ANALISIS MENENTUKAN REKOMENDASI FILM DENGAN MENGGUNAKAN TEKNIK PENGGALIAN DATA BERDASARKAN RIWAYAT TONTON

Abdurrahman
Program Studi Teknik Informatika, FTI, Institut Teknologi Budi Utomo Jakarta
sl.abdurrahman@gmail.com

Abstrak
Dunia perfilman berkembang pesat dengan beredarnya berbagai jenis film. Penonton seringkali kebingungan menentukan film apa yang akan ditonton. Oleh karena itu, dibutuhkan analisis untuk memberikan rekomendasi film sesuai dengan genre yang sering ditonton. Penelitian ini bertujuan untuk menganalisis rekomendasi genre film menggunakan algoritma Apriori. Penelitian ini melibatkan beberapa tahapan dalam Knowledge Discovery in Databases (KDD), yaitu seleksi data, pra-pemrosesan data, transformasi data, data mining, evaluasi, dan interpretasi hasil. Penelitian ini berhasil mengidentifikasi pola preferensi penonton berdasarkan genre film yang mereka tonton menggunakan algoritma Apriori. Hasil penelitian ini dapat digunakan oleh industri perfilman untuk memberikan rekomendasi film atau strategi pemasaran yang lebih efektif.
Kata Kunci : Rekomendasi Film, Penggalian Data, Algoritma Apriori

1. PENDAHULUAN

Penggalian data atau data mining merupakan suatu proses untuk menemukan informasi yang menarik dan tersembunyi dari suatu kumpulan data yang berukuran besar yang tersimpan dalam suatu basis data, gudang data atau tempat penyimpanan data lainnya. Teknik-teknik penambangan data yang digunakan bertugas untuk menemukan pola baru dan bermakna di dalam basis data yang mungkin masih belum diketahui. Tan, Steinbach, dan Kumar (2006:) [1]

Algoritma Apriori adalah algoritma yang berpengaruh untuk aturan asosiasi. Algoritma apriori termasuk jenis aturan asosiasi pada penggalian data. Aturan yang menyatakan asosiasi antara beberapa atribut sering disebut analisis keranjang pasar. Analisis asosiasi atau *association rule mining* adalah teknik penggalian data untuk menemukan aturan suatu kombinasi item.

Salah satu tahap analisis asosiasi yang menarik perhatian banyak peneliti untuk menghasilkan algoritma yang efisien adalah analisis pola frekuensi tinggi (*frequent pattern mining*). Penting tidaknya suatu asosiasi dapat diketahui dengan dua tolak ukur, yaitu: *support* dan *confidence*. nilai penunjang (*support*) adalah persentase kombinasi item tersebut dalam basis data, sedangkan nilai kepastian (*confidence*) adalah kuatnya hubungan antar-item dalam aturan asosiasi. Tanna & Ghodasara (2014) [2].

2. METODOLOGI

Metode penulisan penelitian ini menggunakan metode kualitatif, pendekatan penelitian ini memanfaatkan data. Semua data diperoleh dari teknik pengumpulan data.

Gambar 1 Tahapan Implementasi data

Sumber Data : Hasil Olahan Data Penelitian

Tahapan Implementasi data antara lain :

1. Proses Pengumpulan data : proses pengumpulan data yang digunakan adalah data kualitatif. Pengumpulan data dilakukan adalah dengan melakukan pengumpulan data dari kaggle.com. proses pengumpulan data dari berbagai sumber penelitian dan dari data movie.
2. Pada proses pengumpulan data terdapat beberapa dataset dengan isi data yang tidak sama, sehingga harus mencari ulang data yang sesuai.
3. Analisis Data: dilakukan untuk mengetahui informasi dari dataset yang dikumpulkan.
4. Tahap praproses data atau sering dikenal dengan istilah *data preprocessing* adalah proses untuk merubah data mentah kedalam bentuk yang mudah dipahami. Berikut adalah beberapa langkah praproses data :
   a. Kategorisasi : data dikelompokan berdasarkan tipe data yaitu tipe data Integer, Float dan tipe data object.
   b. Pembersihan : data yang sudah di kategorisasi di bersihkan melalui beberapa proses seperti mengisi nilai yang kosong, menghapus kolom yang tidak diperlukan, dll
   c. Transformasi : mengubah format data yang sesuai dengan kebutuhan analisis
5. Penerapan Algoritma Apriori : penerapan algoritma dilakukan untuk mendapatkan hasil Analisis dari data movie.
   a. Perhitungan nilai support
   $$Support(A \cap B) = \frac{\text{Jumlah Transaksi Menggunakan A dan B}}{\text{Total Transaksi}} \times 100\%$$
   b. Perhitungan nilai confidence
   $$Confidence\ (A|B) = \frac{\text{Jumlah Transaksi Menggunakan A dan B}}{\text{Total Transaksi Mengandung A}} \times 100\%$$
   c. Perhitungan nilai lift
   $$\text{Lift ratio} = \frac{\text{Support (A}\cap\text{B)x Support (B)}}{\text{Support (A)}}$$

3. HASIL DAN PEMBAHASAN

Tahapan dalam Knowledge Discovery in Databases (KDD) :

1. Seleksi Data
2. Pra-pemrosesan data
3. Transformasi data
4. Data Mining
5. Evaluasi dan Interpretasi

3.1 Prosedur Pembentukan dataset

Sebelum melakukan tahapan KDD ada beberapa prosedur yang harus dipersiapkan :

1. Menginstall Library
- %pip install plotly
- %pip install mlxtend pandas
- %pip install matplotlib seaborn pandas
- %pip install word cloud
- %pip install dask
2. Mengimport library
- Import numpy as np
- import matplotlib.pyplot as plt
- *import seaborn as sns*
- *from scipy (Scientific python) import stats*
- from sklearn.preprocessing import LabelEncoder
- from mlxtend.frequent_patterns import apriori, association_rules
- from wordcloud import WordCloud

3.1. Data Selection

Sumber data yang diambil dari kaggle.com dalam format file .CSV (*Comma Separated Values*). Terdapat beberapa dataset yang digunakan sebagai target data untuk analisis ini.

1. Dataset I Pada dataset pertama dengan nama file movies.csv memiliki beberapa rincian yaitu:

a. Baris: 722.408 baris
b. Fields: 20 kolom

2. Dataset II :
   Pada dataset pertama dengan nama file netflix_dataset.csv memiliki beberapa rincian yaitu:

a. Baris: 7.791 baris
b. Fields: 11 kolom

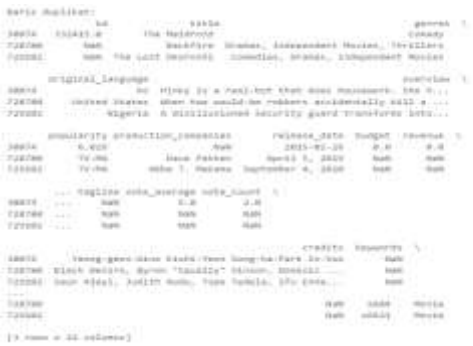
Tabel yang akan diganti nama kolom adalah tabel dua. Berikut adalah beberapa kolom yang diganti nama kolom/field:

1. 'Title': 'title',
2. 'Director': 'production_companies',
3. 'Cast':'credits',
4. 'Country': 'original_language',
5. 'Release_Date': 'release_date',
6. 'Duration': 'runtime',
7. 'Type': 'genres',
8. 'Description': 'overview',
9. 'Rating': 'popularity'

ISSN 2406-9841
e-IISN 2986-2973

3.2 Data Cleaning

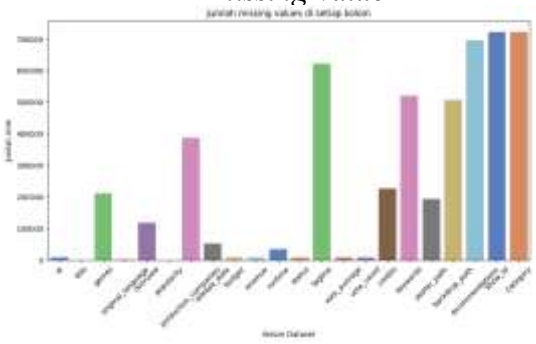
1. Data Duplikat



Gambar 2. Tampilan baris data duplikat
Sumber Data : Hasil Olahan Data Penelitian

Dari gambar diatas terdapat 3 baris data duplikat pada data pada dataset yaitu pada baris 30074, baris 728780 dan baris 729102.

1. Mencari *missing value*



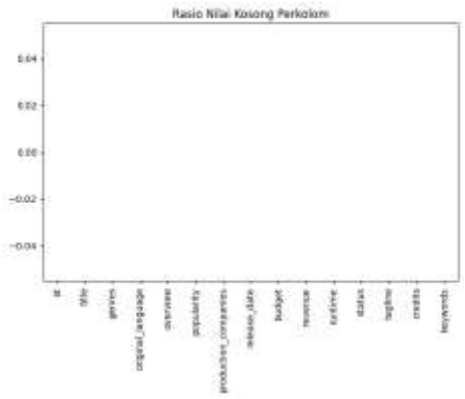
Gambar 3. Visualisasi data yang memiliki *missing value*
Sumber Data : Hasil Olahan Data Penelitian

Langkah-langkah untuk membersihkan data yang memiliki *missing value* sangat penting, karena menentukan jumlah data yang akan digunakan untuk tahap data mining selanjutnya.

Untuk mencegah terhapusnya data yang penting, digunakan metode ffill:

df.fillna(method='ffill',inplace=True)

Metode ffill berfungsi untuk mengisi nilai yang kosong dengan nilai sebelumnya. Setelah langkah pengisian data tersebut, barulah dilakukan penghapusan *missing value* yang memiliki banyak nilai kosong.



Gambar 4. Visualisasi data *missing value* setelah dibersihkan
Sumber Data : Hasil Olahan Data Penelitian

2. Menghapus data kolom yang tidak diperlukan

```
<class 'pandas.core.frame.DataFrame'>
Index: 0 entries
Data columns (total 22 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   id                    0 non-null      float64
 1   title                 0 non-null      object
 2   genres                0 non-null      object
 3   original_language     0 non-null      object
 4   overview              0 non-null      object
 5   popularity            0 non-null      object
 6   production_companies  0 non-null      object
 7   release_date          0 non-null      object
 8   budget                0 non-null      float64
 9   revenue               0 non-null      float64
 10  runtime               0 non-null      object
 11  status                0 non-null      object
 12  tagline               0 non-null      object
 13  vote_average          0 non-null      float64
 14  vote_count            0 non-null      float64
 15  credits               0 non-null      object
 16  keywords              0 non-null      object
 17  poster_path           0 non-null      object
 18  backdrop_path         0 non-null      object
 19  recommendations       0 non-null      object
 20  Show_Id               0 non-null      object
 21  Category              0 non-null      object
dtypes: float64(5), object(17)
memory usage: 0.0+ bytes
```

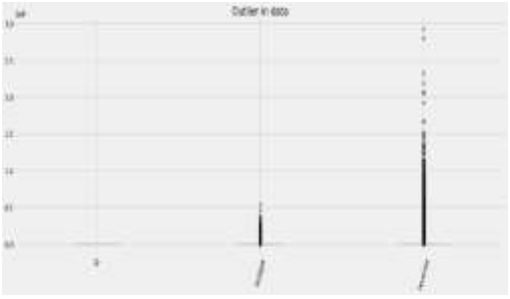
Gambar 5. Info data kolom sebelum dihapus
Sumber Data : Hasil Olahan Data Penelitian

```
<class 'pandas.core.frame.DataFrame'>
Index: 0 entries
Data columns (total 15 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   id                    0 non-null      float64
 1   title                 0 non-null      object
 2   genres                0 non-null      object
 3   original_language     0 non-null      object
 4   overview              0 non-null      object
 5   popularity            0 non-null      object
 6   production_companies  0 non-null      object
 7   release_date          0 non-null      object
 8   budget                0 non-null      float64
 9   revenue               0 non-null      float64
 10  runtime               0 non-null      object
 11  status                0 non-null      object
 12  tagline               0 non-null      object
 13  credits               0 non-null      object
 14  keywords              0 non-null      object
dtypes: float64(3), object(12)
memory usage: 0.0+ bytes
```

Gambar 6. Info data setelah hapus kolom
Sumber Data : Hasil Olahan Data Penelitian

Info diatas menampilkan kolom setelah dihapus adalah sebagai berikut: id, title, original_language, overview, popularity, production_companies, release_date, budget, revenue, runtime, status, tagline, vote_average, vote_count, credits, keywords

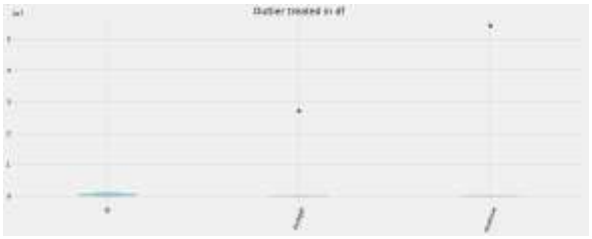
3. Menghapus Outlier



Gambar 7. Visualisasi data sebelum dihapus outlier
Sumber Data : Hasil Olahan Data Penelitian

Visualisasi diatas terdapat dua titik koodinat : yaitu koordinat y jumlah panjang data outliers hingga 3.0 dan koordinat x adalah nama kolom/field, terdapat 2 kolom outliers data yaitu kolom budget dan kolom revenue.

- Pada data budget memiliki outliers 2 outliers yang menyimpang hingga 0,5.
- Pada data revenue memiliki outliers yang menyimpang jauh yaitu 3.0



Gambar 8. Visualisasi data setelah dihapus outlier
Sumber Data : Hasil Olahan Data Penelitian

Visualisasi diatas setelah dihapus outliersnya tidak terdapat nilai pada koordinat x maupun y.

3.3 Data Transformation

Pada tranformasi data berdasarkan genre dari film

Gambar 9. Transformasi Data
Sumber Data : Hasil Olahan Data Penelitian

Proses encoding : mengubah teks menjadi kode biner yang bisa dimengerti oleh komputer.
Proses encoding atau transformasi data yaitu dengan cara menampilkan data film dengan menampilkan nilai 0 dan nilai 1. Nilai 0 adalah genre yang kosong dan tidak memiliki nilai dan nilai 1 adalah nilai yang berisi genre. Ketika bernilai 1 maka pada baris tersebut terdapat genre film tersebut.

3.4 Data Mining

1. Penerapan Algoritma Apriori

Terdapat 3 nilai yang dihitung pada penerapan algoritma apriori :

1. Perhitungan nilai support

$$Support(A \cap B) = \frac{\text{Jumlah Transaksi Mengandung } A \text{ dan } B}{\text{Total Transaksi}} \times 100\%$$

Menampilkan nilai support

	support	itemsets
0	0.224044	(Action)
1	0.169399	(Adventure)
2	0.087432	(Animation)
3	0.229508	(Comedy)
4	0.076503	(Crime)

Gambar 10. Hasil Nilai Support dengan nillai minimal 0,01
Sumber Data : Hasil Olahan Data Penelitian

Data di atas adalah hasil dari proses data mining menggunakan algoritma Apriori untuk menemukan itemset yang sering muncul dalam dataset. Mari kita bedah setiap bagian dari data tersebut:

a. **Support**: Support adalah ukuran seberapa sering item atau itemset muncul dalam dataset dan nilai itemset dibagi dengan 100. Nilai 0.224044 berarti bahwa itemset "Action" muncul dalam sekitar 22.4% dari total transaksi atau data dalam dataset.
b. **Itemsets**: Itemset adalah kumpulan item yang sering muncul bersama dalam dataset. Dalam contoh ini, itemset yang ditemukan adalah "Action", yang menunjukkan bahwa genre "Action" sering muncul dalam dataset.

Secara keseluruhan, data ini menunjukkan bahwa genre "Action" cukup populer dan muncul dalam sekitar 22.4% dari data yang dianalisis. Informasi ini dapat digunakan untuk membuat aturan asosiasi yang berguna

dalam memahami pola atau hubungan antar item dalam dataset.

2. Perhitungan nilai *confidence*

$$Confidence\ (A|B) = \frac{Jumlah\ Transaksi\ Menggunakan\ A\ dan\ B}{Total\ Transaksi\ Mengandung\ A} \times 100\%$$

Menampilkan nilai *confidence*

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(Animation)	(Family)	0.087432	0.103825	0.060109	0.687500	6.621711
1	(History)	(Drama)	0.054645	0.393443	0.043716	0.800000	2.033333
2	(Music)	(Drama)	0.010929	0.393443	0.010929	1.000000	2.541667
3	(Romance)	(Drama)	0.092896	0.393443	0.065574	0.705882	1.794118
4	(Action, Animation)	(Adventure)	0.027322	0.169399	0.021858	0.800000	4.722581

Gambar 11. Hasil nilai perhitungan confidence dengan nilai minimal 0,6
Sumber Data : Hasil Olahan Data Penelitian

Hasil perhitungan nilai confidence pada data di atas:

a. Antecedents: Ini adalah item atau itemset awal. Dalam contoh ini, antecedent-nya adalah "(Animation)".
b. Consequents: Ini adalah item atau itemset yang muncul setelah antecedent. Dalam contoh ini, consequent-nya adalah "(Family)".
c. Antecedent Support: Ini adalah nilai support untuk antecedent, yaitu frekuensi kemunculan item atau itemset dalam dataset. Nilai 0.087432 berarti itemset "Animation" muncul dalam sekitar 8.7% dari total transaksi atau data dalam dataset.
d. Consequent Support: Ini adalah nilai support untuk consequent, yaitu frekuensi kemunculan item atau itemset dalam dataset. Nilai 0.103825 berarti itemset "Family" muncul dalam sekitar 10.4% dari total transaksi atau data dalam dataset.
e. Support: Ini adalah nilai support untuk aturan asosiasi lengkap (antecedent dan consequent muncul bersama-sama). Nilai 0.060109 berarti bahwa itemset "(Animation, Family)" muncul bersama-sama dalam sekitar 6.0% dari total transaksi atau data dalam dataset.
f. Confidence: Confidence mengukur seberapa sering consequent muncul dalam transaksi yang juga mengandung antecedent. Nilai confidence dihitung sebagai berikut: Nilai Confidence sekitar 68.75% dari transaksi yang mengandung "Animation" juga mengandung "Family".
g. Lift: Lift mengukur kekuatan aturan asosiasi dibandingkan dengan kemunculan consequent secara acak. Nilai lift dihitung adalah: Nilai lift sebesar 6.621711 berarti bahwa kemunculan "Family" dalam transaksi yang mengandung "Animation" adalah sekitar 6.62 kali lebih tinggi daripada kemunculannya secara acak.

Jadi, data tersebut menunjukkan bahwa terdapat hubungan yang cukup kuat antara "Animation" dan "Family". Jika sebuah transaksi mengandung "Animation", maka kemungkinan besar transaksi tersebut juga mengandung "Family" dengan confidence sebesar 68.75% dan lift sebesar 6.62

3. Perhitungan nilai lift

$$\text{Lift ratio} = \frac{\text{Support (A}\cap\text{B)x Support (B)}}{\text{Support (A)}}$$

Menampilkan nilai Lift

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(Adventure)	(Action)	0.169399	0.224044	0.098361	0.580645	2.591660
1	(Action)	(Adventure)	0.224044	0.169399	0.098361	0.439024	2.591660
2	(Action)	(Animation)	0.224044	0.087432	0.027322	0.121951	1.394817
3	(Animation)	(Action)	0.087432	0.224044	0.027322	0.312500	1.394817
4	(Fantasy)	(Action)	0.114754	0.224044	0.043716	0.380952	1.700340

Gambar 12. Hasil perhitungan nilai lift dengan batas minimal nilai 1.0
Sumber Data : Hasil Olahan Data Penelitian

Berikut adalah cara menghitung nilai lift dari data di atas:

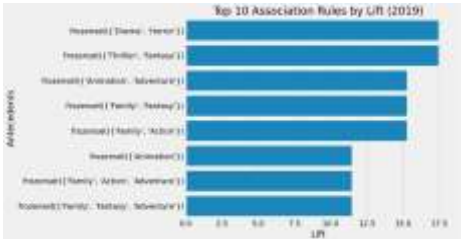
a. **Antecedent Support (Adventure)**: Proporsi transaksi yang mengandung genre "Adventure". Misalnya, jika ada 100 transaksi dan 16,939 di antaranya mengandung "Adventure", maka antecedent support adalah 16,939 / 100 = 0.169399.
b. **Consequent Support (Action)**: Proporsi transaksi yang mengandung genre "Action". Misalnya, jika ada 100 transaksi dan 22,404 di antaranya mengandung "Action", maka consequent support adalah 22,404 / 100 = 0.224044.
c. **Support (Adventure -> Action)**: Proporsi transaksi yang mengandung kedua genre "Adventure" dan "Action". Misalnya, jika ada 100 transaksi dan 9,836 di antaranya mengandung kedua genre tersebut, maka support adalah 9,836 / 100 = 0.098361.
d. **Confidence (Adventure -> Action)**: Kemungkinan bahwa genre "Action" akan

ISSN 2406-9841
e-IISN 2986-2973

muncul dalam transaksi yang sudah mengandung genre "Adventure".
Confidence dihitung sebagai support dari aturan dibagi dengan antecedent support.
Confidence = Support / Antecedent Support = 0.098361 / 0.169399 = 0.580645.

e. **Lift (Adventure -> Action)**: Ukuran seberapa banyak genre "Adventure" meningkatkan kemungkinan genre "Action".
Lift dihitung sebagai confidence dibagi dengan consequent support.
Lift = Confidence / Consequent Support = 0.580645 / 0.224044 = 2.591660.

3.4.1 Hasil Analisis Algoritma Apriori



Gambar 13. Hasil genre terbesar dari perhitungan nilai lift diatas 1.0

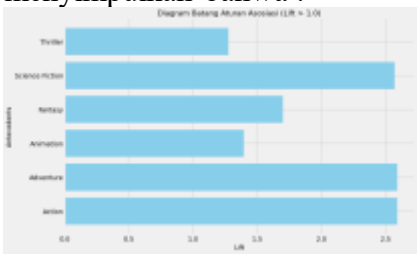
Pada sumbu atau titik koordinat x menampilkan nilai lift dan sumbu atau titik koordinat y menampilkan data nilai pertama atau Antecedents.

- Pada data teratas terdapat kombinasi data genre drama dan horor dengan nilai 17.5
- Pada data teratas ke 2 terdapat genre Thiller dan Fantasi dengan nilai 17.5
- Data ke 3 terdapat genre Animation dan Adventure dengan nilai diatas 15.0
- Data ke 4 terdapat genre Family dan Fantasi dengan nilai diatas 15.0
- Data ke 5 terdapat genre Family dan Action dengan nilai diatas 15.0
- Data ke 6 terdapat genre Animation dengan nilai diatas 10.0
- Data ke 7 terdapat genre Family, Action dan Adventure dengan nilai diatas diatas 10.0
- Data ke 8 terdapat genre Family, Fantasi dan Adventure dengan nilai diatas 10.0

3.5 Interprestasi/Evaluasi

3.5.1 Kesimpulan dari data

Pada tahap evaluasi menyimpulkan bahwa :



Gambar 14. Hasil nilai perhitungan lift dengan nilai pertama lebih dari 1.0

Pada hasil visualisasi data di atas, dapat dilihat bahwa genre film dengan nilai antencenden terbesar atau yang paling banyak ditonton memiliki nilai di atas 1. Titik koordinat y menunjukkan nilai antencenden, sementara titik koordinat x menunjukkan nilai lift.

- Genre Action, Adventure, dan Science Fiction memiliki nilai lift diatas 2.5
- Genre Fantasy memiliki nilai lift diatas 1.5
- Animation dan Thiller memiliki nilai lift diatas 1.0

3.5.2 Rekomendasi berdasarkan Analisis

Gambar 15. Rekomendasi berdasarkan hasil analisis genre

Hasil di atas menunjukkan pola preferensi penonton berdasarkan genre film yang mereka tonton. Dalam data tersebut, "tingkat kepercayaan" merujuk pada seberapa besar kemungkinan seseorang yang menonton satu genre film juga akan menonton genre film lain. Mari kita analisis beberapa temuan utama:

1. Action dan Adventure:
   - Jika seseorang menonton film Action, ada kecenderungan

ISSN 2406-9841
e-IISN 2986-2973

sebesar 43.90% bahwa mereka juga akan menonton film Adventure.
- Sebaliknya, jika seseorang menonton film Adventure, kemungkinan mereka juga menonton film Action lebih besar, yaitu sebesar 58.06%.
- Hal ini menunjukkan bahwa kedua genre ini memiliki hubungan yang cukup erat dalam preferensi penonton.

2. Action dan Animation:
   - Jika seseorang menonton film Action, ada kemungkinan sebesar 12.20% bahwa mereka juga akan menonton film Animation.
   - Namun, jika seseorang menonton film Animation, kemungkinan mereka menonton film Action adalah sebesar 31.25%.
   - Walaupun tingkat kepercayaan ini lebih rendah dibandingkan dengan Adventure, ada keterkaitan yang tetap signifikan antara kedua genre ini.
3. Fantasy dan Action:
   - Jika seseorang menonton film Fantasy, kemungkinan mereka juga akan menonton film Action adalah sebesar 38.10%.
   - Hal ini menunjukkan bahwa meskipun tidak sekuat hubungan antara Action dan Adventure, ada ketertarikan yang cukup signifikan di kalangan penonton film Fantasy terhadap film Action.

4. KESIMPULAN DAN SARAN

Dapat disimpulkan bahwa terdapat pola preferensi penonton film yang cukup signifikan berdasarkan genre. Temuan utama menunjukkan bahwa genre Action dan Adventure memiliki keterkaitan yang erat, dengan penonton yang menyukai satu genre memiliki kemungkinan besar untuk juga menyukai genre lainnya.

Sementara itu, hubungan antara genre Action dan Animation, serta Fantasy dan Action, juga menunjukkan keterkaitan yang signifikan meskipun tidak sekuat hubungan antara Action dan Adventure. Pola-pola preferensi ini dapat memberikan wawasan berharga bagi industri perfilman dalam

menyusun rekomendasi film atau strategi pemasaran yang lebih efektif dan tepat sasaran.

5. DAFTAR PUSTAKA

[1] Arhami & Nasir. *Data mining algoritma dan impelementasi*. Aceh:2020
[2] Mustika. *Data Mining dan Aplikasinya*. Bandung:2021
[3] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. *Introduction to Data Mining*. Boston:2006

[4] Purnama. *Pengantar Machine Learning*. Bandung:2019