The Project and Data Management Plan

Project Title: Movies Recommendation System

Research Question: How can the MovieLens dataset be leveraged to build an effective system that predicts user preferences and delivers personalized movie recommendations? This research explores the development of a machine learning-based system designed to enhance user experience through tailored recommendations.

Aims and Objectives: In order to give consumers tailored movie recommendations, the project's goal is to build and construct a system that uses the MovieLens dataset. Among the goals are:

1. Examining the dataset to identify trends and user preferences.
2. Applying assessment criteria like RMSE, precision, and recall to compare the effectiveness of three machine learning models: Content-Based Filtering, Collaborative Filtering, and Matrix Factorization.
3. Developing visuals to examine user-movie interactions, such as clustering and heatmaps.
4. Creating an intuitive user interface to showcase the recommendation system.

To learn more about the dataset, the project will start with exploratory data analysis (EDA). The application and comparison of the three models will come next. After that, the model with the highest performance will be prepared for deployment. Although the strategy is detailed.

Background: Movie recommendation systems play a crucial role in modern digital platforms by predicting user preferences and suggesting content, thereby enhancing personalized experiences. Streaming services such as Netflix, Amazon Prime, and Hulu widely utilize these technologies to improve customer engagement and satisfaction. The MovieLens dataset, a widely recognized benchmark in recommendation system research, is ideal for developing and evaluating recommendation models due to its rich collection of user ratings, movie metadata, and tagging information. Collaborative filtering, one of the most popular techniques, leverages user-item interactions to identify patterns and make product recommendations based on comparable users' interests (Koren et al., 2009). By lowering dimensionality and revealing latent variables in the data, matrix factorization techniques like Singular Value Decomposition (SVD) enhance suggestions even more (Koren,2008).

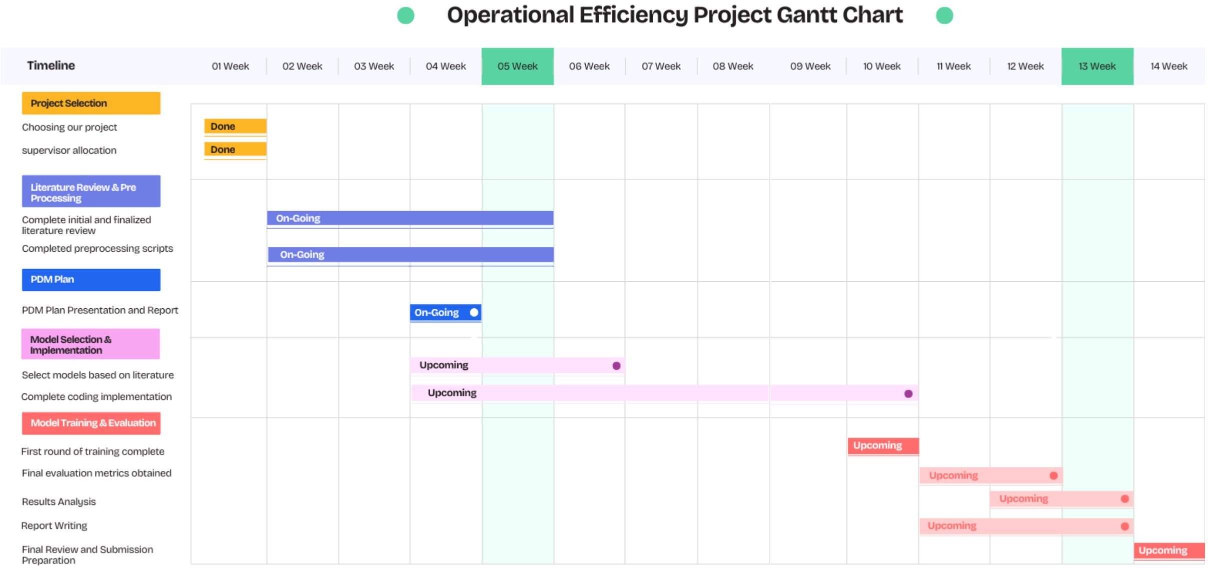
Data: The MovieLens dataset, a well-known standard in recommendation system research, served as the dataset for this study. It is maintained and made publicly available by the University of Minnesota's GroupLens research lab (Harper & Konstan, 2015). Time stamps, ratings (on a scale of 1 to 5), user IDs, and movie IDs are among the specific details concerning user-movie interactions that are included in the collection. It also offers movie metadata, such as release years, genres, and titles. The most widely used version of the dataset, which comes in various formats, includes 100,000 user ratings for 1,700 films. This sample size maintains computational economy while being enough for training and assessing machine learning models. The dataset is perfect for creating and evaluating recommendation algorithms including content-based filtering, matrix factorization, and collaborative filtering because of its rich content and organized nature.

Data Ethics: There is no personally identifiable information (PII) in the publicly accessible MovieLens dataset that was used for this project. It ensures that there are no ethical issues with the use of personal data or privacy by including anonymised user IDs, movie ratings, and metadata. There is no need for further anonymization because the dataset has already been gathered and shared for research purposes.

The project's main goal is to use this dataset to train and assess machine learning models like matrix factorization and collaborative filtering. Standard procedures like unit testing and cross-validation will be used for code testing, guaranteeing that assessments are carried out under strict control and without the involvement of outside human beings. This study does not require ethical approval from the University of Hertfordshire Ethics Committee, as it does not involve the use of personal data or the collection of new information from individuals. The dataset is publicly available, anonymized, and used solely for academic research, minimizing ethical concerns.

Project Plan: Literature review, dataset investigation, data preprocessing, model implementation, evaluation, visualization, system development, testing, report writing, and presentation preparation are among the main responsibilities that make up the project. To guarantee effective progress, these tasks overlap; for example, performing the literature review while examining the dataset.

There are 17-week milestones, such as finishing the dataset preprocessing by Week 4, implementing the model by Week 7, and submitting the final report by Week 15. Week 8's interim report, Week 15's final report submission, and Week 17's final presentation are all important evaluation points. For efficient project management, a Gantt chart will graphically depict the timetable, highlighting task durations and overlaps.



Data Management Plan: The MovieLens dataset will be gathered from the GroupLens publicly accessible source and saved locally in CSV format, with backups made on cloud storage platforms such as Google Drive to offer redundancy. To protect data and code, regular backups will be made every week for local storage and every two weeks for cloud storage.
GitHub will be used to manage version control, with at least weekly commits to monitor code development and preserve an unambiguous history of changes. To guarantee effective project management, the repository will be arranged into structured folders for data, code, documentation, and outcomes. Throughout the project lifecycle, this data management plan facilitates a seamless and well-organized workflow by guaranteeing safe storage, frequent backups, and methodical version control.

References: Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. 30-37 in Computer, 42(8).

1. Konstan, J. A., and Harper, F. M. (2015). The context and history of the MovieLens datasets. Interactive Intelligent Systems Transactions (ACM), 5(4), 1–19.

2. Y. Koren (2008). The neighborhood meets factorization: a multi-dimensional collaborative filtering model. The 14th ACM SIGKDD international conference on data mining and knowledge discovery proceedings.
3. Semeraro, G., Lops, P., & de Gemmis, M. (2011). Current developments and trends in content-based recommender systems. Handbook of Recommender Systems, 73-105.