Abstract - The main goal and contribution of this paper is to analyze Netflix Dataset and classify whether a title is a show or a movie using two machine learning models Random Forrest Classifier and Logistic Regression. The dataset was taken from Kaggle [1], it initially has 12 features, a few of which influence the type of content it and 8807 samples. Data Preprocessing and exploratory data analysis of the dataset are described in this study. Following this, two supervised learning methods were used , Random Forest method used for classification, and Logistic Regression is a statistical model that can be used for binary classification, it was found both methods were quite suitable in terms of accuracy of classification.

Introduction

Netflix has established itself as one of the largest streaming services globally, boasting over 282.7 million subscribers worldwide as of late 2024 [2]. With a vast library of movies, TV shows, and original content, Netflix relies on sophisticated content classification systems to enhance user experience and personalize recommendations [3]. This study focuses on the classification of Netflix content—whether a title is a Movie or TV Show—using machine learning models. By applying supervised learning techniques, this research aims to identify key features that influence content classification, such as release year, duration, genre, country, and rating. Understanding these factors can improve search optimization, recommendation systems, and content analytics for streaming platforms.

This study implements a comprehensive machine learning pipeline to classify Netflix titles as either Movies or TV Shows. The process includes data preprocessing, feature engineering, exploratory data analysis (EDA) using Python libraries such as NumPy, Pandas, Matplotlib, and Seaborn, feature selection through Chi-Square tests and Random Forest feature importance, and the training and evaluation of two supervised classification models: Random Forest Classifier and Logistic Regression. The models' performance is assessed based on accuracy, precision, recall, F1-score, and generalization ability, utilizing cross-validation and hyperparameter tuning to identify the most effective classification approach and to gain insight into the features that most significantly distinguish between Movies and TV Shows.

Literature Review

Netflix and machine learning-based classification are closely interlinked topics, extensively covered in academic papers and industry research. This section explores relevant studies in the field. A 2024 Netflix study on content recommendations achieved 96.25% accuracy using Random Forest after feature selection [4]. Similarly, in the broader streaming industry, Random Forest achieved 87% accuracy when

classifying content types based on a mix of numeric and categorical features [5]. These highlight the effectiveness of Random Forest in this domain.

While Logistic Regression (LR) is widely used for binary classification, studies suggest that its performance can be inferior to ensemble models like Random Forest. For instance, a 2022 study on movie classification found that Logistic Regression achieved 65% accuracy, whereas Random Forest reached 74% accuracy [6]. However, Logistic Regression performs well when data exhibits a clear linear boundary, such as duration thresholds where movies tend to be longer than TV shows.

Despite extensive research on recommendation systems and regression-based streaming analytics, there is a lack of studies applying Random Forest and Logistic Regression to the binary classification of Movies vs. TV Shows. This project aims to fill this gap by applying RF and LR for direct classification of Netflix content, rather than focusing on user preference.

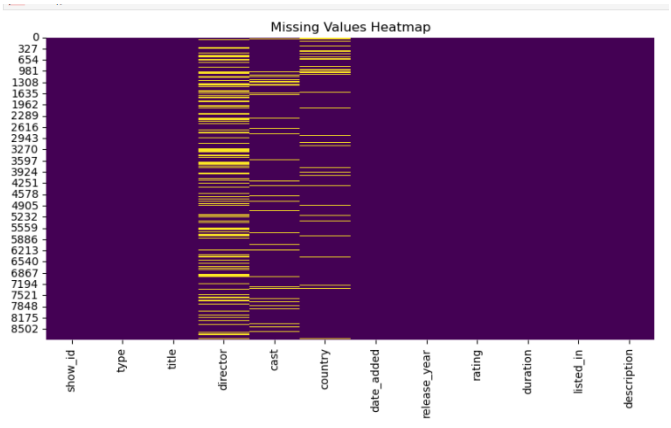
Data Exploration And Preprocessing

The dataset used in this study was obtained from Kaggle and contains 8,807 records of Netflix titles with 12 features such as title, type (Movie or TV Show), director, cast, country, release year, content rating, genre, and description (see figure on the right). The primary goal of this section is to explore the dataset, analyse key trends, and preprocess the data for classification. The Python programming language and relevant libraries such as NumPy, Pandas, Matplotlib, and Seaborn were used for data analysis and visualization.

```
<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
```

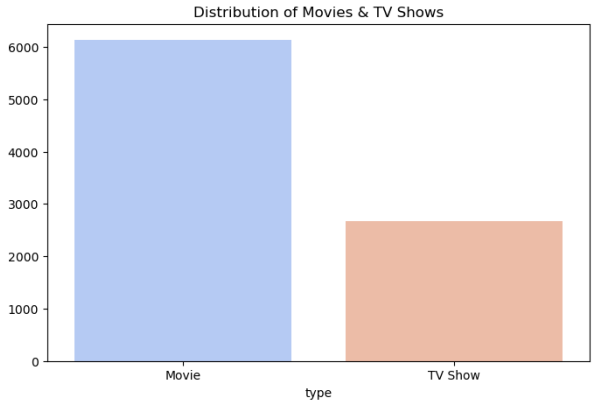
Missing Data Visualization

A missing values heatmap (pictured on the right) was generated to visualize gaps in the dataset. Some key findings include:Director and Cast had a large number of missing values, making them unreliable for classification.Also,Date Added had some missing values (only 10) and Country had 831 missing values.



Handling Strategy: Some columns with missing values were dropped, while others were imputed (covered in next section).

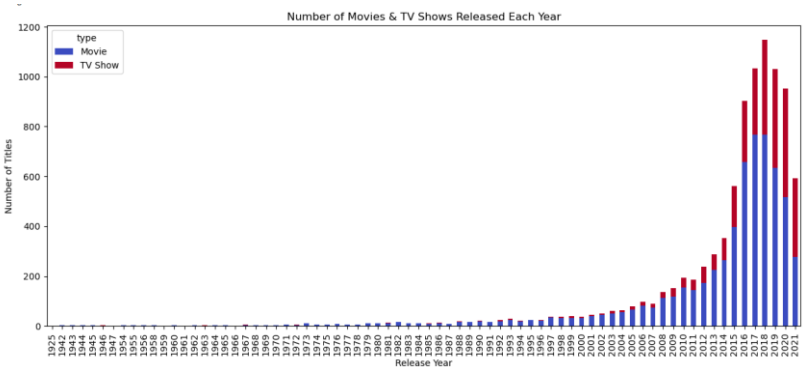
To understand patterns in the data, various visualizations were generated:



Distribution of Netflix Content:

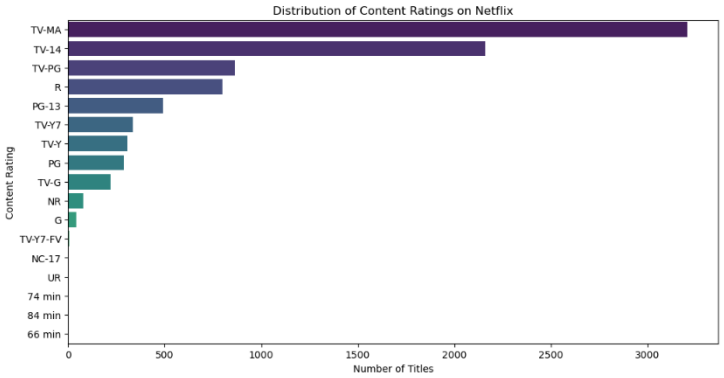
Netflix has a significantly higher number of movies (~6,000) compared to TV shows (~3,000) (see graph).This class imbalance is accounted for in model evaluation to prevent bias in classification.

Movies & TV Shows Released Over Time:



The majority of content was released after 2010, likely because of Netflix's doing more original productions.

Distribution of Content Ratings:



TV Shows tend to have stricter ratings (e.g., TV-MA, TV-14).WherasMovies have a broader range (e.g., G, PG, R, NC-17).This feature is expected to be useful in distinguishing between a movie and a TV show.

Data Preprocessing

To prepare the dataset for machine learning, several data cleaning steps were performed. Irrelevant columns such as Title, Director, Cast, and Description were dropped, as they contain unstructured text data unsuitable for classification without Natural Language Processing. Show_ID was also removed since it is a unique identifier and does not contribute to pattern recognition. The duration feature required transformation, as movies are recorded in minutes while TV shows are listed in seasons. To standardize this, seasons were converted into minutes using the assumption that *one season equals ten 60-minute episodes (600 minutes)*.

Handling missing values was also necessary. Numerical attributes like *release year* were filled with the median to prevent data loss, while *country* missing values were replaced with the most common country in Netflix's catalogue. Lastly, *rating* was imputed using the mode.

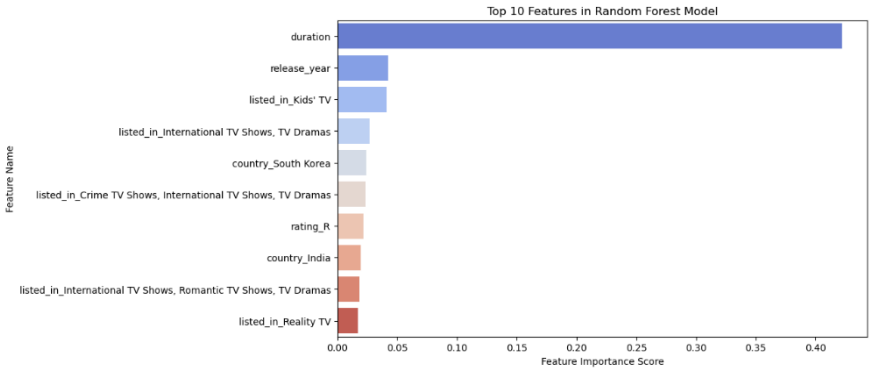
Feature Selection and Model Training

Top Features by Chi-Square Test:

	Feature	Chi2 Score	p-value
1185	listed_in_Kids' TV	504.043348	1.253794e-111
7	rating_R	343.017633	1.405052e-76
1175	listed_in_International TV Shows, TV Dramas	277.223842	3.023986e-62
1007	listed_in_Crime TV Shows, International TV Sho...	252.021674	9.412466e-57
266	country_India	227.641640	1.948376e-51
1196	listed_in_Kids' TV, TV Comedies	226.819507	2.944257e-51
450	country_South Korea	225.994373	4.455905e-51
1209	listed_in_Reality TV	217.655082	2.936853e-49
1156	listed_in_International TV Shows, Romantic TV ...	215.363976	9.282536e-49
6	rating_PG-13	213.870494	1.965484e-48

To determine the most relevant features, two feature selection techniques were applied: The Chi-Square test is a statistical method used to determine whether there is a significant association between categorical variables. It is commonly applied in feature selection to assess whether specific categorical features contribute meaningfully to a classification problem. In our dataset, we applied the Chi-Square test to evaluate the importance of categorical variables such as content rating, country, and genre ("listed_in") in distinguishing between movies and TV shows(cite table above). The higher the score the more likely it was to be used.The most significant features included: listed_in_Kids' TV (Chi2 Score: 504.04) and rating_R (Chi2 Score: 343.01), These features were retained for modeling due to their strong association with the target variable.

Feature Importance from Random Forest



This step was performed after training the Random Forest classifier (covered later in the paper) but is presented here for clarity. One advantage of Random Forest is its ability to provide feature importance analysis, which helps identify key predictors [7].

The top-ranked features based on the model's decision-making were duration and release year, with duration being the most dominant, having an importance score of 0.4221. This analysis confirms the expected hypothesis: duration is a highly influential factor in distinguishing between a movie and a TV show, as movies typically have a fixed runtime, whereas TV shows consist of multiple episodes spanning multiple seasons.

After selecting the final features, the target variable ('type') was converted into binary format (0 = Movie, 1 = TV Show). One-hot encoding was then applied to categorical variables—Country, Rating, and Listed_in (Genres)—to ensure they were numerically represented for machine learning models.

Next, the dataset was split into 80% training and 20% testing using stratified sampling, preserving the original class distribution. Since Logistic Regression requires feature scaling, Standardization was applied using Sklearn's StandardScaler() to ensure consistent numerical ranges.

The final feature set (X) included Release Year, Duration, One-Hot Encoded Countries, Ratings, and Genres, with 'type' as the target variable for classification.

Model Training and Evaluation

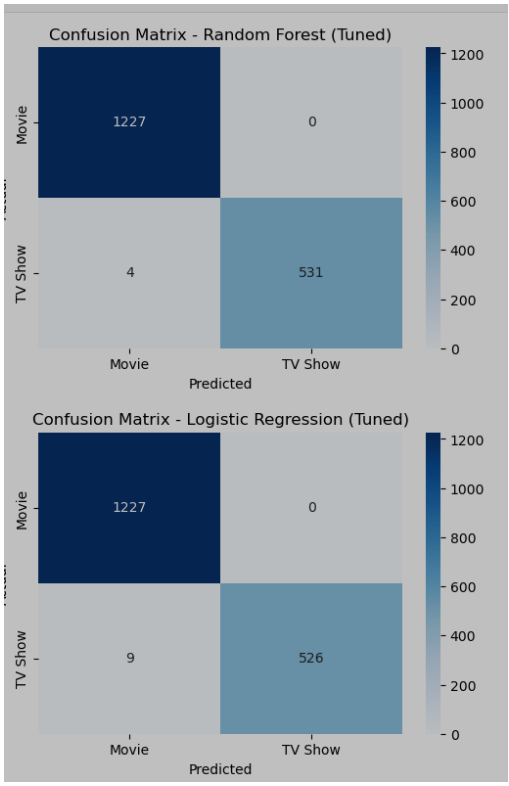
To classify whether a Netflix title is a Movie or TV Show, two supervised machine learning models were implemented: Random Forest Classifier (RF) and Logistic Regression (LR). Random Forest Classifier employs ensemble learning to aggregate predictions from multiple decision trees, in turn enhancing accuracy. Its suitability for Netflix content classification stems from handling mixed data types (e.g., numerical *duration* and categorical *genre*) and capturing complex feature interactions [8].Logistic Regression estimates class probabilities via linear feature combinations, offering interpretable coefficients (e.g., quantifying *release year*'s impact). Its simplicity makes it ideal for baseline comparisons in binary tasks like movie/TV classification[9].

Implementation and Hyperparameter Tuning

Metric	Random Forest	Tuned RF	Logistic Regression	Tuned LR
Accuracy	1	0.9958	0.9925	0.9990
Precision	1	1	0.99	0.99
Recall	1	0.95	0.99	0.99
F1 Score	1	0.97	0.99	0.99
CV accuracy	0.9954	0.9985	0.9913	0.9990

Using Sklearn, both Logistic Regression and Random Forest models were trained and performed well. A range of metrics was obtained from the default and fine-tuned models, along with cross validation scores. The table above provides a detailed breakdown of these results. The Random Forest model initially achieved an accuracy of 100%, indicating possible overfitting, while Logistic Regression achieved 99.25%. To optimize performance, GridSearchCV was applied to both models for hyperparameter tuning. After fine-tuning, Random Forest accuracy decreased slightly to 99.58%, likely due to a reduction in overfitting, while Logistic Regression improved to 99.90%. To evaluate generalization to unseen data, cross-validation (CV) scores were computed. Results show that Random Forest maintained a CV accuracy of approximately 99.85%, whereas Logistic Regression's CV accuracy improved from 99.13% to 99.90% after tuning.

Confusion Matrix



To analyze the models' performance in distinguishing Movies vs. TV Shows, confusion matrices were made for both tuned models (see figure on the right). Both models correctly classify 1,227 movies. Logistic Regression performs slightly worse in correctly predicting TV Shows, with 526 correct classifications compared to Random Forest's 531. Finally, Random Forest misclassifies 4 TV Shows as Movies, whereas Logistic Regression misclassifies 9.

Model Selection

Based on the evaluation metrics, Tuned Random Forest seems to be the most suitable classification task. The key reasons for this is the higher recalls as it is slightly better at identifying TV Shows correctly, minimizing false negatives. Also with a Higher F1-Score, the balance between precision and recall is superior in Random Forest. Although Logistic Regression performed well, its slightly lower recall and F1-score indicate that it may miss more TV Show classifications compared to Random Forest. Therefore, the final model used for classification is the Tuned Random Forest model.

Conclusion and Future Work

This study successfully applied Random Forest and Logistic Regression models to classify Netflix titles as either Movies or TV Shows. The Tuned Random Forest model demonstrated the best performance, achieving high accuracy, recall, and F1-score. However, the model's exceptionally high-performance metrics suggest overfitting, which could limit its generalization to new data.

Additionally, the assumption that 1 season equals 600 minutes (ten 60-minute episodes) was used to normalize the duration feature. While this was a quick solution to handle making that category uniform, TV show duration is variable. In terms of future work, I could explore more flexible conversion methods, such as analyzing the actual runtime of different TV series which then gives more accurate feature engineering. Other methods that could be addressed next time would be feature reduction and regularization to improve the model generalization. Also, the use of Hybrid Models and other architectures such as KNN or Neural Network and Random Forest Ensembles would be something I would like to explore rather than evaluating only 2 models. Implementing these improvements, would greatly enhance my analysis and could contribute to the area of classification within the streaming domain.

References

[1] Kaggle Netflix Dataset: https://www.kaggle.com/datasets/anandshaw2001/netflix-movies-and-tv-shows

[2] Netflix Subscribers Statistics 2025[Users by Country]:

https://evoca.tv/netflix-user-statistics/

[3] Netflix's Algorithm: How Does Netflix Use AI to Personalize Recommendations?

https://litslink.com/blog/all-about-netflix-artificial-intelligence-the-truth-behind-personalized-content

[4] ([PDF] Classification of Movie Recommendation on Netflix Using Random Forest Algorithm(**Alifia Salwa Salsabila**,**Atika Sari**, **Eko Hari Rachmawanto)**

https://www.researchgate.net/publication/382630304_Classification_of_Movie_Recommendation_on_Netflix_Using_Random_Forest_Algorithm

[5] **Predicting Popularity of Video Streaming Services with Representation Learning: A Survey and a Real-World Case Study(Sidney Loyola de Sá , Antonio A de A Rocha , Aline Paes):**, https://pmc.ncbi.nlm.nih.gov/articles/PMC8588537/]

[6]Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets (Kaitlin Kirasich , Trace Smith , Bivin Sadler)

https://scholar.smu.edu/cgi/viewcontent.cgi?params=/context/datasciencereview/article/1041/&path_info=Report.pdf]).

[7] Random Forest Classifiers :A Survey and Future Research Directions(Vrushali Y Kulkarni, Dr Pradeep K Sinha)

https://adiwijaya.staff.telkomuniversity.ac.id/files/2014/02/Random-Forest-Classifiers_A-Survey-and-Future.pdf

[8] *The Elements of Statistical Learning*. Springer. Hastie, T., Tibshirani, R., & Friedman, J. (2009).

[9] *Applied Logistic Regression*. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Wiley.