Internship Project Report

Project Title: Movie Recommendation System
Name: Divyadharshini.S
College: Avinashilingam Institute For Home Science And Higher Education For Women
Internship Duration: 2 Weeks
Submission Date: 23 June 2025

1. Introduction

Finding the right movie to watch can be difficult, especially with so many options available online. That's why I decided to work on a movie recommendation system for my internship project. The idea was to suggest movies similar to a user's favorite movie by analyzing content like genres, keywords, cast, and more. This project gave me a great opportunity to learn how real-world recommendation systems work.

2. Abstract

In this project, I built a simple movie recommendation system using a method called content-based filtering. I used the TMDB 5000 dataset, which contains information about thousands of movies, including their overviews, genres, cast, and keywords. After cleaning and processing the data, I combined important details into a new feature called "tags". These tags were then converted into numerical form using TF-IDF vectorization. Finally, I used cosine similarity to measure how closely related two movies are, and the system suggests the top 5 movies that are most similar to the one the user enters.

3.Tools & Technologies Used

- **Python** – For writing and running the code

- **Google Colab** – For working on the project online

- **Pandas & NumPy** – For data analysis and handling

- **Scikit-learn** – For TF-IDF vectorizer and similarity calculations

- **TMDB Dataset** – Used for training the recommendation system

4. Steps I Followed

1. **Loaded the Dataset** – I merged two files (tmdb_5000_movies.csv and tmdb_5000_credits.csv) using the movie titles.

2. **Selected Important Features** – I picked columns like overview, genres, cast, crew, and keywords.

3. **Cleaned the Data** – Since many of these columns had text stored as code-like strings, I extracted just the useful words (like the top 3 actors, director, etc.).

4. **Created Tags** – I combined all the cleaned information into a single column called tags.

5. **Vectorized the Tags** – I used TF-IDF to convert the text into numbers the model could understand.

6. **Calculated Similarity** – Using cosine similarity, I found how similar one movie is to another.

7. **Built the Recommender Function** – Finally, I created a function where users can enter a movie name and get recommendations based on content similarity.

5. Conclusion

This project was a great learning experience. I not only understood how recommendation systems work but also practiced working with real-world datasets. The model is easy to use, and the results are relevant and fast. In the future, this system can be improved further by including user ratings or reviews, or by building a collaborative filtering version. Overall, I enjoyed building it and learned a lot about data cleaning, vectorization, and similarity-based machine learning.