# Project Documentation

## 1. Dataset Description

- Source: Netflix dataset (contains metadata about Movies and TV Shows).  
- Columns used:  
 • country → Country of release  
 • listed\_in → Genre/category of the title  
 • type → Target variable (Movie or TV Show)  
  
The dataset provides categorical features which we transform into numerical values for machine learning.

## 2. Data Cleaning & Transformation

- Handling missing values: Dropping rows where critical columns (country, listed\_in, type) are missing.  
- Label Encoding:  
 • Converted country into numeric labels using LabelEncoder.  
 • Converted listed\_in (genre) into numeric labels.  
 • Encoded target column type (Movie=0, TV Show=1).  
- Final dataset prepared for model training.

## 3. Exploratory Data Analysis (EDA) Insights

- Distribution of types: More Movies than TV Shows in the dataset.  
- Country impact: Certain countries dominate the dataset (e.g., USA, India).  
- Genre distribution: Some genres are strongly correlated with one type (e.g., Documentaries lean more towards TV Shows).  
- Visualizations: Bar plots, count plots, and correlation checks performed.

## 4. Feature Selection Process

- Features chosen:  
 • country (encoded)  
 • genre (encoded)  
- Target: type (Movie or TV Show).  
- Selected after testing, as they had the strongest influence on prediction.

## 5. Model Development, Evaluation & Deployment

- Model used: RandomForestClassifier  
- Training: Dataset split into train/test sets.  
- Evaluation: Accuracy measured (fill in your result).  
- Deployment:  
 • Streamlit app built.  
 • App takes Country and Genre as input.  
 • Outputs prediction of Movie or TV Show.

## 6. Code Documentation

- Each code file contains inline comments in present continuous tense, explaining each step.  
 Example:  
 # Loading the dataset  
 # Encoding categorical features  
 # Training the Random Forest model  
 # Deploying with Streamlit