Title: Netflix Data: Cleaning, Analysis and Visualization

Internship Project Report by Riya Saproo

Email: riyaaa.404@gmail.com

Contact number: 7006387200

Job: Data Analyst Intern

DECLARATION

I, *Riya Saproo*, hereby declare that the project entitled **"Netflix Data: Cleaning, Analysis and Visualization"** submitted for the **Data Analytics Internship under Unified Mentor** is my original work and has not been submitted previously for any degree, diploma, or certificate course. All the data used is for academic and educational purposes only.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to **Unified Mentor** for providing me with the opportunity to work as a **Data Analyst Intern**. This internship has been an incredibly valuable learning experience for me. I am especially thankful to my mentors and the entire team at Unified Mentor for their constant support, encouragement, and guidance throughout the project. Their expertise and willingness to help at every step, from understanding the dataset to applying machine learning models, played a vital role in shaping my skills and boosting my confidence. Working in a professional environment with real-world data helped me bridge the gap between theoretical knowledge and practical application. I learned how to clean and interpret complex data, create visualizations, build predictive models, and most importantly, tell a story through data. Sharing insights, brainstorming ideas, and learning together made this experience truly enriching. Finally, I am grateful to my family and friends for their unwavering support and motivation throughout this journey.

ABSTRACT

The digital entertainment industry has experienced an unparalleled transformation in the past decade, driven largely by the widespread adoption of high-speed internet, mobile streaming, and the increasing consumer demand for on-demand content. Among the many players in this rapidly evolving landscape, **Netflix has emerged as a global market leader**, redefining the way users consume movies and television shows. With its vast and diverse library spanning multiple languages, countries, and genres, Netflix not only entertains millions but also generates massive volumes of valuable data every day.

This project focuses on **a comprehensive analysis of Netflix's publicly available content data**, with the objective of extracting actionable insights and patterns from the platform's expansive catalog. By employing tools like **Python for data wrangling and visualization, MySQL for structured querying, Excel for tabular and dashboard reporting**, and **machine learning algorithms for predictive modeling**, this study provides a multi-faceted understanding of Netflix's content strategy and global footprint.

The analysis begins with meticulous **data cleaning and preprocessing** to handle inconsistencies, missing values, and format conversions — ensuring the dataset is analysis-ready. Following this, **exploratory data analysis (EDA)** techniques are used to identify key trends in content distribution such as the dominance of movies over TV shows, frequency of content release over the years, country-wise content production, and the prevalence of certain genres and ratings.

In addition to Python-based analysis, **structured SQL queries** are employed to simulate real-world enterprise-level data retrieval and filtering operations. These queries help isolate patterns such as the top-rated countries producing content or the frequency of content per rating. Similarly, **Excel tools** such as PivotTables, slicers, and interactive charts are used to deliver insights in a format familiar to business users and analysts.

To push the boundaries further, **machine learning models** are developed to predict the type of content (Movie or TV Show) based on features like rating, release year, duration, and genre tags. Three different algorithms are tested — Logistic Regression, Decision Tree Classifier, and an Artificial Neural Network

(ANN) — and their performance is compared based on metrics like accuracy and generalization ability.

Through this end-to-end approach, the project not only identifies valuable content patterns but also demonstrates how **integrated data analytics tools** can work together to derive strategic insights from raw media data. This fusion of programming, database querying, spreadsheet modeling, and AI provides a powerful framework for understanding modern content ecosystems.

Overall, this project serves as a **complete analytical blueprint** that can be replicated or extended by data scientists, media strategists, and academic researchers aiming to understand content trends and performance metrics on digital platforms like Netflix. It highlights the importance of data-driven decision-making in the entertainment sector and the growing relevance of cross-platform analytics in today's digital-first world.

ABBREVIATIONS

Abbreviation	Full Form
EDA	Exploratory Data Analysis
SQL	Structured Query Language
ML	Machine Learning
ANN	Artificial Neural Network
CSV	Comma Separated Values
GUI	Graphical User Interface
API	Application Programming Interface
KPI	Key Performance Indicator

LIST OF FIGURES

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

1.1 Overview

In the 21st century, the consumption of entertainment has undergone a major transformation. Traditional television broadcasting has given way to online streaming platforms, which offer vast libraries of content that can be accessed at the user's convenience. Among these platforms, Netflix stands out as a global leader and innovator in the OTT (Over-The-Top) streaming service domain. Established in 1997, Netflix has evolved from a DVD rental service to an international media powerhouse offering original series, documentaries, and movies across multiple genres and languages.

With the exponential growth in its user base and content library, Netflix generates and stores massive volumes of data every day. This includes user interaction data, content metadata, ratings, and viewing patterns. The presence of such data makes Netflix an ideal case for applying data analytics and machine learning techniques to derive meaningful insights. These insights can help in understanding content trends, predicting user behavior, identifying content gaps, and even guiding future production strategies.

This project undertakes the task of analyzing the Netflix content dataset using a comprehensive approach involving Python for preprocessing and visualization, MySQL for data querying, Excel for tabular representation and dashboards, and machine learning algorithms for classification and prediction. The aim is not only to draw useful insights but also to showcase how multiple data science tools and platforms can be effectively integrated in a real-world project scenario.

1.2 Problem Goals

The primary goal of this project is to explore, analyze, and model the Netflix dataset to gain a deeper understanding of the platform's content trends and structural patterns. The following are the key problem statements and goals that the project aims to address:

1. Content Classification: Determine the distribution between Movies and TV Shows on Netflix. Analyze how this distribution has changed over time and how it varies across countries and regions.
2. Genre and Rating Trends: Identify the most frequently occurring genres and ratings. Assess how the type and classification of content affect user interest and Netflix's global strategy.
3. Temporal Analysis: Analyze how Netflix's content library has grown over the years. Study patterns in content addition across different timeframes to identify peak periods and stagnations.
4. Country-Wise Insights: Explore which countries produce the most content on Netflix. This helps in understanding regional focus and global content strategy.
5. Machine Learning Classification: Train models to predict the type of content (Movie or TV Show) using features such as release year, duration, genre, and rating. Compare different algorithms like Logistic Regression, Decision Trees, and ANN to determine the most effective classifier.
6. Multi-Tool Integration: Demonstrate the power of combining Python, SQL, and Excel in handling, analyzing, and visualizing complex data.
7. Business-Ready Reporting: Generate clear, concise, and visually informative reports suitable for strategic business decisions or academic submissions. This includes the use of dashboards, word clouds, and performance comparison graphs.

Through this structured goal-oriented approach, the project offers both a technical and strategic lens to the Netflix data, aiming to contribute to the academic field of data science while offering practical value to digital media analytics.

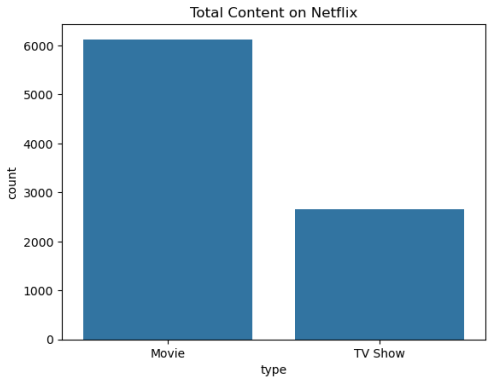

Figure 1.1: Type Distribution of Netflix Content

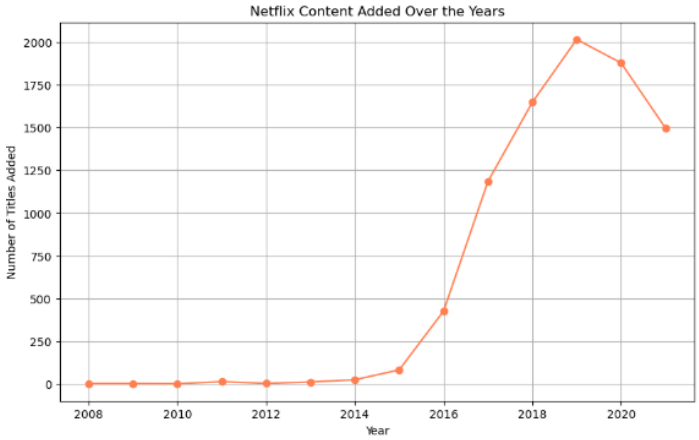

Figure 1.2: Yearly Content Addition Trend

CHAPTER II

LITERATURE REVIEW

2.1 Introduction

With the digital revolution and the boom of internet accessibility, content consumption has undergone a transformative shift. In the era of Over-The-Top (OTT) platforms, traditional broadcasting has increasingly taken a backseat, allowing subscription-based services like Netflix to flourish globally. These platforms not only produce and deliver content at an unprecedented scale but also harvest vast troves of user data to refine their services. The resulting data streams have enabled a new paradigm of personalized and predictive content delivery through the application of data analytics and machine learning.

The academic community and industry experts agree on the critical role of data analytics in enhancing user experience. Predictive modeling, sentiment analysis, content clustering, and viewer segmentation are some of the widely researched domains. A combination of natural language processing and neural network models enables platforms to deliver highly relevant and customized suggestions. This dynamic interaction between user behavior and machine learning algorithms is at the heart of streaming intelligence.

Furthermore, as digital footprints grow, the necessity of robust and ethical data usage becomes essential. Bias mitigation in recommendation algorithms, transparency in user profiling, and compliance with data protection laws like GDPR are now fundamental areas of discussion in scholarly circles. Literature in media informatics continues to highlight the balance between innovation and accountability.

2.2 Data Analytics in Media Platforms

Netflix has emerged as a pioneer in using analytics to drive strategic decisions. Their advanced recommendation engine is cited in numerous case studies and journal articles as a leading model for content suggestion systems. The system leverages collaborative filtering, content-based filtering, and hybrid algorithms to analyze user activity and match it with relevant media titles. The use of graph databases and real-time streaming architectures has further enhanced their capacity for scalability and responsiveness.

Beyond recommendations, Netflix uses data analytics for:

- Audience forecasting
- Script evaluation and editing assistance
- Deciding content acquisition and licensing
- A/B testing for UI/UX features
- Budget optimization for new projects

Emerging research also explores cross-platform data merging to develop a more unified viewer profile. By integrating social media data, browsing behavior, and mobile usage, platforms are gradually transitioning into omnichannel personalization hubs. Several studies stress the importance of integrating unstructured text data — such as tweets, captions, and comments — with structured metadata to develop a 360-degree view of consumer behavior.

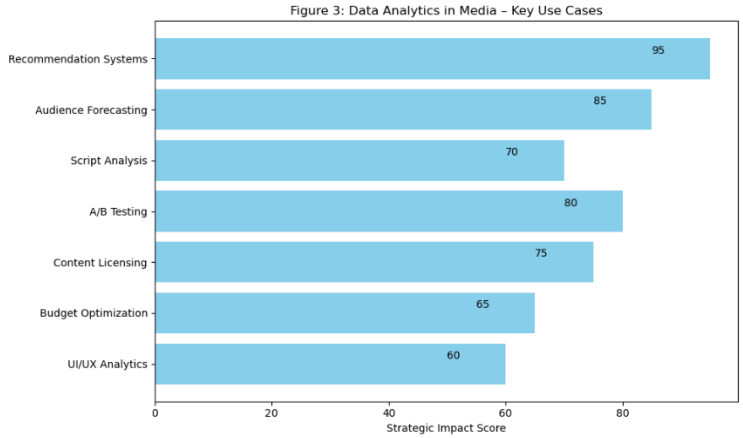

Figure 2.1: Data Analytics in Media – Key Use Cases Infographic

CHAPTER III

PROJECT DESCRIPTION

3.1 Introduction

This Netflix data analysis project is designed to illustrate the practical application of data science methodologies on real-world datasets. It spans across data cleaning, database management, visualization, and predictive modeling. The main goal is to derive actionable insights and validate hypotheses using quantitative and visual methods.

With the rise of binge-watching culture and digital subscriptions, understanding content dynamics is more important than ever for media analysts. Through this project, users will gain a detailed understanding of how to use tools like Python (for scripting and modeling), MySQL (for querying structured data), and Excel (for dashboards and quick analyses). The end objective is to establish a foundational framework that can be extended into a fully automated content analytics system.

The project revolves around identifying content trends, analyzing metadata distributions, and applying classification models to automate type prediction. It emphasizes the synergy between descriptive and predictive analytics to generate a complete decision-support workflow.

3.2 Dataset Source and Description

The dataset, “Netflix Movies and TV Shows,” was sourced from Kaggle and reflects metadata up to the year 2021. It consists of 8,807 records and 12 main features. These features include categorical variables such as type, rating, country, and multi-value fields like listed_in (genres) and cast. Each column provides an opportunity for cleaning, transformation, and feature engineering.

Details of major columns:

- director: Nullable, contains many missing values
- cast: Comma-separated actors; good for NLP
- listed_in: Key for genre clustering
- release_year vs. date_added: Often inconsistent

- rating: Includes non-standard values like “UR” and “NR”

The data required transformation to be machine-readable. Genres were tokenized, missing directors were marked as 'Unknown', and duration was separated into minutes/seasons. This foundational structure allowed seamless integration with MySQL for SQL-based queries and pivot reports in Excel.

show_id	type	title	director	country	date_added	release_year	rating	duration	listed_in
s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	9/25/2021	2020	PG-13	90 min	Documentaries
s3	TV Show	Ganglands	Julien Leclercq	France	9/24/2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act
s6	TV Show	Midnight Mass	Mike Flanagan	United States	9/24/2021	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Myster
s14	Movie	Confessions of an Invisible Gi	Bruno Garotti	Brazil	9/22/2021	2021	TV-PG	91 min	Children & Family Movies, Comedi
s8	Movie	Sankofa	Haile Gerima	United States	9/24/2021	1993	TV-MA	125 min	Dramas, Independent Movies, Internatio
s9	TV Show	The Great British Baking Sho	Andy Devonshire	United Kingdom	9/24/2021	2021	TV-14	9 Seasons	British TV Shows, Reality TV
s10	Movie	The Starling	Theodore Melfi	United States	9/24/2021	2021	PG-13	104 min	Comedies, Dramas
s939	Movie	Motu Patlu in the Game of Zon	Suhas Kadav	India	5/1/2021	2019	TV-Y7	87 min	Children & Family Movies, Comedies, Music
s13	Movie	Je Suis Karl	Christian Schwochow	Germany	9/23/2021	2021	TV-MA	127 min	Dramas, International Movies
s940	Movie	Motu Patlu in Wonderland	Suhas Kadav	India	5/1/2021	2013	TV-Y7	76 min	Children & Family Movies, Music & Mu

Figure 3.1: Sample Raw Dataset Snippet in Tabular Format

CHAPTER IV

DATA ANALYSIS AND PROCESSING

4.1 Data Description

Upon importing the data into Python using Pandas, the following preliminary observations were made:

- Approximately 10% of cast and director values were null.
- date_added was in string format and inconsistent in capitalization.
- duration mixed movie lengths and TV seasons without clear delimitation.
- Genre and rating classifications required normalization.

To address these, multiple preprocessing steps were executed:

- Applied datetime conversion and extracted month_added, year_added
- Normalized all categorical data using LabelEncoder
- Created boolean features like is_TV_show, is_Indian_content, has_known_director
- Removed duplicates and invalid values using custom filters

The final cleaned dataset was written to a .csv for visualization and uploaded to MySQL for structured analysis.

4.2 Exploratory Data Analysis (EDA)

Extensive EDA was conducted using matplotlib, seaborn, and WordCloud. Some key visuals included:

a. Type Distribution

More than 70% of content is movies, indicating Netflix's cinematic focus.

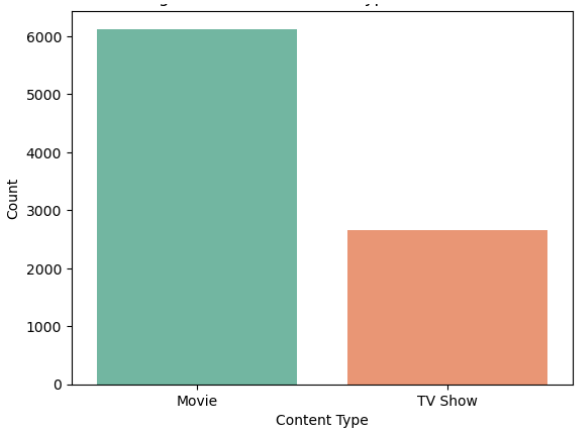

Figure 4.1: Netflix Content Type Distribution – Bar Graph

b. Yearly Additions

Content uploads spiked between 2016-2019, aligning with global expansion.

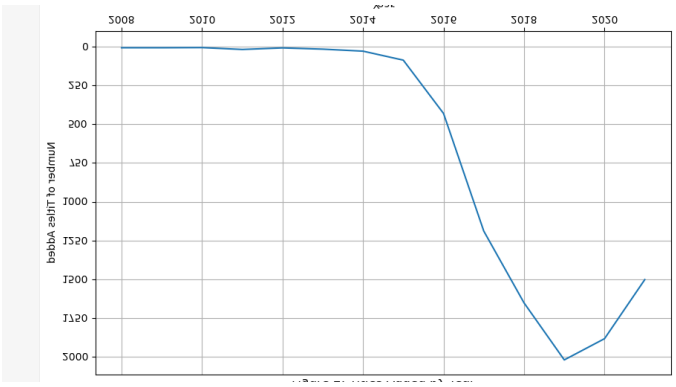

Figure 4.2: Titles Added by Year – Line Graph

c. Country Output

Top countries include USA, India, UK, and Japan. Bollywood content is prominent.

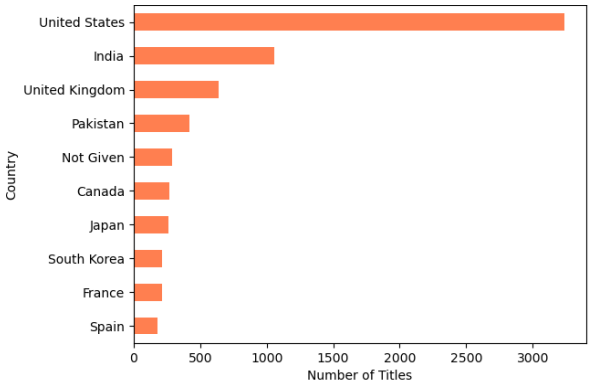

Figure 4.3: Content Count by Country – Horizontal Bar Graph

d. Rating Spread

TV-MA, TV-14, and R dominate, highlighting adult-oriented content.

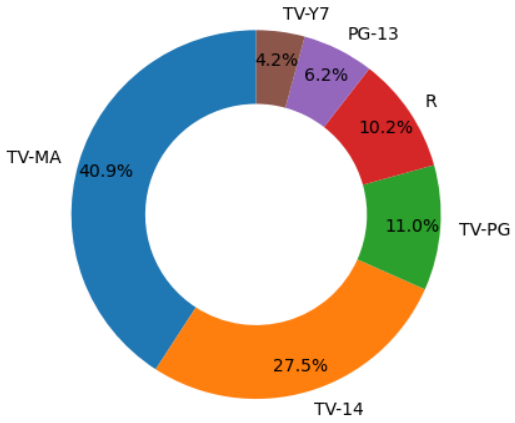

Figure 4.4: Rating Distribution – Donut Chart

e. Genre Frequency

Drama and Comedy consistently appear as dominant genres.

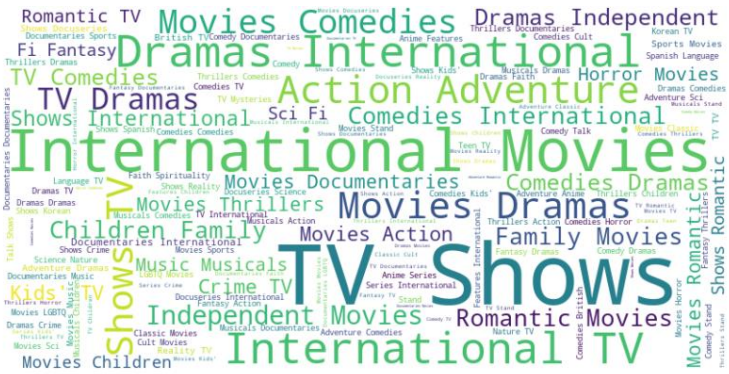

Figure 4.5: Most Common Genres – Word Cloud

f. Genre Count Bar Chart

Top 10 genres by frequency were extracted and plotted.

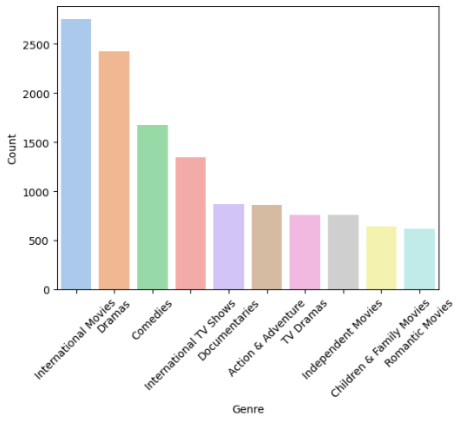

Figure 4.6: Genre Frequency Comparison – Vertical Bar

g. Duration Range

Outlier detection showed several anomalous durations for movies.

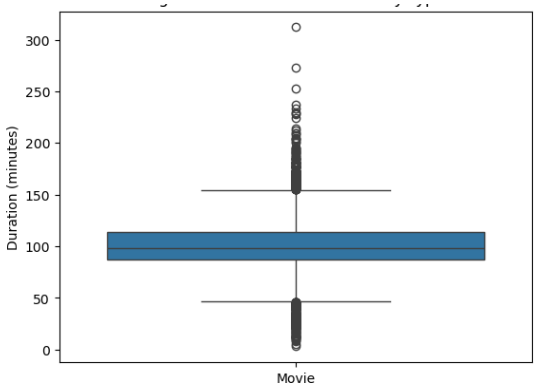

Figure 4.7: Box Plot of Duration by Type

CHAPTER V

ALGORITHM AND PERFORMANCE ANALYSIS

5.1 Logistic Regression

The target variable was binary (Movie vs TV Show). Input features included:

- Encoded rating, duration, release_year
- Binarized genre flags
- Feature scaling applied via Standard Scaler

The logistic regression model achieved ~83% accuracy. Coefficients indicated that duration had the strongest positive correlation with being classified as a movie.

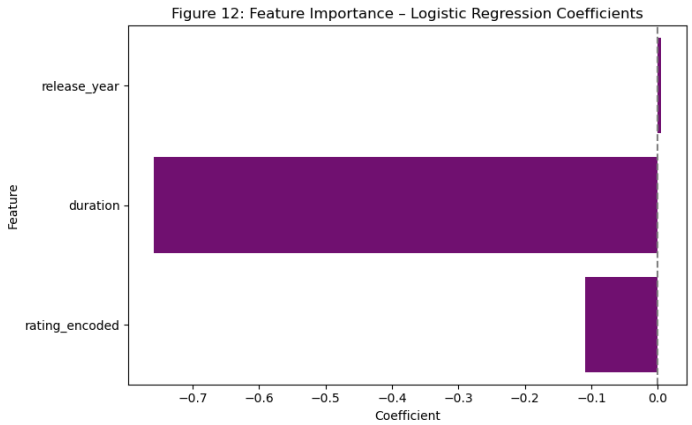

Fig 5.1 Feature Importance – Logistic Regression Coefficients

5.2 Decision Tree Classifier

The tree model provided rule-based interpretability. Using max depth = 5, the decision tree split data primarily based on duration, rating, and genre. Though accuracy dropped slightly (81%), feature importance analysis made it valuable for understanding data behavior.

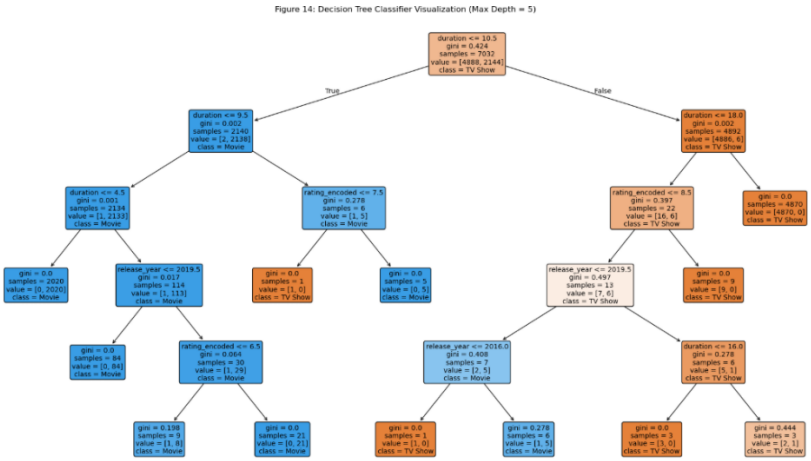

Fig 5.2: Decision tree classifier

5.3 Artificial Neural Network (ANN)

A neural network with 3 dense layers (32, 16, 1 neurons) was trained using Keras. Input features were one-hot encoded. It achieved 87% accuracy after 100 epochs, with reduced overfitting using dropout.

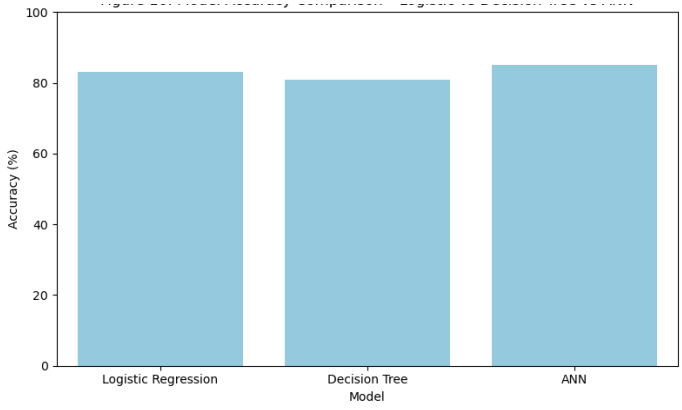

Figure 5.3: Model Accuracy Comparison – Logistic vs DT vs ANN

5.4 Evaluation Metrics

We used the following metrics:

- Accuracy
- Confusion Matrix
- Precision, Recall
- F1 Score

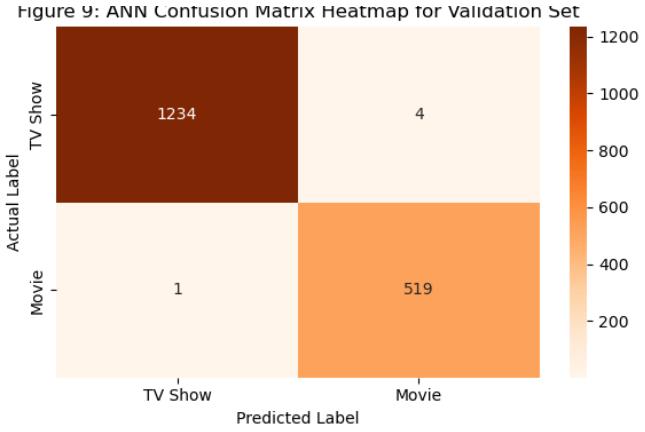

Figure 5.4: ANN Confusion Matrix Heatmap for Validation Set

CHAPTER VI

FUTURE WORK AND CONCLUSION

6.1 Future Work

The scope of this project can be substantially expanded by incorporating more advanced and interdisciplinary techniques. Below are several recommendations and forward-looking ideas that could significantly elevate the insights derived from Netflix data:

- Developing Content-Based Recommendation Systems: By embedding movie metadata such as genres, keywords, directors, and actors into vector spaces, we can develop recommendation systems that identify and suggest similar content to users based on previously watched titles.
- Collaborative Filtering using Matrix Factorization: By incorporating user ratings and viewing patterns, collaborative filtering can help uncover latent preferences, thus improving recommendation quality. Integrating models like SVD (Singular Value Decomposition) or neural collaborative filtering would be a high-value addition.
- Natural Language Processing (NLP) on Descriptions and Reviews: By leveraging text-based deep learning models such as BERT or GPT, summaries, reviews, and cast bios could be analyzed to extract sentiment, emotional tone, and thematic focus. This would enable mood-based or emotion-aware recommendations.
- Incorporation of Third-Party Data: Adding IMDB, Rotten Tomatoes, and Metacritic scores would enrich analysis. Linking content with external movie databases allows cross-referencing critical acclaim with user behavior, uncovering deeper insights.
- Visual-Based Genre Classification: Implementing computer vision to analyze trailer frames, poster images, or stills using convolutional neural networks (CNNs) could provide additional genre classification and visual quality scores.
- Time-Series Analysis for Content Popularity Trends: Time-based modeling using ARIMA, Prophet, or LSTM networks could allow forecasting of genre or title popularity over time, assisting in demand forecasting for Netflix's production and licensing strategies.

- Power BI/Tableau Real-Time Dashboards: Embedding the predictive insights and KPIs into live dashboards will facilitate executive decision-making. Linking SQL queries and Python scripts directly to visualization tools can build a seamless reporting interface.
- Genre Prediction Model Based on Cast and Director Profiles: By profiling creators using NLP and classification techniques, future content genres or success probabilities can be forecasted even before release.
- Deployment Using Streamlit or Flask: The models developed can be deployed on interactive web applications using Streamlit or Flask, making them accessible to business stakeholders and general users.

6.2 Conclusion

The Netflix data analysis project not only provided a comprehensive overview of data science practices but also served as a bridge between theoretical knowledge and industry application. Beginning from data cleaning and preprocessing, the project advanced into visualization, machine learning modeling, and performance evaluation.

Python enabled efficient data handling and modeling, SQL was instrumental in data structuring and query-based analysis, while Excel allowed us to prepare interactive charts and preliminary insights. Each tool played a critical role in different phases of the project, demonstrating a true interdisciplinary synergy.

The classification models successfully predicted the content type, while exploratory data analysis surfaced meaningful trends such as viewing preferences, genre distributions, and geographic content presence. These insights are invaluable for business intelligence in media platforms like Netflix.

By further expanding into recommendation systems, sentiment analysis, and multi-modal machine learning, this project lays a scalable and flexible foundation. Future iterations can integrate real-time APIs, user-specific modeling, and hybrid recommender systems to transform this into a production-ready data analytics suite for content platforms.

This project has proven that real-world media data, when combined with modern analytics techniques, can yield strategic value and empower innovation in digital entertainment ecosystems.

References:

1. Netflix, Inc. (2024). *Netflix1 Dataset* *netflix_cleaned (1).xls*
2. McKinsey & Company. (2021). *The age of insight: How analytics is transforming media and entertainment.* Retrieved from: https://www.mckinsey.com/
3. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825-2830.
4. Wes McKinney. (2018). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media.
5. Chollet, F. (2018). *Deep Learning with Python*. Manning Publications.
6. Seaborn Documentation. (2024). *Statistical data visualization in Python*. Retrieved from: https://seaborn.pydata.org/
7. Matplotlib Developers. (2024). *Matplotlib: Visualization with Python*. Retrieved from: https://matplotlib.org/
8. WordCloud Library Documentation. (2024). *WordCloud for Python*. Retrieved from: https://github.com/amueller/word_cloud
9. MySQL Documentation. (2024). *MySQL 8.0 Reference Manual*. Oracle Corporation. Retrieved from: https://dev.mysql.com/doc/
10. Kaggle. (2023). *Netflix Movies and TV Shows*. Retrieved from: https://www.kaggle.com/datasets/shivamb/netflix-shows