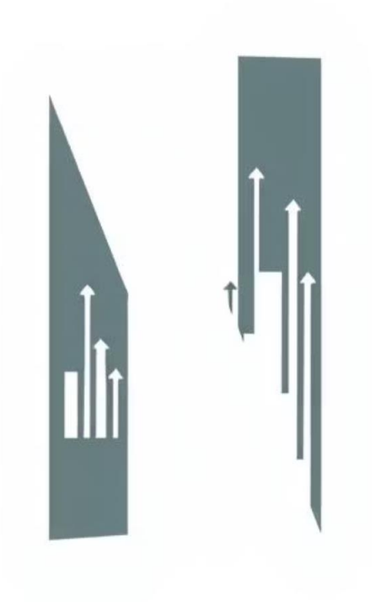
Data Mining Analysis of Netflix Movies & TV Shows

Classification • Clustering • Recommendation System

Presented by : Sultan | 56189

Course : Data Mining – Semester Project

Instructor : Sir Tajamul Shahzad



Introduction to the Project

Project Goal

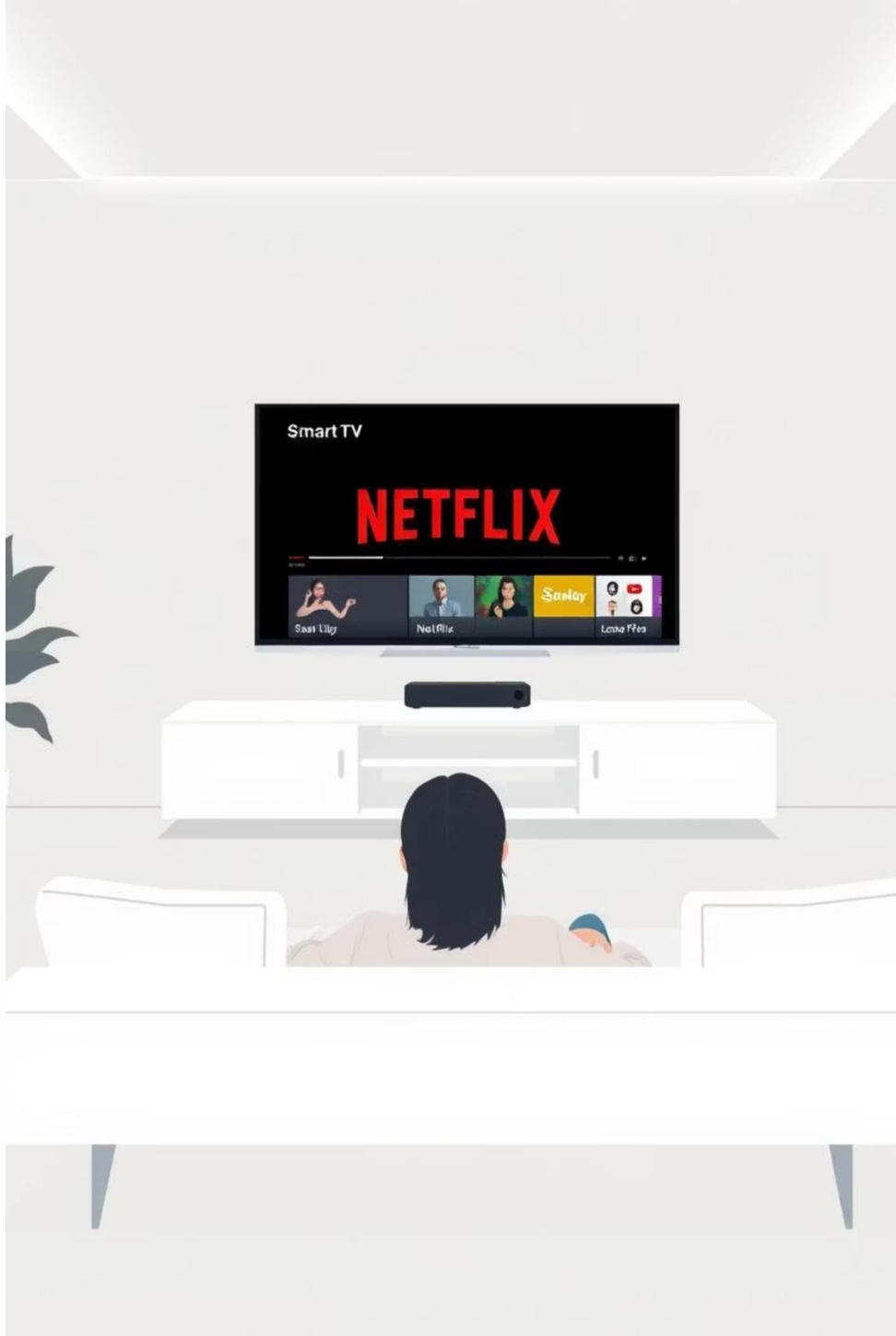
Extract insights from Netflix dataset.

Key Tasks

Content classification, clustering, recommendation system.

Dataset Origin

Kaggle: Netflix Movies & TV Shows metadata.



Understanding the Netflix Dataset

The Netflix dataset is a rich source of information, crucial for our analysis. It contains detailed metadata for each title.

- **Key Attributes:** Title, type, director, cast, country, release year, rating, duration, genres, description.
- **Analytical Potential:** Supports diverse analyses including Exploratory Data Analysis (EDA), classification, clustering, and the development of recommendation systems.
- **Scale:** Over 8,000+ titles, spanning various countries and genres, providing a comprehensive view of Netflix's content library.

CRISP-DM Workflow for Data Mining

01

Business Understanding

Improve content discovery on Netflix.

02

Data Understanding

Explore dataset structure and patterns.

03

Data Preparation

Cleaning, encoding, TF-IDF transformation.

04

Modeling

Classification, clustering, recommendations.

05

Evaluation

Assessing models: Accuracy, F1-score, Silhouette score.

06

Deployment

Saving machine learning models using Joblib.

Exploratory Data Analysis (EDA) Insights

Key observations from initial data exploration:

Content Distribution

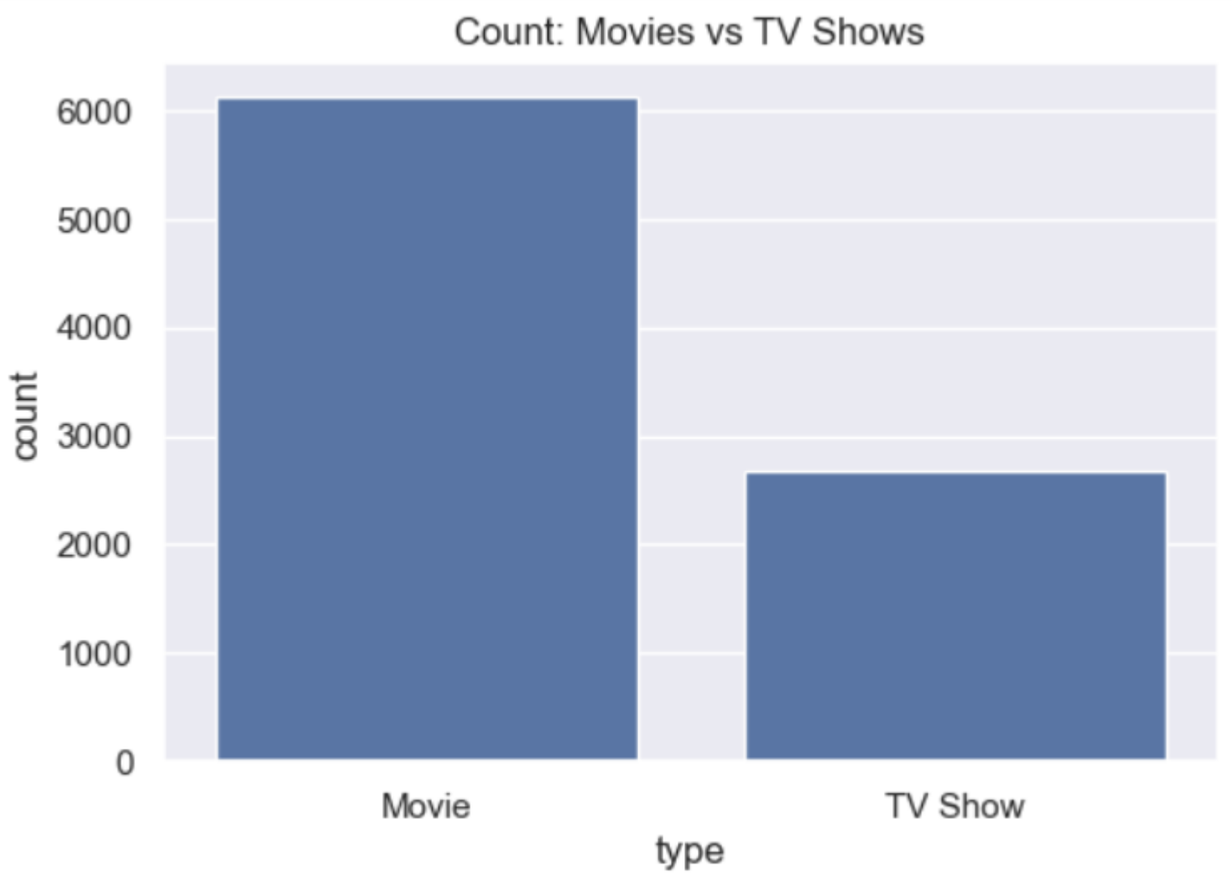
More Movies than TV Shows.

Top Content Producers

USA, India, UK lead in Netflix content production.

Dominant Genres

Drama, International Movies, Comedies are most popular.



Data Cleaning & Preparation Steps

Rigorous preparation ensures data quality for modeling:

1

Handle Missing Values

Strategically filled null entries for completeness.

2

Standardize Durations

Converted "120 min" to numeric minutes for consistency.

3

Encode Categorical Data

OneHotEncoder for features, LabelEncoder for target variables.

4

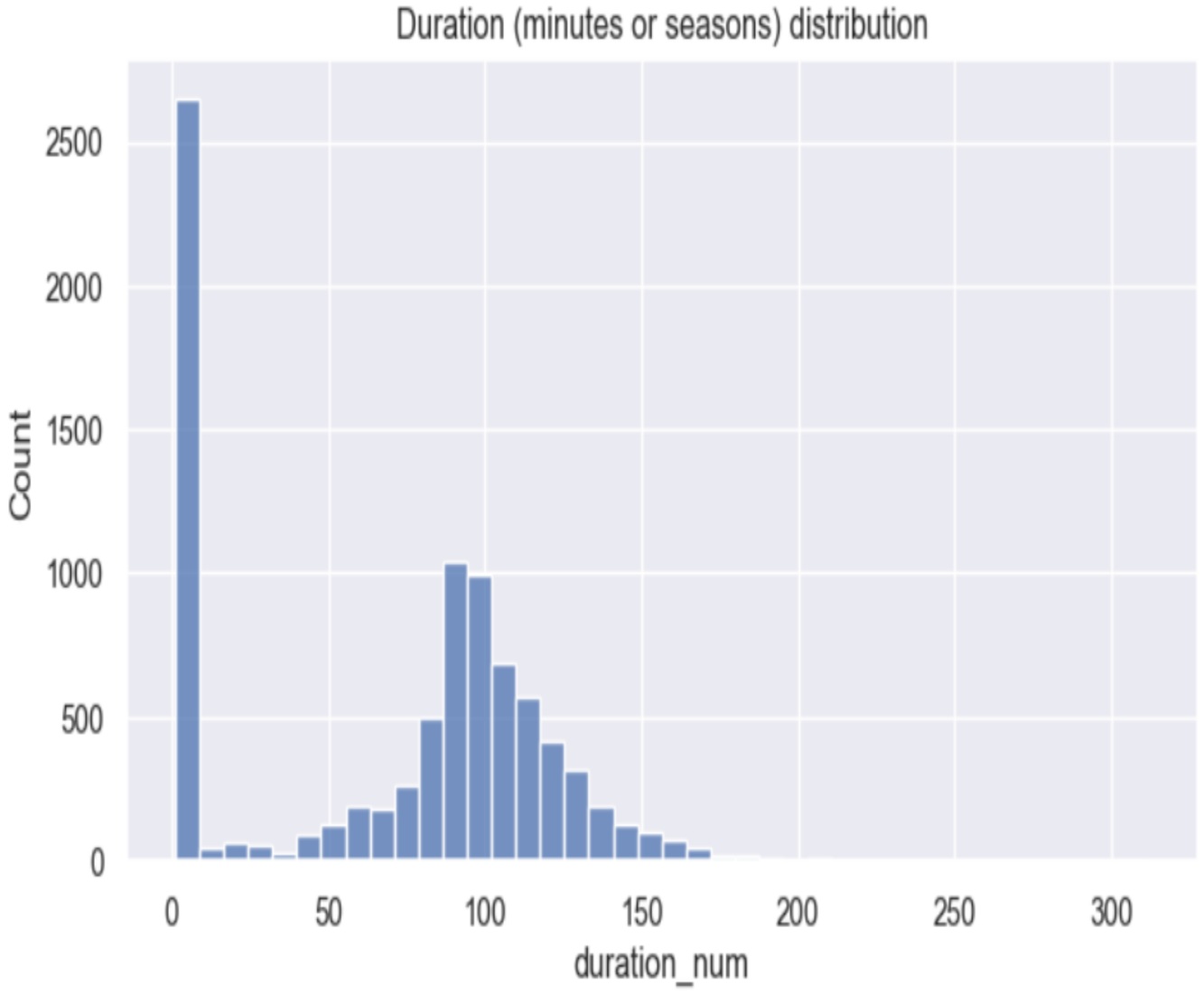
Text Feature Extraction

TF-IDF applied to descriptions for semantic representation.

5

Split Dataset

Divided into training and testing sets for model validation.



Classification Models: Movie vs. TV Show

Three machine learning models were trained to predict content type:

1

Logistic Regression

A baseline model for binary classification.

2

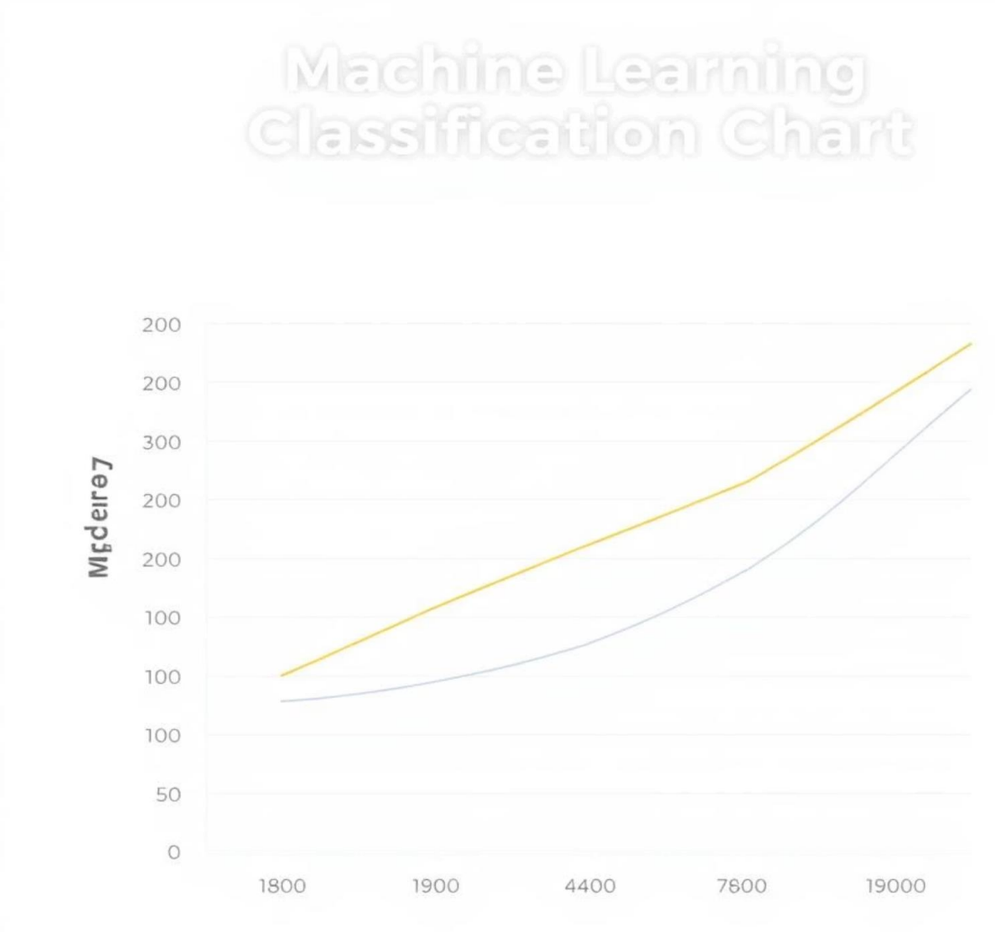
Random Forest

Ensemble method for robust prediction.

3

Gradient Boosting

Powerful algorithm for high accuracy.



Evaluation Metrics: Accuracy, Precision, Recall, and F1 Score were used to assess model performance. The model with the highest F1 Score was selected as optimal.

Clustering Analysis: Discovering Content Groups

K-Means clustering applied to categorize similar Netflix titles.

- ## K-Means Algorithm

  Clustered TF-IDF vectors derived from content descriptions.

- ## Dimensionality Reduction

  TruncatedSVD used for effective visualization of clusters.

- ## Identified Clusters

  Groups emerged, such as drama-focused, kids' content, and thrillers.

- ## Cluster Quality

  Silhouette score determined the cohesion and separation of clusters.

Content-Based Recommendation System

A personalized recommendation engine built on content similarity.

TF-IDF Vectors

Descriptions transformed into vectors for semantic comparison.

Cosine Similarity

Measures textual similarity between titles.

User Input

System takes a title, finds similar content.

Personalized Suggestions

Enhances user experience with Netflix-like personalization.

Deployment: Making Models Accessible

Trained models are saved for seamless integration and reusability.

Saved Models

tfidf.pkl and kmeans.pkl are stored.

Serialization

Joblib used for efficient model saving.

Integration Possibilities:

- Web applications
- Mobile applications
- Recommendation APIs

Ensures that models can be utilized without the need for retraining, streamlining future development and implementation.

THE END

THANK YOU FOR YOUR TIME