**Report: Movie Recommendation System with Predictive Feature Selection**

1. Introduction

This report outlines the development of a content-based movie recommendation system using the TMDB 5000 Movies dataset. The goal was to predict the most likely movies a user would enjoy based on similarity in genres, plot keywords, cast, and directors. The project also incorporated three feature selection methods (Filter, Wrapper, and Embedded) to optimize the model.

2. Dataset Selection

Dataset Used

Source: TMDB 5000 Movies

Size: 4,806 movies

Key Features:

title, genres, keywords, cast, crew, overview

Why This Dataset?

Contains rich metadata (genres, actors, directors, plot summaries).

Ideal for content-based filtering (no user ratings required).

Publicly available and well-structured.

**3. Objective**

The primary objective was to:

Predict the most similar movies based on content features (genres, keywords, cast, crew).

Identify key influential features using Filter, Wrapper, and Embedded methods.

**4. Methodology**

(1) Data Preprocessing

a. Data Cleaning

Merged the movies and credits datasets on title.

Removed missing values and duplicates.

Selected relevant columns: id, title, genres, keywords, cast, crew, overview.

b. Feature Extraction

Genres & Keywords: Extracted from JSON format into lists.

Cast: Selected the top 3 actors per movie.

Crew: Extracted director names only.

Text Processing:

Combined overview, genres, keywords, cast, and crew into a single tags column.

Applied stemming (PorterStemmer) to reduce words to root forms.

(2) Feature Selection

To improve recommendation quality, three feature selection methods were applied:

a. Filter Method: Mutual Information

Goal: Identify words most predictive of movie genres.

Process:

Calculated Mutual Information (MI) between words and genre labels.

Selected top words (e.g., "war", "hero", "scifi").

Findings: Words like "battle" and "space" were highly predictive of Action/Sci-Fi genres.

b. Wrapper Method: Recursive Feature Elimination (RFE)

Goal: Select the best 20 features for genre prediction.

Process:

Used Logistic Regression with RFE.

Iteratively removed the least important features.

Findings: Director and lead actor names were among the top selected features.

c. Embedded Method: Lasso Regression

Goal: Use L1 regularization to eliminate irrelevant features.

Process:

Trained a Lasso Regression model to predict movie ratings.

Features with non-zero coefficients were retained.

Findings: Words like "epic" and "romance" were key predictors of ratings.

(3) Model Building

Content-Based Recommender

Vectorization: Converted movie tags into TF-IDF vectors (5,000 most frequent words).

Similarity Calculation: Used cosine similarity to find the closest matches.

Recommendation Function:

Takes a movie title as input.

Returns the top 5 most similar movies based on cosine similarity.

5. Results & Findings

Key Insights

Feature Selection Impact

Filter Method (MI): Highlighted genre-specific keywords.

Wrapper Method (RFE): Prioritized cast and director names.

Embedded Method (Lasso): Emphasized descriptive plot terms.

Recommendation Examples

Input: "The Dark Knight Rises" → Output:

Ex.1

1. Batman Begins

2. The Dark Knight

3. Inception

4. Man of Steel

5. Watchmen

Ex.2

Input: "Toy Story" → Output:

1. Toy Story 2

2. Toy Story 3

3. Monsters, Inc.

4. Finding Nemo

5. Up

Performance Observations

Movies with shared directors/actors ranked higher in similarity.

Genre keywords (e.g., "animation", "superhero") improved relevance.

6. Conclusion

Successfully built a content-based recommender using TF-IDF and cosine similarity.

Verified that directors, lead actors, and genre keywords are the most influential features.

Feature selection improved interpretability and recommendation quality.