Nikhil Khanchandani

KDD Process Netflix Dataset Report

1. Data Selection: The dataset used in this project is the Netflix Shows dataset obtained from Kaggle. It contains a variety of information about Netflix content, including title, release year, genre, country of origin, and rating. For our analysis, we focused on the key attributes that allow us to gain insights into global Netflix content trends. These included 'title', 'listed_in' (genre), 'release_year', 'country', 'rating', and 'duration'. This step ensures that only the relevant data is selected for the task.

2. Data Preprocessing: In this phase, we cleaned the data by filling missing values and removing rows with missing essential data like 'release_year' and 'country'. Missing values in the 'rating' column were replaced with 'Not Rated', ensuring that our dataset was complete and ready for analysis. Furthermore, categorical variables such as 'listed_in' and 'country' were prepared for transformation in the next phase.

3. Data Transformation: We engineered new features like 'content_age', which measures how old each piece of content is, by subtracting the release year from the current year. Additionally, we applied one-hot encoding to the categorical columns, transforming 'listed_in' (genres) and 'country' into numerical representations. This ensured that the data was suitable for our analysis and modeling.

4. Data Mining: We used data mining techniques to explore trends in Netflix content production. By grouping the data by 'country' and 'listed_in', we identified which countries produce the most content and which genres dominate Netflix’s catalog.

- The United States leads in content production, followed by countries like India and the UK.
- In terms of genres, 'Dramas' and 'Comedies' were the most common across Netflix.

5. Interpretation: Our analysis revealed several key insights about Netflix's content strategy. The dominance of the US in content production aligns with Netflix's market presence in the country, but the growing influence of countries like India reflects Netflix's global expansion efforts. The high prevalence of genres like 'Dramas' and 'Comedies' suggests that these genres resonate with a wide audience, driving viewership across regions.

6. Knowledge: From this analysis, we gained actionable insights into how Netflix tailors its content offerings to different regions and preferences. These findings can help inform content creation strategies, especially in identifying growth areas for Netflix, such as expanding content in non-English-speaking markets. Additionally, genres like 'Action' and 'Thriller' could be explored further for their potential in underrepresented regions.

Conclusion: The KDD methodology provided a structured approach for extracting valuable insights from the Netflix dataset. By following each phase systematically, we were able to uncover patterns and trends that reflect Netflix's content strategy and its global market expansion. Future analyses could focus on exploring specific genres in more depth or applying machine learning models to predict the success of content based on these attributes.