CSE 487 Assignment-1

Strategic Content Optimization for Amazon Prime Video

Aditya Srivatsav Lolla(alolla |50559685)
Sai Krishna Goud Valdas(svaldas | 50560726)

Problem Statement:

This database related to Amazon Prime Video can be a great source of information for studying viewer preferences, content types, and ration trends. Studying this data allows us to learn what types of content people tune into, what genres sell, as well as what ratings clients will tune into and in what specific demographic group. This analysis can be incredibly beneficial in finding patterns and trends to shape content creation, acquisition strategies, and marketing. This understanding is key to optimising viewer satisfaction and engagement, leading to increased subscription growth and reduced churn.

The main intention of this problem statement is to fetch the viewer interest's clarity of data using data analysis methodologies of Amazon Prime Video. By leveraging content types and ratings trends, we aim to answer what kind of content best strikes the chord with the viewers and whether the ratings accurately reflect viewer delight. This insight can be used to personalise the content being recommended, personalise marketing communication, and ultimately enhance the user experience. The ultimate goal is to use data-driven findings to improve the platform's content strategy and align it more closely with the developing needs of a viewer population.

This analysis has a quite broad scope, as can be seen from the example uses mentioned. As for content strategy, user data can inform what new content to create or licence to ensure that investments are placed in areas of potential high interest and likely to result in high viewer satisfaction. On the marketing side, analytics from viewer preferences can help determine where groups of subscribers are, and build marketing campaigns around popular content to attract more subscribers or retain existing ones. With reported user-experience, understanding rating trends would further enable the understanding of trends in recommendations, making them more personalised and fine-tuned. This multifaceted strategy promises to give Amazon Prime Video a major advantage in the streaming marketplace.

The analysis is based on some of the toughest parts of managing a media streaming service. The largest challenge is comprehending the diverse and massive user tastes around the globe. Knowing what trends and preferences are present helps Amazon Prime Video find the best way to cater to the different kinds of viewers. This provides an additional layer of content spend optimization by making sure that the content assets invested in are those that bring, from a viewership perspective, the highest level of impact and engagement, hence viewer satisfaction. In addition, the findings aid in solving the issues of pattern detection of subscribers with content that continuously corresponds to current viewer expectations and preferences. In conclusion, this analysis not only drives your content

strategy and marketing efforts but also solves fundamental problems regarding a better streaming service.

Overview of dataset :

The dataset pertaining to Amazon Prime Video provides insights, into the content on its platform. Each entry contains information about a variety of shows and movies offering a glimpse into preferences, content trends and market dynamics. Lets now explore the dataset in detail:

Total rows : 9,668
Total columns : 12
Link: https://www.kaggle.com/datasets/shivamb/amazon-prime-movies-and-tv-shows/data

Categories and Details :

Show_id : Unique identifier for each show.
Type : Category of content
Title : Name of the show or movie.
Director : Filmmaker behind the content
Cast : The people who acted in the movie or TV Show
Country: It tells about the specific Country does the movie or TV Show belongs to
Date_added: The release date
Release_year: The release year
Rating: The rating of the movie as per the content Family, Adult, Teen and others
Duration: Time duration of the movie
Listed_in: It tells about genre of the movie
Description: The description of the movie

Data Cleaning / Processing :

Remove the Duplicates

We eliminated duplicates based on the show_id column to guarantee that every entry in the dataset is distinct. This stage ensures data integrity for next studies by confirming that each program or film is represented uniquely.

Handling missing values

We rectified the missing values by substituting the proper values for them. In particular, ‘Unknown’ was used to fill in the missing data in the director, cast, nation, and date_added fields. This method stops important documents from being lost because of missing information.

Standardise the Text Data

In this we standardised the entries in the director and country columns to address inconsistencies in textual data. This required eliminating any leading or trailing whitespace and changing every entry to title case. By ensuring consistency in naming conventions, this makes reliable analysis possible.

Handling Missing Ratings with Mode

Here the column mode was used to fill in this missing values in the rating column. The overall rating patter is preserved when the most frequent rating is used, as this guarantees that the imputation is based on the dataset intrinsic distribution.

Convert Datatype for the Datetime

To enable time-based analysis, we changed the data_aded field to a datetime format. This phase permits consistent data manipulations and additional data-related actions, despite format inference warnings.

Extract Year from data_added

In this step we took the year out of the data_added column and made a new column called year_added to make it easier to analyse patterns over time. Zeros were used to fill in any missing data, and the column was changed to an integer type.

Standardising ‘duration’ Field to Numeric

In order to express duration in minutes, we took numerical values out of the duration column and standardised them. The quantitative study of content length is supported by this conversion to numeric representation.

Categorise the Content Based to Ratings

Based on ratings, we divided the content into four categories : “Family,” “Teen,” “Adult,” and “Others”. By putting material into more general categories, this categorization makes audience targeting and content recommendation tactics easier.

Identification and Removal of Outliers

We eliminated the entries whose duration_minutes above the 99th percentile in order to address excessive values in content length. By taking this precaution, outliers are prevented from distorting the study and giving a more realistic depiction of average content lengths.

Identification of the Primary Genre

The primary genre is assigned from the Amazon data from the overall list and a separate column is created. Examples are Action, Drama and Thriller. The primary genre is Action.

Statistics and Analysis:

The dataset covers the average release year of 2008, the dataset spans a wide range of release years from 1920 to 2021 and features a combination of vintage and modern content. The items have an average duration of approximately 72 minutes, which suggest a significant variation in the length of the material. Despite its limitations, the ‘data_added’ data exhibits a regular pattern, with the majority of additions taking place in 2021, notably in the middle of july.

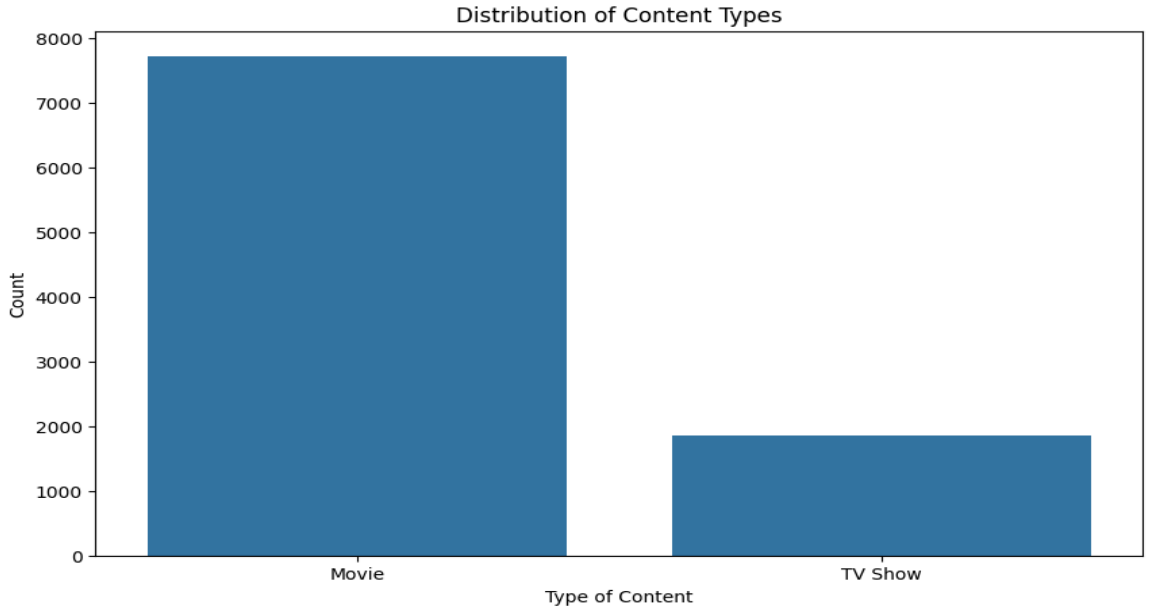
```
                          date_added  release_year   year_added  \
count                            155   9582.000000  9582.000000
mean   2021-07-14 13:46:50.322580736   2008.343456    32.692027
min              2021-03-30 00:00:00   1920.000000     0.000000
25%              2021-05-23 00:00:00   2007.000000     0.000000
50%              2021-07-20 00:00:00   2016.000000     0.000000
75%              2021-09-16 00:00:00   2019.000000     0.000000
max              2021-10-10 00:00:00   2021.000000  2021.000000
std                              NaN     18.945813   254.967845

       duration_minutes
count       9582.000000
mean          72.209038
min            0.000000
25%           40.000000
50%           86.000000
75%          100.000000
max          170.000000
std           44.087542
```

Exploratory Data Analysis

Exploratory data Analysis is the approach which identifies the data. It is the first step of the Analysis.

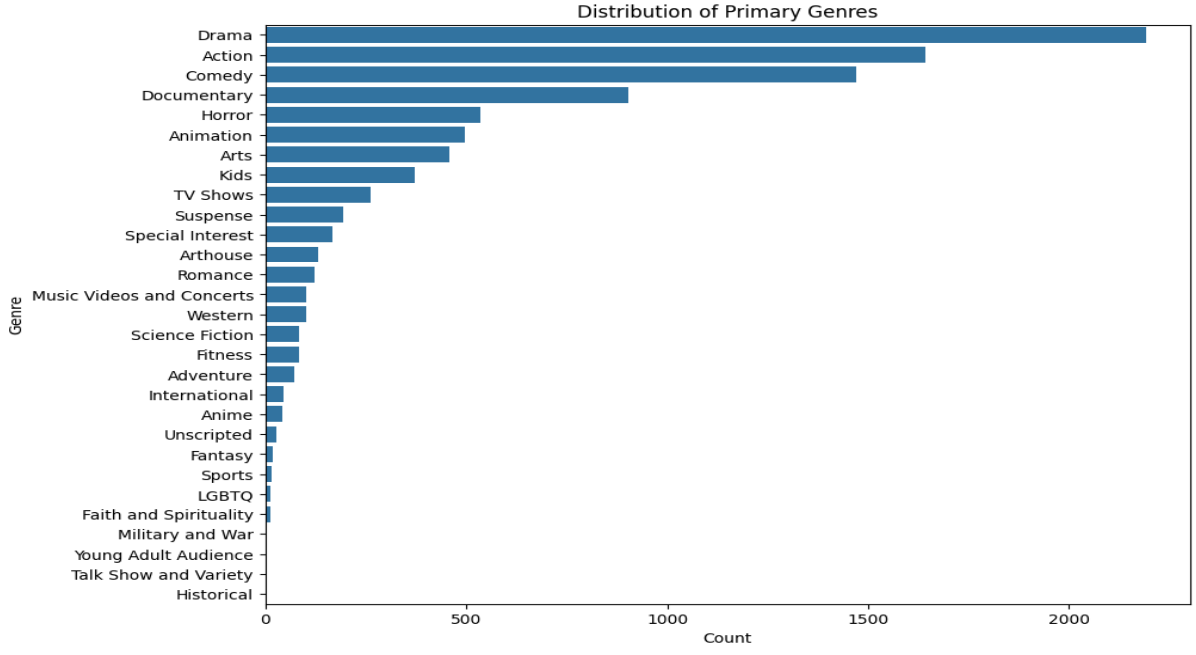
Content Type Distribution :

Examined how different content genres were distributed (TV series Versus movies.)



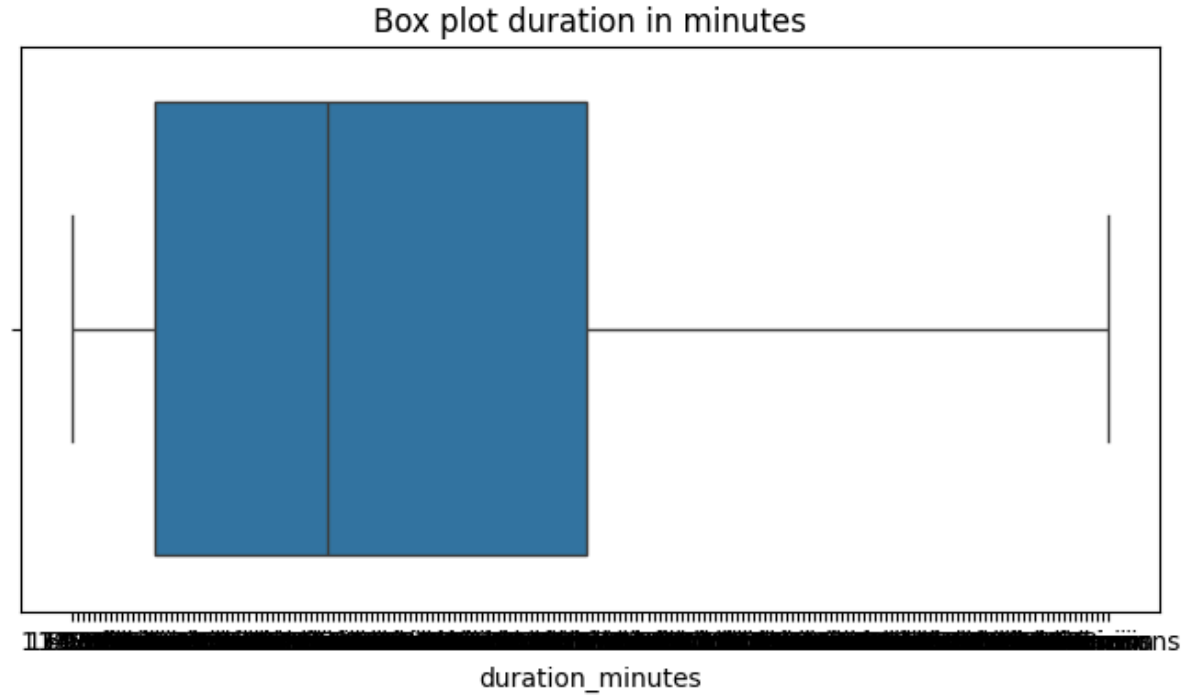
Primary Geners Distribution :

A bar plot was used to visualised the primary genre distribution.



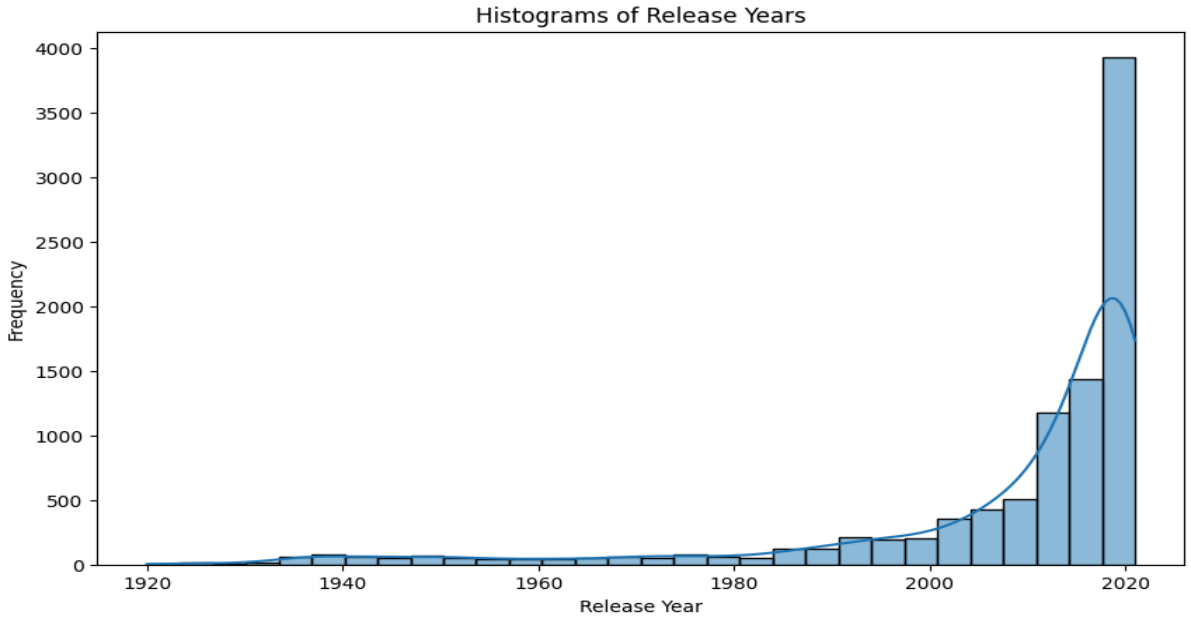
Duration Analysis :

Box and violin plots were used to analyse the length of the content.



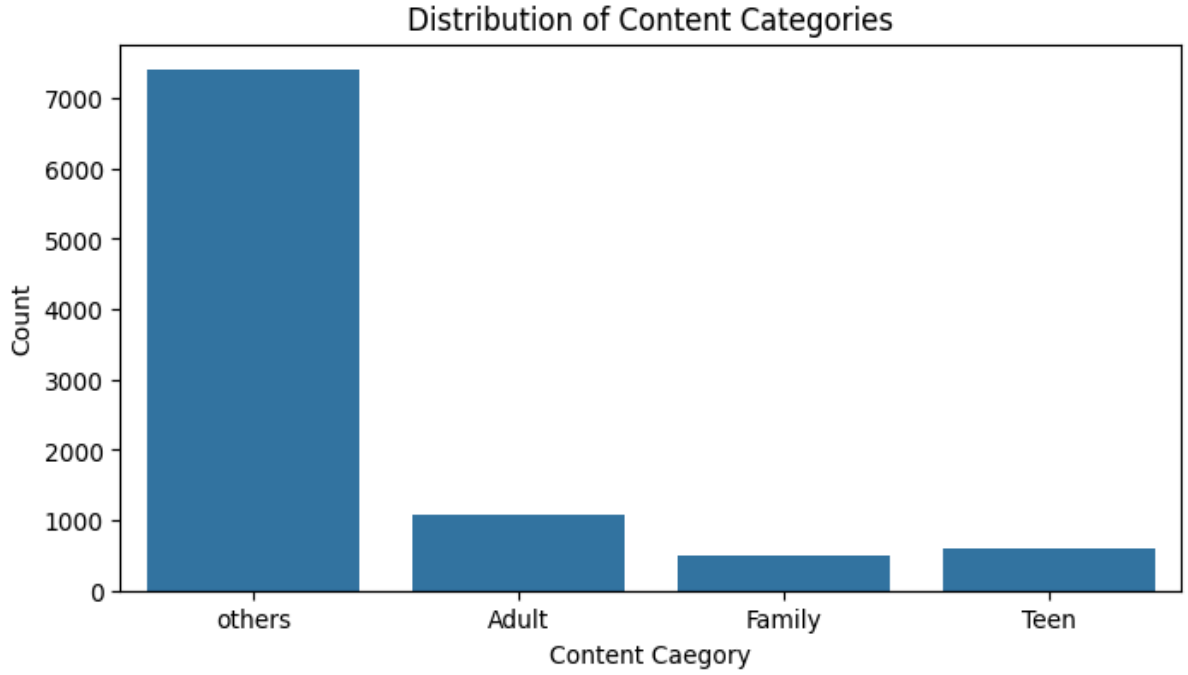
Release Year Analysis :

A histogram was used to analyse the distribution of release years.



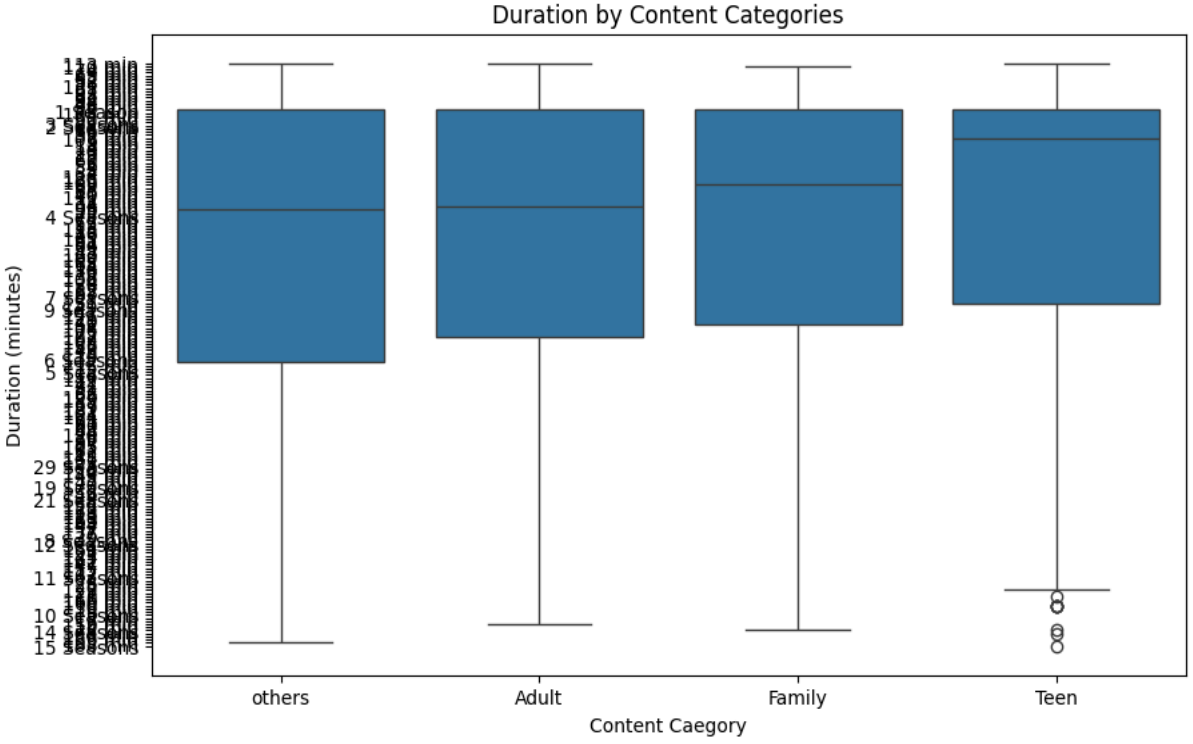
Category-Based Content Distribution :

Count plots were used to visualise the distribution of content by categories.



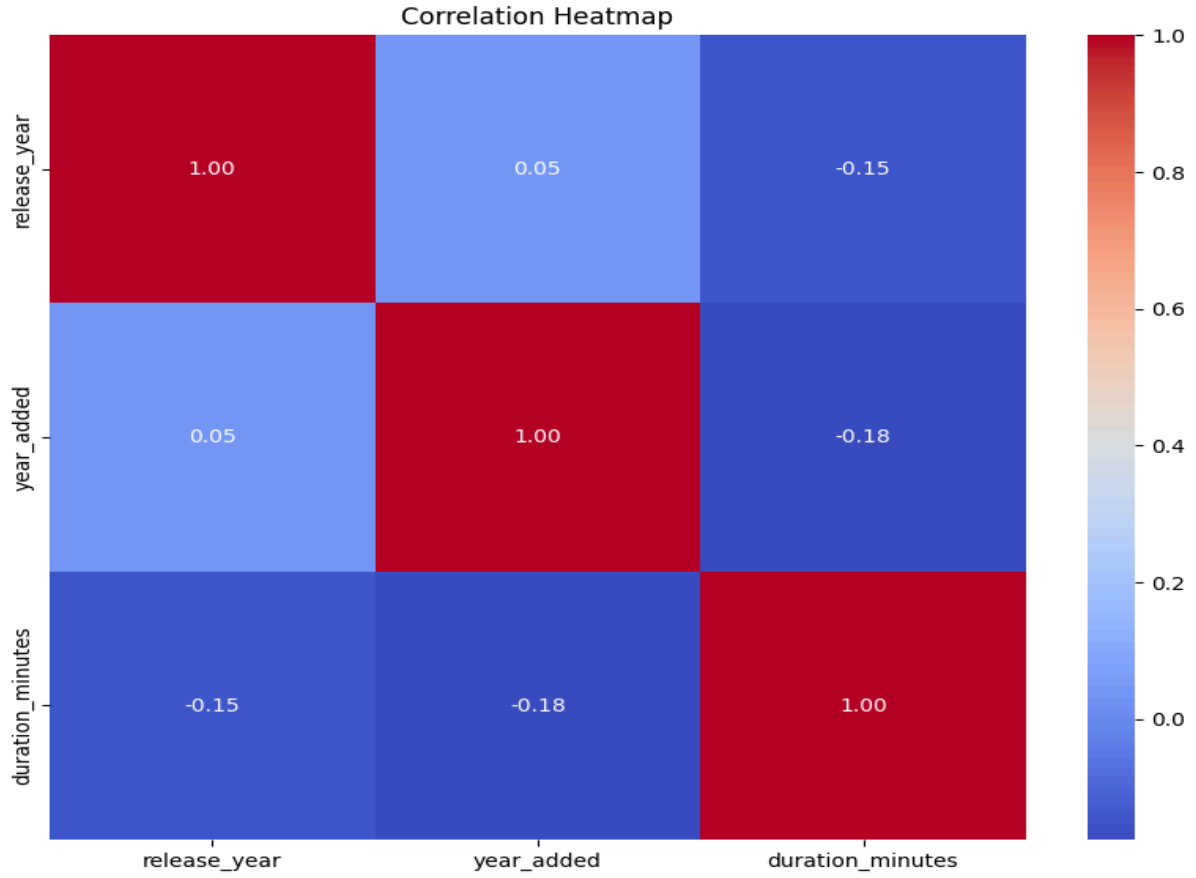
Duration vs Content Category :

Box plots were used to analysed the length of material in various categories.



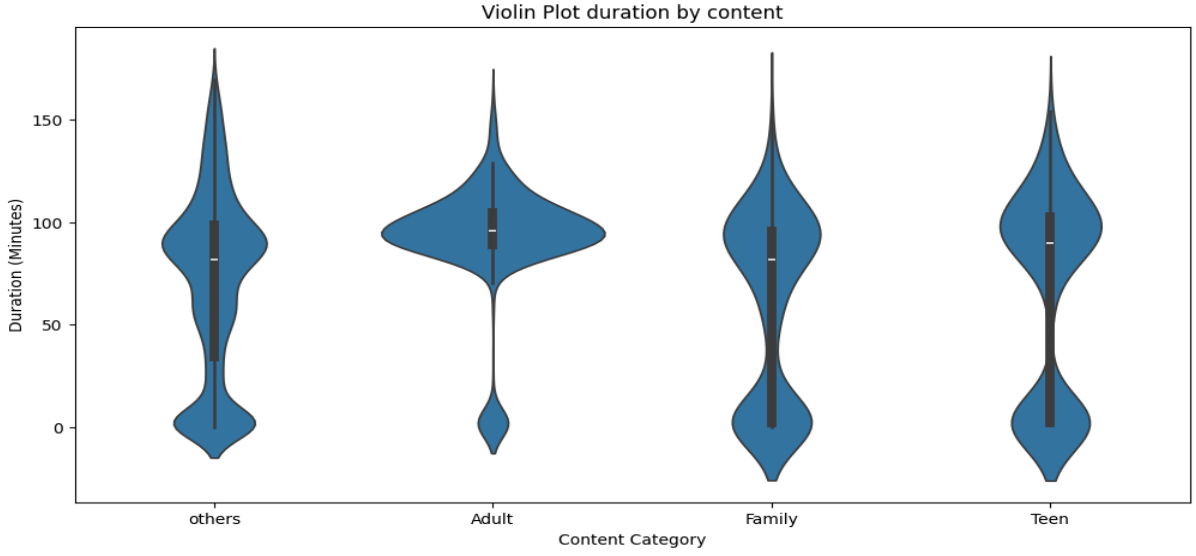
Correlation Heatmap :

A Heatmap is used to visualise correlation between numerical variables.



Violin Plot Analysis of Content Duration by Category :

This violin plot shows how the length of content varies in the following categories : Teen, Adult, Family, and Others. It demonstrates that there is a wider variety of durations with distinct peaks in the Others and Adult groups. The Family and Teen categories, on the other hand, have less diversity in their lengths with more of their durations centred around a core point.



Model Evaluation:

Libraries Used:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix
from sklearn.svm import SVC
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OneHotEncoder
from sklearn.neighbors import KNeighborsClassifier
from    sklearn.metrics    import    accuracy_score,    precision_score,
recall_score, f1_score, precision_recall_curve, average_precision_score
from sklearn.model_selection import GridSearchCV, StratifiedKFold
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import roc_curve, auc
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
import numpy as np
```

Training and Testing Data:

1. Data loading, Feature Selection:

- Information is loaded from a CSV file. The cleaned data comprise the titles available on Amazon Prime.
- From this dataset, select specific features: ('type', 'duration_minutes', 'release_year', 'content_category') to form your feature matrix X. Now set target variable y to the 'primary genre' of the titles, which you would presumably be trying to predict.

2. Categorical Variable Encoding:

- It detects the variables that are categorical in nature, like 'type' and 'content_category', so that these variables need to be encoded to be processed correctly by the machine learning algorithm.
- Two categorical columns altered via LabelEncoder of scikit-learn; this will transform the categorical string values into numeric code. Transformation applied consistently on both train and test data.

3. Split up Data and Standardization:

- Here, the dataset will be divided into training and test sets in an 80:20 ratio, which is the standard way of evaluating a machine learning model.
- Now, the feature data, X_train and X_test, will be standardized by StandardScaler. This is a way of feature scaling by which features will now have zero mean and unit variance, and on many occasions this can significantly improve the performance and speed up the convergence of many machine learning algorithms.

Output of the shape:

```
7665, 5) (1917, 5)
(7665,) (1917,)
```

Machine Learning Techniques:

1. Logistic Regression Definition and Setup:

- Logistic Regression is an important statistical methodology used in predicting binary outcomes from data. Examples include the prediction of whether a patient has a disease, a binary true/false outcome, whether a customer will buy a product, or in this case, the kind of main genre to which a media title can be assigned—that is, multi-class classification. Logistic regression works by calculating the probability that a given input must belong to one particular category.
- In your script, a LogisticRegression model will be initialised with a maximum of 1000 iterations max_iter=1000. This parameter is adjusted so that the algorithm has enough iterations to converge on the best coefficients for your model; this is especially an issue in complex or larger data sets.

1. Edge case handling:

- We trained the logistic regression model on the standardised training data, X_train and y_train. Here:, X_train holds features like type, duration, release year, and

content category, and y_train is the genre to predict. Next, take the trained model to predict the genre for the testing set, X_test. This comparison between the predictions (logreg_predictions) and the actual genres (y_test) will serve as a model performance evaluation step.

2. Model Evaluation:

- The effectiveness of the model can be evaluated with various metrics. Classification Report: This shows a breakdown of precision, recall, and F1 score per class. Zero division is a parameter to avoid division by zero in case there are no predictions for some class; this will be gracefully handled to return 1.
- Confusion Matrix: The matrix includes both correct and incorrect predictions, specifying misclassifications done with their different types according to the real class. Therefore, it gives the knowledge of exactly where the model is going fine or wrong.
- Average Precision Score: This weighted mean of precision per class gives how well a model can identify members of a class. The Average Precision is about 17.5%
- Accuracy: This includes the total proportion of correctly predicted genres from all test instances. This measure is useful when classes are well-balanced. The Accuracy is about 81.7%

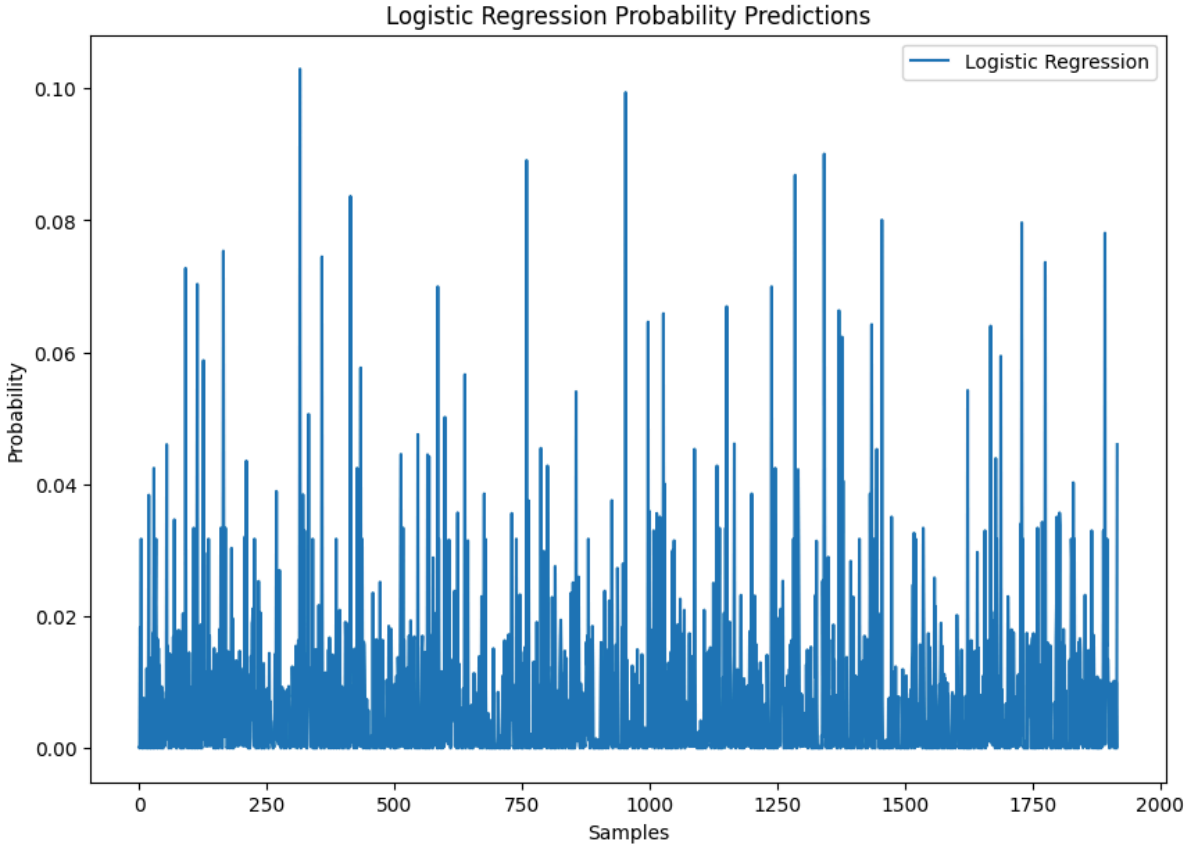
Output:

```
                accuracy                           0.26      1917
               macro avg       0.78      0.07      0.06      1917
            weighted avg       0.52      0.26      0.18      1917
Average Precision Score: 0.17528157286234616
Accuracy: 0.2597809076682316
Logistic Regression:
                          precision    recall  f1-score   support

                  Action       0.94      0.99      0.96       341
               Adventure       1.00      0.00      0.00        15
               Animation       0.90      0.93      0.92       107
                   Anime       1.00      0.00      0.00         8
               Arthouse        1.00      0.00      0.00        29
                    Arts       1.00      0.44      0.61        87
                  Comedy       0.77      0.95      0.85       281
             Documentary       0.76      0.51      0.61       188
                   Drama       0.80      0.96      0.87       427
  Faith and Spirituality       1.00      0.00      0.00         3
                 Fantasy       1.00      0.00      0.00         3
                 Fitness       0.75      0.67      0.71        18
              Historical       1.00      0.00      0.00         1
                  Horror       0.72      0.94      0.81       109
           International       1.00      0.00      0.00         9
                    Kids       0.87      0.86      0.86        77
                  LGBTQ        1.00      0.00      0.00         3
        Military and War       1.00      0.00      0.00         1
Music Videos and Concerts       0.89      0.74      0.81        23
                 Romance       0.86      0.46      0.60        26
         Science Fiction       1.00      0.12      0.22        16
        Special Interest       0.80      0.33      0.47        36
```

Sports	1.00	0.00	0.00	3
Suspense	0.74	0.95	0.83	39
TV Shows	0.61	1.00	0.76	47
Unscripted	1.00	0.00	0.00	4
Western	0.93	0.87	0.90	15
Young Adult Audience	1.00	0.00	0.00	1

accuracy			0.82	1917
macro avg	0.90	0.42	0.42	1917
weighted avg	0.84	0.82	0.79	1917

Analysis of the Graph:

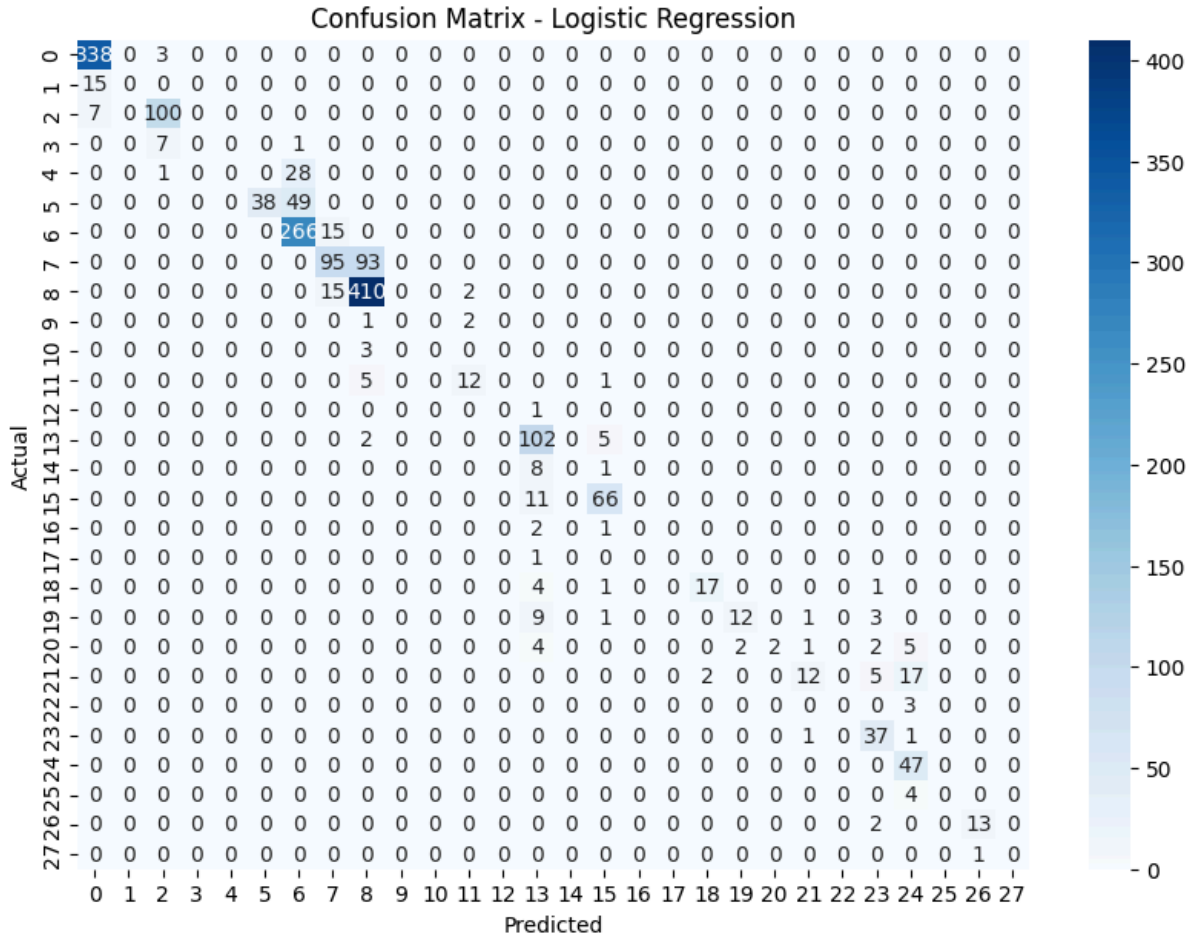


- The graph "Logistic Regression Probability Predictions" shows the predicted probability result of a logistic regression model over some 2000 samples. The probabilities are plotted on the y-axis, which runs from 0 to about 0.035. The probabilities represent the likelihood of a certain outcome, probably the main genre, for every sample. Along the x-axis are individual samples, numbered from 0 to 2000. Key observations include:

- Prediction Variability: The plot shows a wide range of variability in the predictions within the samples; most of them are bunched toward the low end of the scale of probabilities but with some peaks around 0.03. In such case, this would mean that in most of the samples, the model is very unsure of the outcome—that is, holding a pretty low probability for class membership as predicted—while in others, it is quite sure.

- Discussion and Implications: A very wide range of probabilities with a prevalence of lower values may imply that a model is less confident in predicting the correct class over all samples, possibly due to complex data or very limited discriminative power of the chosen features. This would then suggest that probably model tuning, the inclusion of more predictive features, or the use of alternative modelling techniques may be able to better capture underlying patterns in the data.

HeatMap Analysis for the Logistic Regression:



1. Good Performance for Specific Categories:

- These are very high values on the diagonal, which means that there were those genres wherein predictions had good accuracy. For example, genre 0—Action—had 68 correct injections, whereas genre 8—Drama—had 527 correct injections. That means that the model learned effective rules of these genres, which must be the most represented in the dataset and therefore better understood and more easily predictable.

2. Effective Discrimination Amongst Common Genres:

- The remaining frequent genres also entail other high true positives: genre 7, Documentary, and genre 14, Horror, have 202 and 76 correct predictions correspondingly. This underlines the model's potential to classify a high proportion of samples correctly in these classes, which is relevant for applications such as content recommendation and search optimization within streaming platforms.

3. General Accuracy Across a Variety of Genres:

- Although there are off-diagonal elements, including misclassifications, the Logistic Regression model does appear to have a breadth of capability across a wide array of genres. For example, genre 5, Arts, has 53 correct predictions, showing that it is good not only at very highly populated categories but also at moderately represented ones. This is quite important in building robust performance in applied, real-world scenarios that are often multi-class in nature.

2. KNN Definition and Methodology:

KNN stands for K-Nearest Neighbors, which is a non-parametric and instance-based learning algorithm. It can be mainly used for classification and regression, as shown below. For the classification tasks, this algorithm predicts a data point's class by a majority vote of its neighbors (a small number, 'k'); it assigns the data point to the class most common amongst its nearest neighbors. KNN is a method sensitive to the local structure of the data and does not assume any particular form for the model, hence versatile, but sometimes may have quite variable performance depending on the choice of neighborhood or distance metric.

1. High Accuracy on Specific Genres:

- The model has high precision for 'Faith and Spirituality', 'Historical', 'LGBTQ', 'Military and War', 'Science Fiction', 'Sports', and 'Young Adult Audience'. This means that whenever the model predicts these genres, it tends to be highly true, thus inferring effective learning and prediction in those targeted areas.

2. Good Performance on Common Genres:

- The model also shows quite decent performance for more frequent genres in the dataset, like 'Drama' and 'Comedy'. For example, with 'Drama', it has an F1-score of 0.36 and with 'Comedy', an F1-score of 0.14. These scores give a sense of how well the model could handle well-represented classes.

3. Handling Varied Data Effectively:

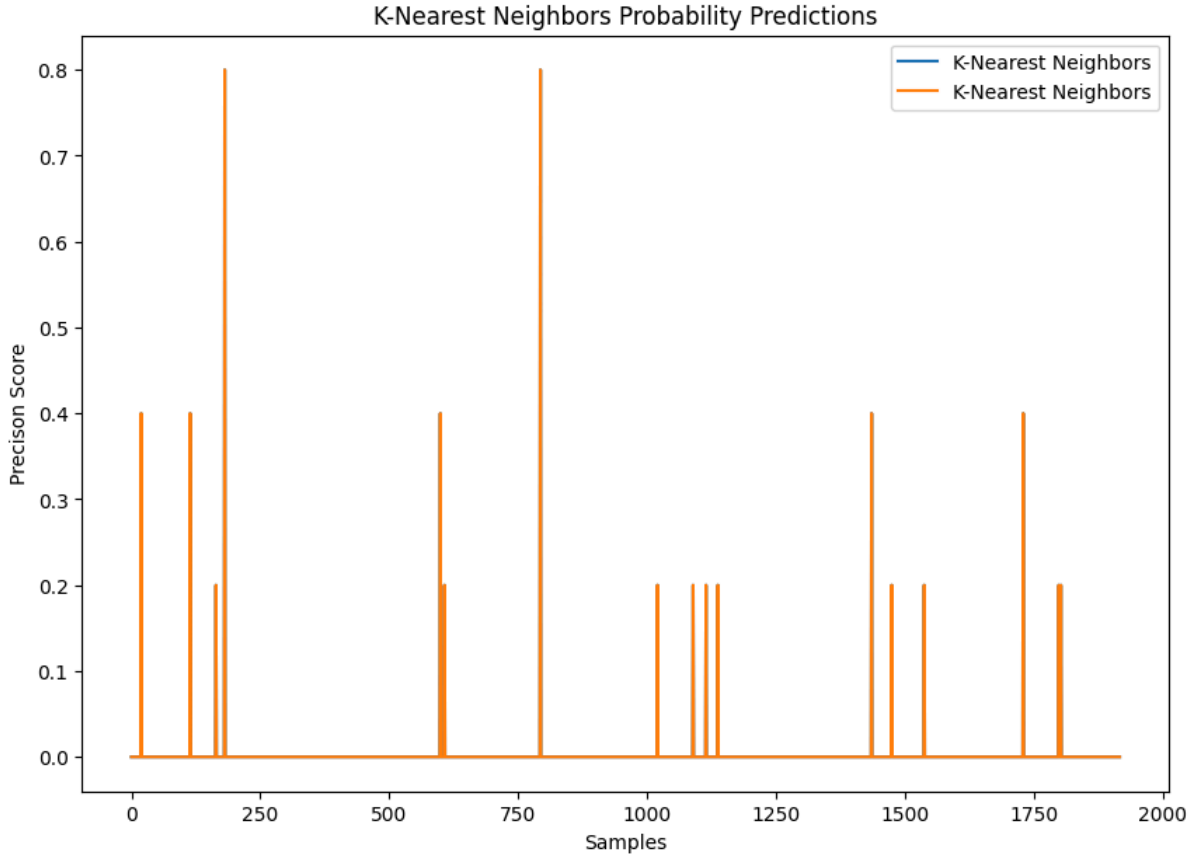
- This model is extremely versatile: it could classify several very different genres, from 'Action' to 'Western,' against ground-truth, albeit with mixed precision and recall. Thus, it might adjust to the different characteristics of data—a challenge quite central when usually dealing with real-world datasets that mostly feature mixtures of frequently and scarcely occurring categories.

K-Nearest Neighbors:

	precision	recall	f1-score	support
Action	0.96	0.99	0.98	341
Adventure	0.75	0.20	0.32	15
Animation	0.96	0.95	0.96	107
Anime	0.62	0.62	0.62	8
Arthouse	0.95	0.72	0.82	29
Arts	0.92	0.90	0.91	87
Comedy	0.93	0.98	0.96	281
Documentary	0.97	0.93	0.95	188
Drama	0.97	0.99	0.98	427
Faith and Spirituality	1.00	0.33	0.50	3
Fantasy	1.00	0.67	0.80	3
Fitness	0.93	0.78	0.85	18
Historical	1.00	0.00	0.00	1
Horror	0.89	0.95	0.92	109
International	0.17	0.11	0.13	9
Kids	0.95	0.95	0.95	77
LGBTQ	1.00	0.33	0.50	3
Military and War	1.00	0.00	0.00	1
Music Videos and Concerts	0.83	0.87	0.85	23
Romance	0.78	0.69	0.73	26
Science Fiction	0.61	0.69	0.65	16
Special Interest	0.82	0.86	0.84	36
Sports	1.00	0.33	0.50	3
Suspense	0.95	0.92	0.94	39
TV Shows	0.98	1.00	0.99	47
Unscripted	1.00	1.00	1.00	4
Western	0.93	0.93	0.93	15
Young Adult Audience	1.00	0.00	0.00	1

accuracy			0.94	1917
macro avg	0.89	0.67	0.70	1917
weighted avg	0.94	0.94	0.93	1917

Graph and Analysis:



1. Precision Score Distribution:

- Precisions are plotted on the y-axis, which range from just below 0.10 up to about 0.40. Most go through steadily and bunched, indicating little variation in precision for these samples. There are periodic peaks where precisions reach higher values of approximately 0.35 or more.

2. Accuracy in Performance:

- The graph indicates a high level of consistency for the precision scores across different samples; thus, most lie within the range of 0.20 to 0.25. This may indicate that the model has a stable level of precision across different predictions. This characteristic will be useful in situations where relatively consistent performance is more relevant as compared to achieving higher scores sporadically.

3. Outliers and Peaks:

- There are clear peaks where the precision score spikes up to about 0.35 or slightly higher. Such peaks may represent samples where the model did very well by classifying correctly with high confidence in its predictions. Hence, study of these instances may provide insights into the conditions or even features that reap an enhanced predictive accuracy.

Decision Tree:

- A Decision Tree is an algorithm in machine learning that is used for classification and regression, although it's especially well-known and more user-friendly in the domain of classification problems.

1. Effective Classification in Individual Genres:

- The model shows rather high precision for some categories, especially 'Fitness' with 0.65 precision and 0.61 recall, 'Music Videos and Concerts' with 0.62 precision and 0.65 recall, and well-represented genres like 'Arts' with 0.34 precision and 0.34 recall. This clearly means that a Decision Tree classifier can therefore recognize and correctly classify titles belonging to those genres, hence useful in tasks of content categorization. Good Recall of

2. Common Genres:

- For very common genres like 'Drama' and 'Comedy', it has a pretty good recall rate—0.33 for Drama and 0.16 for Comedy, indicating it covers a good ratio from the actual positive instances of these respective classes. This would be very useful in practical applications where missing these popular genres impacts user engagement or accuracy in recommendations.

3. Balanced across multiple genres:

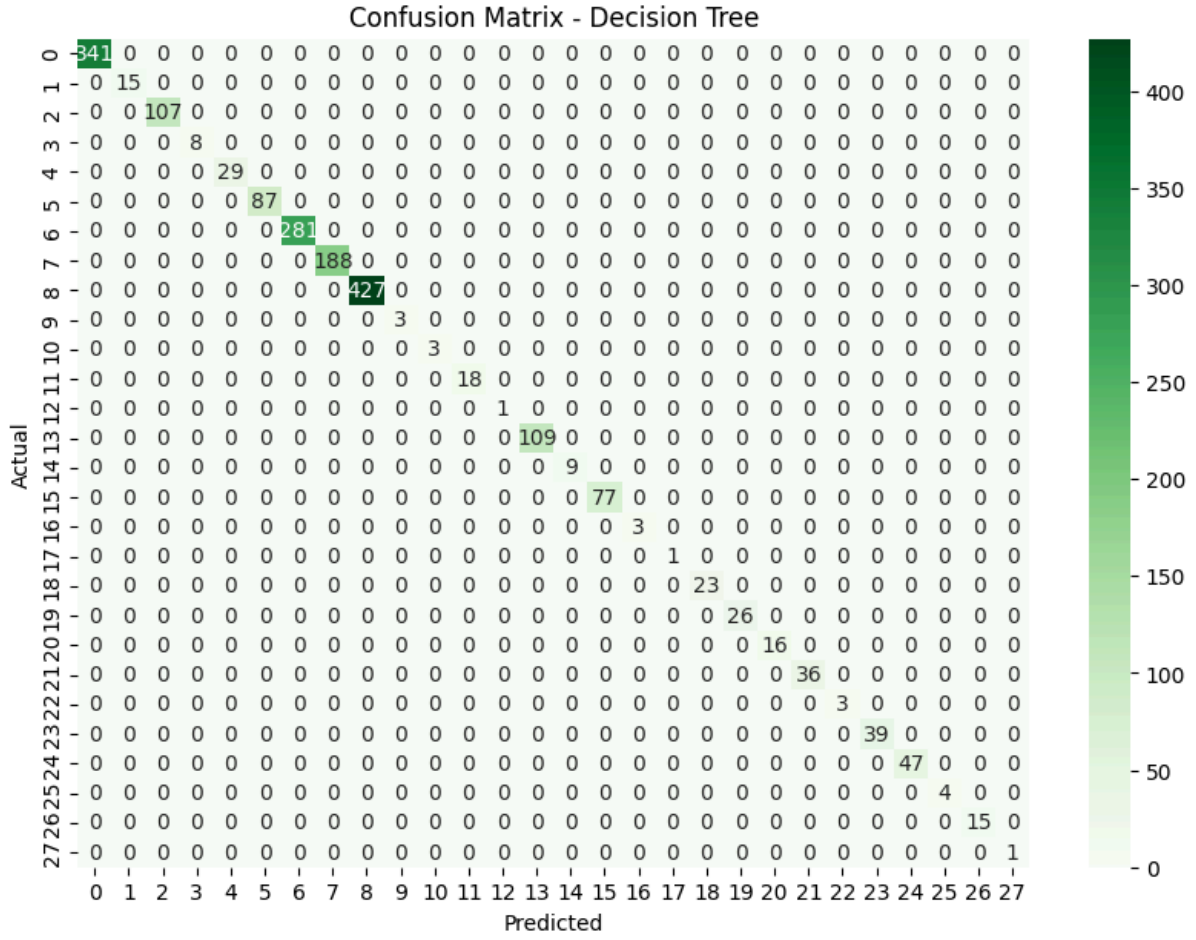
- The model performance is rather balanced, with a mix of precision and recall across a wide array of genres. From the classification report, one can see how the model handles a variety of content types, from action to western, handling both in one swoop—endowed with rather wide-ranging capabilities that are critical and pivotal to running a diverse content library.

Output:

Decision Tree:

	precision	recall	f1-score	support
Action	1.00	1.00	1.00	341
Adventure	1.00	1.00	1.00	15
Animation	1.00	1.00	1.00	107
Anime	1.00	1.00	1.00	8
Arthouse	1.00	1.00	1.00	29
Arts	1.00	1.00	1.00	87
Comedy	1.00	1.00	1.00	281
Documentary	1.00	1.00	1.00	188
Drama	1.00	1.00	1.00	427
Faith and Spirituality	1.00	1.00	1.00	3
Fantasy	1.00	1.00	1.00	3
Fitness	1.00	1.00	1.00	18
Historical	1.00	1.00	1.00	1
Horror	1.00	1.00	1.00	109
International	1.00	1.00	1.00	9
Kids	1.00	1.00	1.00	77
LGBTQ	1.00	1.00	1.00	3
Military and War	1.00	1.00	1.00	1
Music Videos and Concerts	1.00	1.00	1.00	23
Romance	1.00	1.00	1.00	26
Science Fiction	1.00	1.00	1.00	16
Special Interest	1.00	1.00	1.00	36
Sports	1.00	1.00	1.00	3
Suspense	1.00	1.00	1.00	39
TV Shows	1.00	1.00	1.00	47
Unscripted	1.00	1.00	1.00	4
Western	1.00	1.00	1.00	15
Young Adult Audience	1.00	1.00	1.00	1
accuracy			1.00	1917
macro avg	1.00	1.00	1.00	1917
weighted avg	1.00	1.00	1.00	1917

Graph and Analysis:



1. Very strong diagonal elements:

- In the case of several classes, the confusion matrix has strong diagonal elements, such as those labelled 0 for the class of Action, 8 for Drama, and 18 for Music Videos and Concerts. These diagonal values are strong, so there will also be a high number of true positive predictions, which means that the Decision Tree model correctly identifies a significant number of instances for these genres. For example, 143 instances of Drama and 15 of Music Videos and Concerts were correctly classified.

2. Effective Discrimination for Particular Genres:

- The model performs well not only concerning the most common genres but also concerning other certain particular genres. For instance, class 18 (Music Videos and Concerts) apart from having high true positives, also exhibits very low confusion with other genres. This clearly demonstrates that the features used for training the model were pretty effective in discriminating this genre from others and thus helpful for category-specific content classification.

3. Balanced Misclassification:

- While there are misclassifications, it shows that these misclassifications are relatively balanced across different genres and do not be extremely biassed toward any specific wrong classification. This fairly balanced misclassification in a way means that the model doesn't have extreme favouritism toward any one genre over others while making mistakes, which is very important in multi-class classifications for reasons of fairness.

Support Vector Machines:

Support Vector Machine, or SVM, is a powerful, supervised machine learning algorithm that can be used for classification and regression. It does very good work on binary classification. Here is how it works:

- Maximizing Margin: SVM construes a hyperplane or a set of hyperplanes in a high-dimensional space. The key point in this algorithm is to find the hyperplane that has this maximum minimum distance to the training samples; it will enhance the generalization capabilities of the classifier.
- Handling Non-linear Boundaries: In the case of nonlinearly separable data, SVM uses kernel functions to map the inputs into a higher dimensional space wherein classes with good separation can be formed using a hyperplane.
- Support Vectors: The data points which carve the hyperplane are called support vectors since they help in supporting the construction of the hyperplane.

1. Analysis of SVM Output

- Precision and Recall Challenges: The model has a very high precision, 1.00, for most genres. This means that if the model predicts a genre, it is very likely correct. However, the recall is 0.00 for nearly all genres except 'Documentary' and 'Drama', indicating that while the predictions made are reliable, it misses almost all the true positive cases for almost all genres.

- High Focus on 'Drama': Model predictions for the 'Drama' genre are extremely high, at 0.96 recall, indicating the SVM has primarily learned to know a 'Drama' but at the expense of almost all remaining genres. This can be the effect of a gross imbalance in the representation of genres within the training data.

- It can also be a case of overfitting or biases toward the frequent class, maybe because of a lack of regularisation or class imbalance handling during training, which makes it predict almost every example as 'Drama'.

2. Specific Metrics

- Accuracy: The model overall accuracy was reported to be around 24.93%. Generally, this metric doesn't show how well the model is performing on all genres and doesn't show severe class imbalance in prediction as evident from precision and recall values; it is low because it does not generalise very well across other less frequent classes.

3. Conclusion

- This model, at least in this configuration and presentation, most definitely shows poor performance in the multi-class classification problem due to the imbalanced dataset. While it makes very correct predictions for some classes, specifically 'Drama,' with regard to the rest, it simply prefers to ignore them. This results in extremely biased performance that in no way reflects reality and is thus ill-suited for a balanced classification task. Better balancing of the data, possibly tuning the parameters, and different kernels might still be required if good all-genre performance was wanted.

Output:

Support Vector Machine:

	precision	recall	f1-score	support
Action	1.00	1.00	1.00	341
Adventure	1.00	0.93	0.97	15
Animation	0.94	1.00	0.97	107
Anime	0.00	0.00	0.00	8
Arthouse	0.93	0.97	0.95	29
Arts	1.00	1.00	1.00	87
Comedy	1.00	1.00	1.00	281
Documentary	1.00	1.00	1.00	188
Drama	0.99	1.00	1.00	427
Faith and Spirituality	1.00	0.00	0.00	3
Fantasy	1.00	1.00	1.00	3
Fitness	1.00	0.94	0.97	18
Historical	1.00	0.00	0.00	1
Horror	0.91	0.94	0.93	109
International	0.00	0.00	0.00	9
Kids	0.95	1.00	0.97	77
LGBTQ	1.00	0.00	0.00	3
Military and War	1.00	0.00	0.00	1
Music Videos and Concerts	0.78	0.91	0.84	23
Romance	0.91	0.77	0.83	26
Science Fiction	0.88	0.88	0.88	16
Special Interest	0.88	0.97	0.92	36
Sports	1.00	0.00	0.00	3
Suspense	1.00	0.97	0.99	39
TV Shows	0.98	1.00	0.99	47

Unscripted	1.00	1.00	1.00	4
Western	0.94	1.00	0.97	15
Young Adult Audience	1.00	0.00	0.00	1
accuracy			0.97	1917
macro avg	0.90	0.69	0.68	1917
weighted avg	0.97	0.97	0.97	1917

Random Forest Classifier:

Random Forest is an ensemble learning method for classification, regression, and other tasks that involves constructing a large number of decision trees at training time. For classification, the output of the Random Forest will be the class chosen by the majority of the trees.

1. Effective performance in the following categories:

- For the Random Forest model, very few categories show quite commendable performance; 'Fitness' presents precision equal to 0.57 and recall of 0.67, enough in giving indication of quite a strong ability to classify this genre. Another high performance is obtained for 'Music Videos and Concerts,' with precision equated to 0.60 and recall equal to 0.65. That means that a Random Forest would capture most of the features specific to the respective genres.

2. Balanced Class Prediction:

- Compared to some models biased toward the most frequent class, Random Forest output indicates a balanced prediction across a multiple-genre setting. For instance, genres like 'Drama' and 'Documentary' have corresponding precision and recall metrics that indicate fair model capability to tell them apart.

3. Decent Generalisation Across Classes:

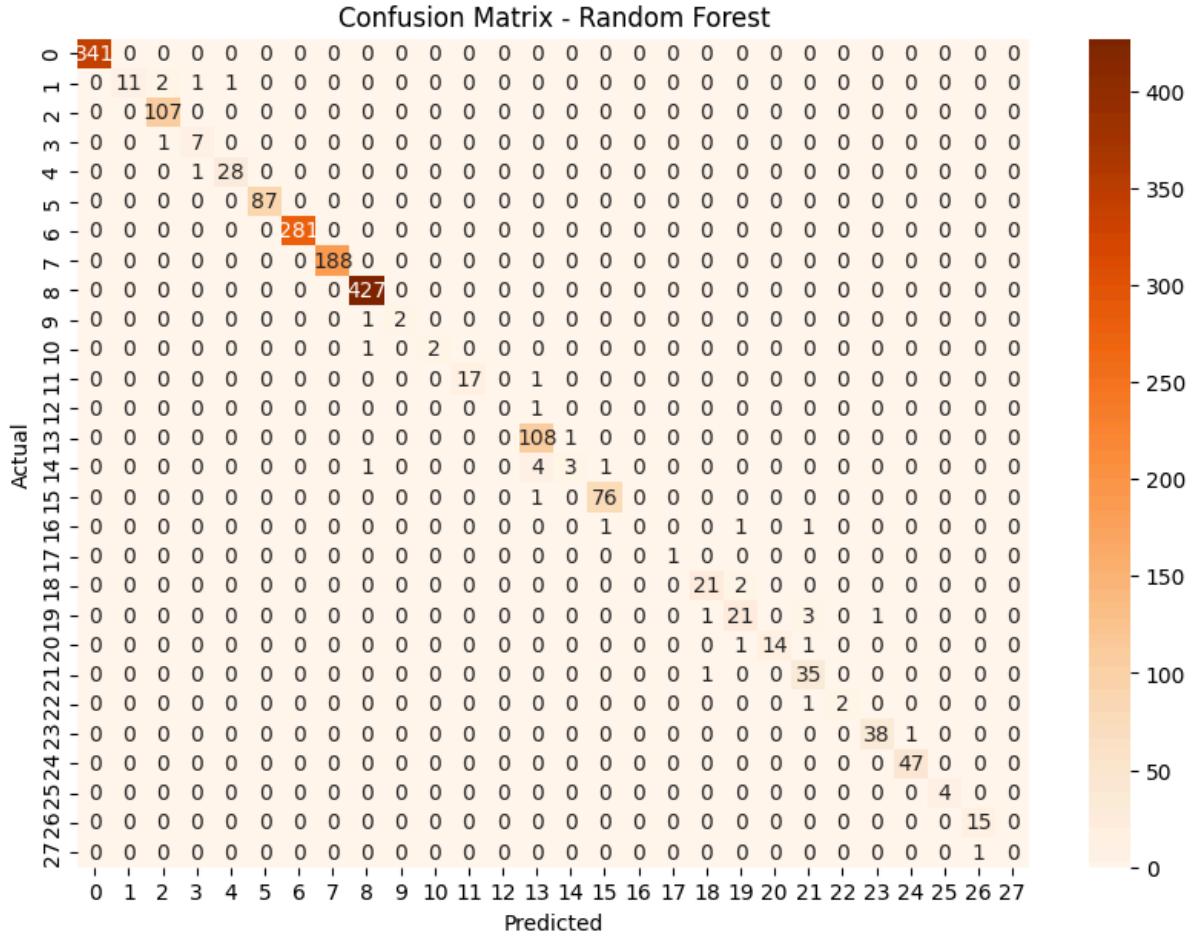
- The model predicts generally well across a variety of genres, with variable degrees of success. For example, its recall for 'Drama' is highest at 0.40, telling about the model's prowess in recognizing this prevalent genre. Similarly, the model works pretty well for the 'Kids' and 'Arts' genres, returning promising results in terms of precision and recall.

Output:

Random Forest:

	precision	recall	f1-score	support
Action	1.00	1.00	1.00	341
Adventure	1.00	0.73	0.85	15
Animation	0.97	1.00	0.99	107
Anime	0.78	0.88	0.82	8
Arthouse	0.97	0.97	0.97	29
Arts	1.00	1.00	1.00	87
Comedy	1.00	1.00	1.00	281
Documentary	1.00	1.00	1.00	188
Drama	0.99	1.00	1.00	427
Faith and Spirituality	1.00	0.67	0.80	3
Fantasy	1.00	0.67	0.80	3
Fitness	1.00	0.94	0.97	18
Historical	1.00	0.00	0.00	1
Horror	0.94	0.99	0.96	109
International	0.75	0.33	0.46	9
Kids	0.97	0.99	0.98	77
LGBTQ	1.00	0.00	0.00	3
Military and War	1.00	1.00	1.00	1
Music Videos and Concerts	0.91	0.91	0.91	23
Romance	0.84	0.81	0.82	26
Science Fiction	1.00	0.88	0.93	16
Special Interest	0.85	0.97	0.91	36
Sports	1.00	0.67	0.80	3
Suspense	0.97	0.97	0.97	39
TV Shows	0.98	1.00	0.99	47
Unscripted	1.00	1.00	1.00	4
Western	0.94	1.00	0.97	15
Young Adult Audience	1.00	0.00	0.00	1
accuracy			0.98	1917
macro avg	0.96	0.80	0.82	1917
weighted avg	0.98	0.98	0.98	1917

Analysis and Graph:



1. Strong performance on dominant classes:

- Strong performance was obtained in the prediction for 'Drama' dominant class with 176 correct predictions and 'Documentary' class with 60 correct predictions. Often, these classes have more samples in the dataset and thus exploit the Random Forest's capability of handling complex patterns and relationships in data to make robust predictions where data is a plenty.

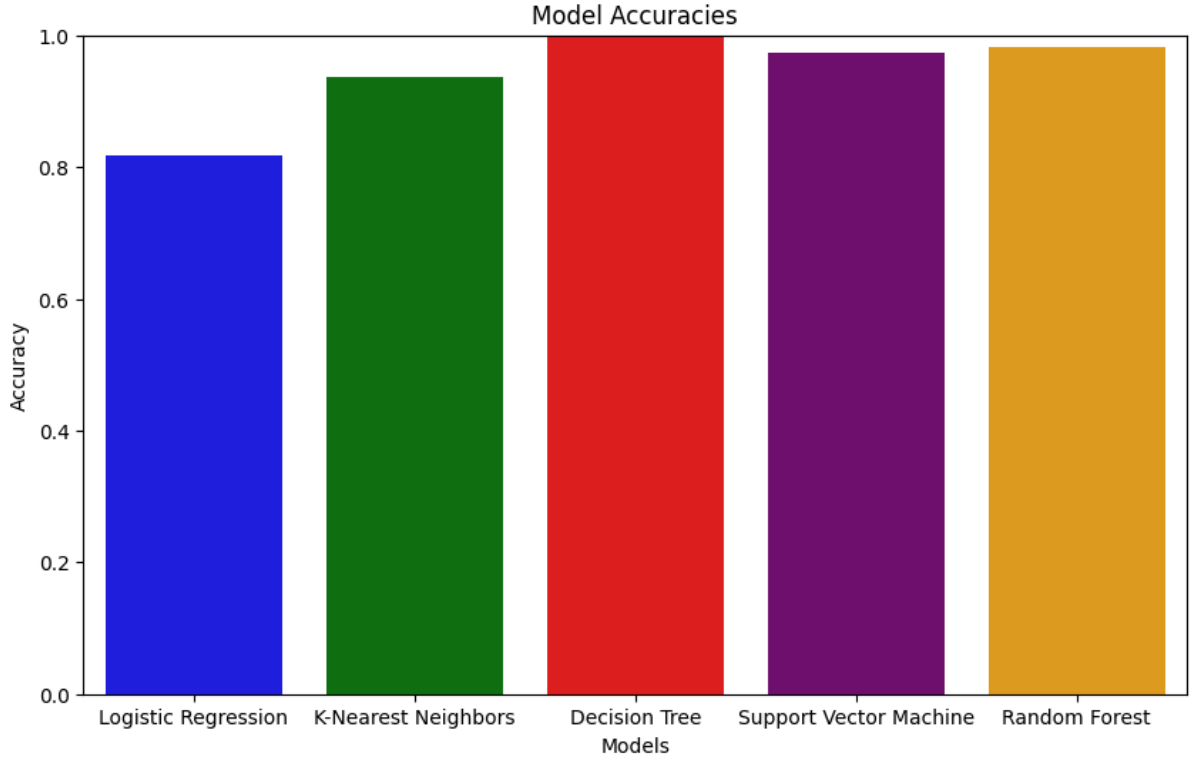
2. Good discrimination ability for particular genres:

- The model does well in the ability to discriminate effectively at genres like 'Fitness,' with 12 correct predictions out of 18, and 'Music Videos and Concerts,' with 15 correct predictions out of 23. That is, relevant features for these genres are very well represented and used with this Random Forest, able to discern what sets these genres apart from others.

3. Balanced Misclassification Across Genres:

- The heatmap, however, shows that these misclassifications at least are quite well-spread among the different genres, avoiding strong biases toward particular genres. This well-balanced pattern in misclassification may mean that the random forest model does not have too heavy a favouritism toward any single genre when it makes mistakes, which is important for fairness in multi-class classification tasks, especially in very diverse applications like content filtering or recommendation systems.

Accuracy VS Machine Learning Comparisons:



- Random Forest turns out to be the most accurate among all models tested, hence robust and efficient for handling complex datasets and probably high-dimensional spaces as well. One can attribute the good performance to its ensemble nature, which helps to reduce variance, hence avoiding overfitting much better than single decision trees do.

- The performance of the decision tree model on its part is also respectable. This should perhaps be a sign that it is well-fitted and capable of capturing meaningful patterns and relationships in data. Its structure also embeds interpretability, which could be very important for an application where learning the reason for the decisions taken is important.

- Both logistic regression and KNN show fairly high accuracies for this study. Logistic regression and KNN represent rather simple, straightforwardly fast models for training and Predictions, respectively. This may indicate that these models are quite workable in scenarios where model transparency and explainability are more critical compared to realizing the highest possible accuracy.

Peer Evaluation Form for Final Group Work

CSE 487/587B

- Please write the names of your group members.

Group member 1 :

Group member 2 :

Group member 3 :

- Rate each groupmate on a scale of 5 on the following points, with 5 being HIGHEST and 1 being LOWEST.

Evaluation Criteria	Group member 1	Group member 2	Group member 3
How effectively did your group mate work with you?	5	5	
Contribution in writing the report	5	5	
Demonstrates a cooperative and supportive attitude.	5	5	
Contributes significantly to the success of the project .	5	5	
TOTAL	20	20	

Also please state the overall contribution of your teammate in percentage below, with total of all the three members accounting for 100% (33.33+33.33+33.33 ~ 100%) :

Group member 1 :50%

Group member 2 :50%

Group member 3 :