

# IMDb Final Summary & Insights Report

## 1. Objective Recap

This subtask focuses on compiling the final summary and key insights obtained from analyzing the IMDb dataset. The dataset contains movie metadata including features such as Title, Genre, Rating, Votes, Year, and possibly others like Director, Duration, etc.





The primary goal was to perform data cleaning, preprocessing, visualization, and feature engineering to uncover meaningful patterns and prepare the dataset for potential modeling tasks.

	precision	recall	f1-score	support
0	0.74	0.89	0.81	104
1	0.85	0.67	0.75	96
accuracy			0.79	200
macro avg	0.80	0.78	0.78	200
weighted avg	0.80	0.79	0.78	200

## 2. Key Insights





- 🎬 **Genre Trends:** The most frequent genres across the dataset were **Drama**, **Comedy**, and **Action**. These genres dominate both high-vote and high-rating segments.
- ★ **Ratings & Votes:** Movies with higher vote counts generally had more stable and higher average ratings, especially when released during or after the 1990s.
- 📅 **Year-wise Distribution:** Most of the top-rated movies were released between **1990 and 2020**, with spikes in activity around 2010–2015.
- 🔧 **Feature Engineering:**
  - Applied **One-Hot Encoding** to the Genre column, which allowed categorical genre data to be used in models.
  - Used **Standardization and MinMax Scaling** on numerical features like Votes and Rating to bring values into a consistent range.
- 🔗 **Correlations:**
  - Votes had a moderate-to-strong positive correlation with Rating, especially in top genres.
  - Movies with a larger audience base often scored well, suggesting popularity aligns with perceived quality.

### 3. Challenges Addressed

-  **Missing Data:** Some movies lacked values for features like Rating or Year. These were handled by either imputation or removal, depending on importance.
-  **Duplicates:** Duplicate entries were dropped to ensure clean, unique records.
-  **Categorical Encoding:** Genre and other categorical features were encoded using One-Hot Encoding for compatibility with machine learning models.
-  **Scaling:** Feature scaling addressed the imbalance in feature ranges (e.g., thousands of votes vs. ratings out of 10), improving model readiness.

### 4. Conclusion & Next Steps

The IMDb dataset is now fully cleaned, preprocessed, and ready for machine learning tasks such as:

-  **Predicting movie ratings**
-  **Genre classification**
-  **Recommendation systems**
-  **Clustering based on genre popularity or ratings**

The insights gained offer a strong foundation for understanding movie trends and can guide further exploration into audience behavior, movie success factors, and time-based trends in cinema.