pc

[Company name]  [Company address]

Project Deliverable 2 & 3: Data Preprocessing, Model and Technique Justification

Team Members:

Aun Ali (Member 1)

Abdullah Mazhar (Member 2)

Deliverable 2: Data Preprocessing

Aun Ali

Data Collection and Cleaning:

Collected raw data consisting of user profiles including attributes such as age, gender, height, weight, BMI, dietary preferences, fitness goals, exercise level, and meal details.

Cleaned data by dealing with missing values, handling outliers, and normalizing data such as height, weight, and BMI to maintain consistency across the dataset.

Feature Transformation:

Transformed categorical attributes such as "Dietary Preferences" and "Fitness Goal" into numerical representations for easier processing.

Combined various attributes (age, gender, BMI, etc.) into a more descriptive full-text representation of the user profile, which would be used for generating embeddings.

Text Preprocessing:

Tokenized the generated text descriptions of each user’s profile.

Removed stopwords, lemmatized words, and converted all text to lowercase to standardize the input for embedding generation.

Embedding Generation:

Used the SentenceTransformer library, specifically the MiniLM model, to generate high-dimensional embeddings for each user profile based on their full description. These embeddings capture semantic meaning that will later be used for recommendation tasks.

Deliverable 3: Model and Technique Justification

Abdullah Mazhar

Justification of Models:

DeepSeek: We chose DeepSeek as it is well-suited for our recommender system, as it is effective for extracting contextual information and patterns from user profiles. This model can efficiently handle large amounts of data and offer personalized recommendations based on user preferences and goals.

Two Additional Models:

LLaMA (Large Language Model): LLaMA is used for generating embeddings that capture user profiles in greater detail, allowing for more precise recommendations based on user similarities.

Omni: Omni model was selected for its ability to handle multiple types of data and complex user inputs, adding versatility to the recommender system.

Recommender System Technique:

The system utilizes contrastive learning to fine-tune the embeddings generated from user profiles. This technique helps in identifying similarities between users, thus enabling better personalization for fitness and nutrition recommendations.

HNSW (Hierarchical Navigable Small World) is used for similarity search, which allows us to quickly retrieve similar user profiles and suggest the most relevant meals and workouts.

Conclusion

We both Aun Ali and Abdullah Mazhar contributed to completing the data preprocessing pipeline and model selection for the recommender system. The first deliverable focused on preparing the dataset and transforming it into a format suitable for model training. The second deliverable justified our choice of models and techniques for ensuring the system is robust and capable of offering personalized recommendations. Each part of the project contributes to building a holistic system that delivers accurate and relevant fitness and nutrition suggestions.