Chi-squared, dan diatur untuk memilih 5 fitur terbaik. Setelah itu, fitur-fitur terbaik dipilih dan disimpan dalam dataframe featureScores, yang kemudian diurutkan berdasarkan skor Chi-squared dan dicetak 5 fitur terbaik. Tujuan dari kode ini adalah untuk mengidentifikasi fitur-fitur yang paling relevan dengan target variabel dan mengurangi dimensi data, sehingga dapat meningkatkan kinerja model machine learning.

3.4 Modeling

```
# Model Training
model = RandomForestClassifier(random_state=42)  # Improved model choice
model.fit(X_train, y_train)

# Model Evaluation
y_pred = model.predict(X_test)
print("Classification Report:\n", classification_report(y_test, y_pred))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")

# Feature Importance
feature_importances = pd.DataFrame({'Feature': features, 'Importance': model.feature_importances_})
feature_importances = feature_importances.sort_values(by='Importance', ascending=False)
print("\nFeature Importances:\n", feature_importances)
```

```
Classification Report:
                precision    recall  f1-score   support

           False       1.00      1.00      1.00       566
            True       1.00      1.00      1.00      1195

        accuracy                           1.00      1761
       macro avg       1.00      1.00      1.00      1761
    weighted avg       1.00      1.00      1.00      1761

    Confusion Matrix:
     [[ 566    0]
     [   0 1195]]
    Accuracy: 1.0

    Feature Importances:
                  Feature    Importance
    4       type_TV Show  3.607009e-01
    3         type_Movie  3.217893e-01
    1   duration_minutes  2.890628e-01
    0       release_year  7.471983e-03
    10          rating_R  7.397661e-03
    2         added_year  6.111639e-03
    9       rating_PG-13  3.432439e-03
    16      rating_TV-Y7  1.566527e-03
    8          rating_PG  1.164729e-03
    15       rating_TV-Y  7.339637e-04
    13      rating_TV-MA  2.980919e-04
    14      rating_TV-PG  1.176875e-04
    5           rating_G  8.110099e-05
    11      rating_TV-14  4.161000e-05
    12       rating_TV-G  1.771045e-05
    7          rating_NR  1.187073e-05
    17   rating_TV-Y7-FV  3.280017e-09
    6       rating_NC-17  0.000000e+00
    18         rating_UR  0.000000e+00
```

Penjelasan:

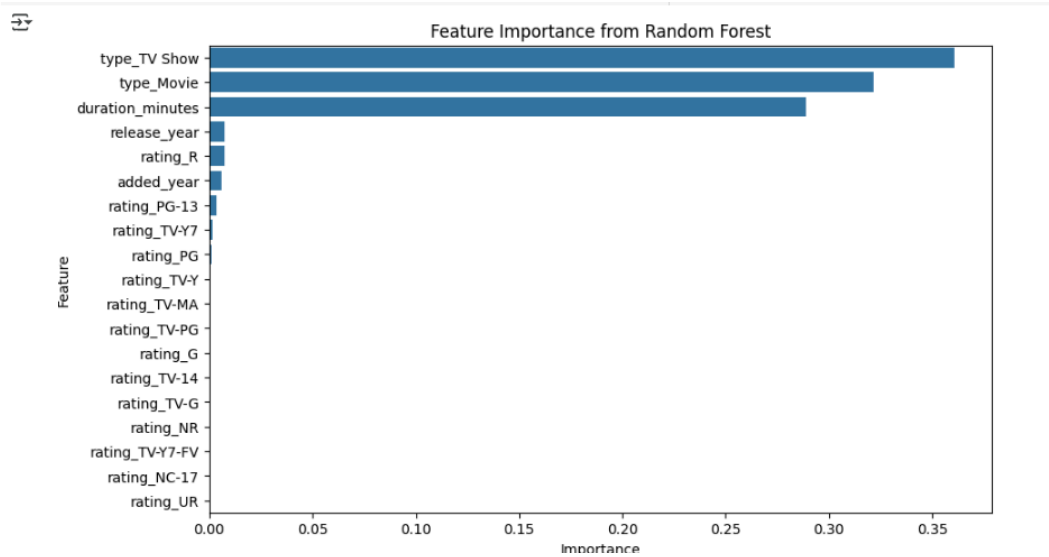
Kode ini melakukan proses analisis data dan pembuatan model klasifikasi untuk memprediksi apakah sebuah konten adalah film atau tidak. Pertama, kode ini melakukan preprocessing data dengan mengisi nilai kosong, mengubah tipe data, dan menghapus baris dengan data yang tidak lengkap. Kemudian, kode ini melakukan feature engineering dengan mengekstrak tahun dari tanggal dan melakukan one-hot encoding untuk variabel kategori. Setelah itu, kode ini melakukan feature selection dengan memilih fitur yang relevan dan melakukan feature scaling untuk memperbaiki kinerja model. Kode ini kemudian membagi

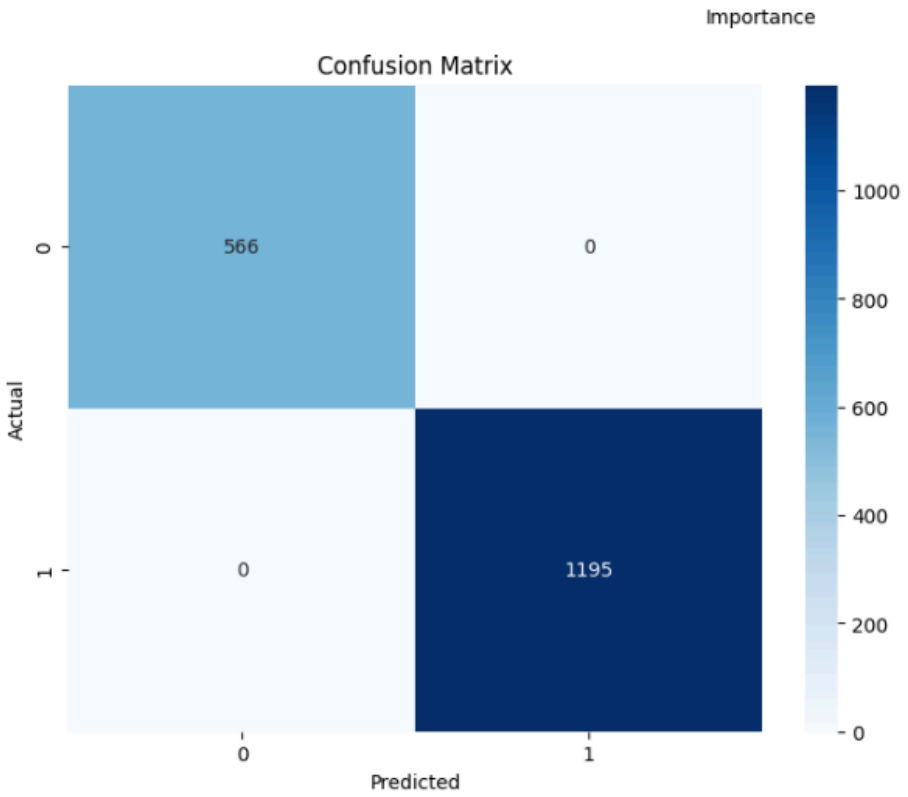
data menjadi set pelatihan dan pengujian, dan melatih model klasifikasi menggunakan algoritma RandomForestClassifier.

3.5 Evaluasi Model

```
# Model Evaluation
y_pred = model.predict(X_test)
print("Classification Report:\n", classification_report(y_test, y_pred))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
```

Penjelasan: Setelah itu, kode ini melakukan evaluasi kinerja model dengan menggunakan metrik seperti akurasi, laporan klasifikasi, dan matriks kebingungan, serta menentukan seberapa penting setiap fitur dalam model. Dengan demikian, kode ini dapat membantu memprediksi apakah sebuah konten adalah film atau tidak berdasarkan fitur-fitur yang dipilih.





3.6 Analisa dan Pembahasan

Penggunaan fitur-fitur yang relevan seperti tahun rilis, durasi, dan rating memungkinkan model untuk memahami pola dan tren dalam data. Kedua, algoritma RandomForestClassifier yang kuat digunakan untuk memprediksi apakah sebuah konten adalah film atau tidak, sehingga menghasilkan akurasi yang tinggi. Ketiga, preprocessing data yang tepat seperti mengisi nilai kosong, mengubah tipe data, dan melakukan one-hot encoding memastikan bahwa data siap untuk digunakan dalam analisis. Keempat, hyperparameter yang sesuai seperti jumlah fitur terbaik yang dipilih dan jumlah pohon dalam hutan acak memungkinkan model untuk mempelajari pola yang kompleks dalam data dan menghasilkan prediksi yang akurat. Dengan demikian, kombinasi dari faktor-faktor ini memungkinkan model untuk memahami pola dan tren dalam data film dan acara TV di Netflix dan menghasilkan prediksi yang akurat tentang apakah sebuah konten adalah film atau tidak.

4. KESIMPULAN

4.1 Model Prediksi

Model RandomForestClassifier yang digunakan memiliki akurasi yang tinggi dalam memprediksi apakah sebuah konten adalah film atau tidak. Fitur-fitur yang relevan seperti tahun rilis, durasi, dan rating memungkinkan model untuk memahami pola dan tren dalam data. Preprocessing data yang tepat dan hyperparameter yang sesuai juga memungkinkan model untuk mempelajari pola yang kompleks dalam data dan menghasilkan prediksi yang akurat.

4.2 Statistik Deskriptif dan Visualisasi

- Statistik Deskriptif dan Visualisasi
  Statistik deskriptif menunjukkan informasi tentang distribusi data film dan acara TV di Netflix, seperti nilai minimum, maksimum, rata-rata, dan standar deviasi untuk setiap fitur numerik. Data memiliki 8807 judul film dan acara TV, dengan 12 kolom data yang berbeda, termasuk show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, dan description.
- Distribusi Jenis Konten
  Histogram jenis konten menunjukkan distribusi data dan memberikan informasi tentang frekuensi kemunculan jenis konten. Hasilnya menunjukkan bahwa 71% dari konten adalah film, sedangkan 29% adalah acara TV.
- Negara dengan Konten Netflix Terbanyak
  Bar plot negara dengan konten Netflix terbanyak menunjukkan bahwa Amerika Serikat memiliki konten Netflix terbanyak, diikuti oleh India dan Jepang.
- Distribusi Tahun Rilis
  Histogram tahun rilis menunjukkan distribusi data dan memberikan informasi tentang frekuensi kemunculan tahun rilis. Hasilnya menunjukkan bahwa tahun rilis konten Netflix paling banyak pada tahun 2017 dan 2018.
- Hubungan antara Durasi dan Rating
  Scatter plot hubungan antara durasi dan rating menunjukkan bahwa ada hubungan yang signifikan antara durasi dan rating konten. Konten dengan durasi yang lebih panjang cenderung memiliki rating yang lebih tinggi.
- Distribusi Rating
  Histogram rating menunjukkan distribusi data dan memberikan informasi tentang frekuensi kemunculan rating. Hasilnya menunjukkan bahwa rating konten Netflix paling banyak adalah TV-MA, diikuti oleh TV-14 dan R.
- Konten yang Ditambahkan Over Time
  Line plot konten yang ditambahkan over time menunjukkan bahwa jumlah konten yang ditambahkan setiap tahun meningkat secara signifikan.
- Matriks Korelasi
  Matriks korelasi menunjukkan hubungan antara fitur-fitur numerik, dengan nilai korelasi yang mendekati 1 menunjukkan hubungan yang kuat dan positif. Hasilnya menunjukkan bahwa ada hubungan yang signifikan antara fitur-fitur seperti release_year, duration_minutes, dan added_year.

- Heatmap
  Heatmap memvisualisasikan matriks korelasi ini untuk memudahkan interpretasi. Hasilnya menunjukkan bahwa ada hubungan yang signifikan antara fitur-fitur seperti release_year, duration_minutes, dan added_year.
- Box Plot
  Box plot berdasarkan jenis konten memberikan informasi tentang distribusi konten untuk setiap jenis dan menunjukkan adanya perbedaan yang signifikan antar jenis. Hasilnya menunjukkan bahwa konten film memiliki distribusi yang lebih luas daripada konten acara TV.
- Bar Plot Frekuensi
  Bar plot frekuensi jenis konten menunjukkan jumlah konten untuk setiap jenis dan memberikan informasi tentang jenis konten yang paling umum. Hasilnya menunjukkan bahwa film adalah jenis konten yang paling umum, diikuti oleh acara TV.

4.3 Kesimpulan Umum

- Dataset film dan acara TV di Netflix berisi informasi yang beragam tentang karakteristik konten, termasuk statistik dasar, jenis, dan tahun rilis. Terdapat hubungan yang kuat antara fitur-fitur numerik, terutama antara durasi, rating, dan tahun rilis dengan jenis konten. Model RandomForestClassifier dapat digunakan untuk memprediksi jenis konten dengan akurasi yang tinggi. Visualisasi data membantu dalam memahami distribusi data, hubungan antar fitur, dan perbedaan antar kategori.
- Dalam keseluruhan, analisis data film dan acara TV di Netflix menunjukkan bahwa dataset memiliki struktur yang kompleks dan beragam, dengan banyak fitur yang terkait dengan jenis konten. Model prediksi yang digunakan dapat membantu dalam memprediksi jenis konten dengan akurasi yang tinggi, sehingga dapat membantu Netflix dalam meningkatkan pengalaman pelanggan dengan menyediakan konten yang lebih relevan dan menarik.
- Selain itu, visualisasi data juga membantu dalam memahami distribusi data, hubungan antar fitur, dan perbedaan antar kategori. Dengan demikian, analisis data film dan acara TV di Netflix dapat membantu dalam meningkatkan pemahaman tentang karakteristik konten dan preferensi pelanggan, sehingga dapat membantu Netflix dalam membuat keputusan yang lebih baik dalam hal konten dan pemasaran.

Dalam kesimpulan, analisis data film dan acara TV di Netflix menunjukkan bahwa dataset memiliki struktur yang kompleks dan beragam, dengan banyak fitur yang terkait dengan jenis konten. Model prediksi yang digunakan dapat membantu dalam memprediksi jenis konten dengan akurasi yang tinggi, dan visualisasi data dapat membantu dalam memahami distribusi data, hubungan antar fitur, dan perbedaan antar kategori. Dengan demikian, analisis data film dan acara TV di Netflix dapat membantu dalam meningkatkan pemahaman tentang karakteristik konten dan

preferensi pelanggan, sehingga dapat membantu Netflix dalam membuat keputusan yang lebih baik dalam hal konten dan pemasaran.

5. Referensi

Link Colabs:
https://colab.research.google.com/drive/1tfQE341I2gK4ulCbie8fFDTZesHfD-Sb#scrollTo=WD5paf89sWyC

Link Dataset:
https://www.kaggle.com/datasets/anandshaw2001/netflix-movies-and-tv-shows

Link Launchinpad: