

## Netflix Project – Summary Report

Netflix Data Analysis, Visualization & Machine Learning – Project Summary

### 1. Introduction

This project delivers a full end-to-end analysis of the Netflix Titles Dataset, including data cleaning, exploratory data analysis (EDA), visualization, machine learning modeling, and a content-based recommendation system. The goal is to extract insights about Netflix's catalog and build practical ML components.

---

### 2. Data Cleaning & Preparation

The following steps were applied:

- Duplicate removal using key metadata
  - Standardization of text fields
  - Conversion of 'date\_added' to datetime (year, month, day)
  - Parsing duration into duration\_min (numeric)
  - Cleaning missing values
  - Extracting primary country
  - Splitting genres into genres\_list and num\_genres
  - Creating combined NLP field text\_all
- 

### 3. Exploratory Data Analysis (EDA)

Major insights include:

- Movies constitute ~70% of all titles; TV shows ~30%
- Most frequent ratings: TV-MA, TV-14, TV-PG

- Top countries: United States, India, United Kingdom
  - Popular genres: Drama, Comedy, Documentaries, International Movies
  - Sharp increase in content additions from 2015 onward
- 

#### **4. Machine Learning Components**

The ML experiments include:

##### **A. Text-based Feature Engineering:**

- TF-IDF transformation on text\_all
- Dimensionality reduction (optional)

##### **B. Classification models:**

- Logistic Regression
- Random Forest
- Hyperparameter tuning via GridSearchCV

##### **C. Evaluation:**

- Accuracy Score
  - Classification Report
  - Confusion Matrix
- 

#### **5. Content-Based Recommendation System**

A TF-IDF-based content recommendation system was implemented using:

- TfidfVectorizer on text\_all
- Cosine Similarity to retrieve close matches

The system produces high-quality content recommendations and is deployment-ready.

---

## 7. Deployment (Streamlit Application)

A Streamlit application was created with:

- **Search box for movie/show title**
- **Top-N recommendations**
- **Clean interface for presenting metadata**

Files required:

- netflix\_cleaned.csv
  - netflix\_vectorizer.pkl
  - netflix\_tfidf\_matrix.npz
  - streamlit\_app.py
- 

## 7. Conclusion

This project demonstrates full-stack data science execution:

**data preprocessing → visualization → ML modeling → recommendation engine → deployment.**

It highlights readiness for real-world analytics and ML responsibilities.

■ End of Report –