Module 2: Group Task

Topic: Data for Al Project Simulation: Groups design a mock Al project (building a recommendation system) by detailing the data requirements, preparation steps, ethical concerns, and governance measures.

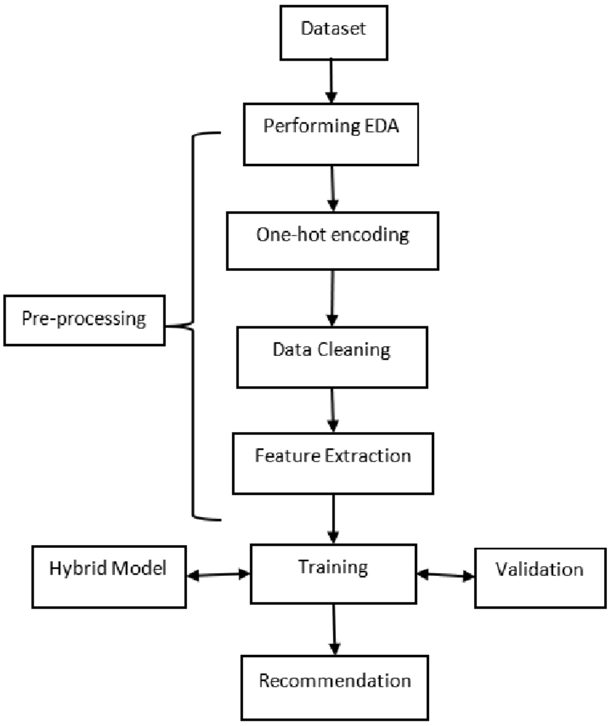
Data for AI Project Simulation – Movie Recommendation System

Introduction

In this simulation, we design a mock AI project in the form of a **Movie Recommendation System**. The objective of this system is to recommend movies to users based on their viewing history, ratings, search behaviour, and preferences. While the idea of recommending movies may seem simple, the process involves collecting structured and unstructured data, cleaning and transforming it, extracting meaningful features, training machine learning models, and implementing ethical and governance safeguards.

Data is the foundation of any AI project. Without high-quality and well-prepared data, even the most advanced algorithms cannot produce reliable results. Therefore, this simulation not only focuses on model development but also emphasizes data preparation, ethical considerations such as privacy and bias, and governance practices to ensure responsible AI usage.

Project Overview: Movie Recommendation System



- The goal of this AI system is to recommend movies to users based on their preferences, watch history, and ratings.
- The system learns patterns from past behaviour and predicts future interests.
- For example, if a user frequently watches action movies and rates them highly, the system will recommend similar action-based content. The system continuously improves as more user data becomes available.
- This type of AI system typically uses collaborative filtering, content-based filtering, or a hybrid approach.

1. Data Requirements

To build an effective recommendation system, different categories of data are required.

The system primarily needs user data, movie data, and interaction data.

User data includes information such as user ID, age group, watch history, ratings given to movies, and browsing activity. This data helps the model understand individual preferences and behaviour patterns.

Movie data includes features such as genre, director, cast, release year, duration, and overall ratings. These attributes help the system identify similarities between movies.

Interaction data captures user actions such as clicks, likes, time spent watching, and skipped content. This behavioural information is highly valuable because it reflects real user interests beyond simple ratings.

The data may exist in structured form (tables), semi-structured form (JSON logs), or unstructured form (text reviews).

2. Data Preparation Steps

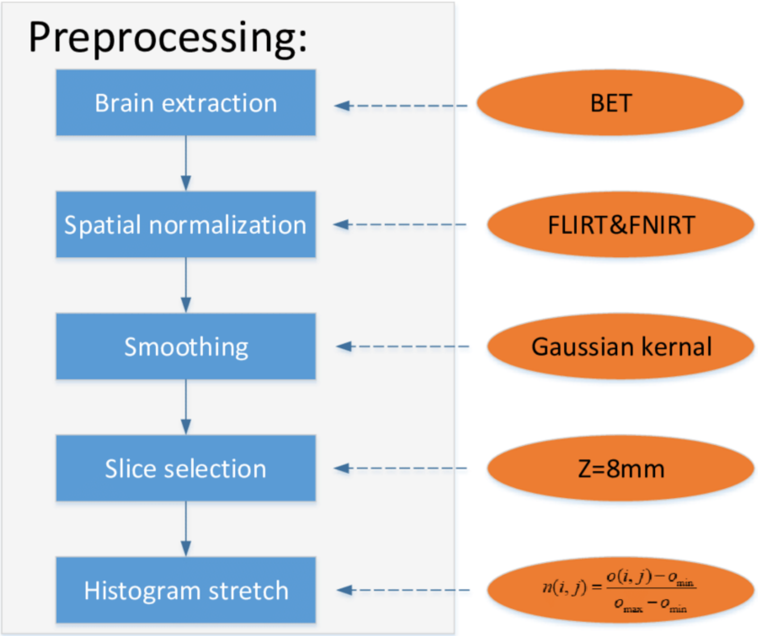
Raw data cannot be directly used for model training. It must undergo preprocessing to improve quality and usability.

The first step is data cleaning, where missing ratings are handled, duplicate records are removed, and inconsistent formats are corrected. For example, genre names must be standardized.

Next comes data transformation, where categorical features such as movie genres are converted into numerical values using encoding methods. Numerical data may also be normalized to maintain consistency.

Feature engineering plays an important role in improving model performance. New features such as "favourite genre," "average rating per category," or "recent viewing trend" are created to help the system better understand user preferences.

Finally, the dataset is divided into training and testing sets to evaluate model accuracy before deployment.



3. Ethical Concerns

Ethical considerations are essential when building AI systems.

Privacy is a primary concern because user watch history and personal information are sensitive. The system must collect data only with user consent and clearly explain its purpose.

Bias can occur if the system favours certain genres, languages, or demographic groups unfairly. This may reduce diversity in recommendations.

Transparency is also important. Users should understand why certain movies are suggested to them. Explainable AI techniques can improve trust.

Security measures must be implemented to protect data from breaches or unauthorized access.

4. Governance Measures

To ensure responsible AI development, governance policies must be applied.

Organizations should collect only necessary data and anonymize personal identifiers. Sensitive information should be encrypted during storage and transmission. Compliance with data protection regulations such as GDPR must be ensured.

Regular audits should be conducted to monitor fairness, detect bias, and maintain system accuracy. Clear access control policies should restrict data usage to authorized personnel only.

Proper governance builds trust and ensures that AI systems operate ethically and responsibly.

Conclusion

The Movie Recommendation System simulation demonstrates how data is the foundation of an AI project. From data collection and preprocessing to model development and ethical governance, each stage is crucial for delivering accurate and fair recommendations. Responsible handling of data not only improves personalization but also protects user privacy and promotes trust in AI-driven technologies.