Building a Health Recommendation System

1. Introduction

The task is to build a personalised health recommendation system that provides tailored health tips to users based on their profiles, which include features such as age, gender, and medical conditions (e.g., diabetes, asthma, hypertension). The system uses a machine learning model to identify users with similar profiles and recommends relevant health tips such as diet recommendations, exercise routines, and lifestyle modifications. This kind of recommendation system is essential for promoting healthy habits and delivering user-specific advice, particularly for managing chronic conditions.

2. Key Preprocessing Steps

- Handling Missing Data: Any missing values in critical fields (such as medical conditions or age) were imputed or dropped, depending on the extent of missingness. For this project, there were no missing values, so no values were dropped.
- Handling Duplicate Data: You would first identify any duplicate rows in the dataset, as duplicate records could distort the model's learning. After identifying them, you remove these duplicates to retain only unique user profiles
- Checking for outliers: In this analysis, we used a boxplot to identify outliers and found that there are none present, so there is no need for removal or transformation.
- Removing irrelevant features: In the preprocessing stage, we remove irrelevant features such as names, doctor details, billing amounts, and other non-essential attributes from the dataset. This helps streamline the data, ensuring that only relevant information is retained for building the health recommendation system. By focusing on pertinent features, we improve the model's efficiency and accuracy in generating recommendations.
- Encoding Categorical Variables: Since features like gender, blood type, medical conditions, and admission type are categorical, we applied Label Encoding to transform these into numerical values. For example, "Male" and "Female" were encoded as 0 and 1, and each medical condition was assigned a unique integer value.

3.Model Choice and Rationale

For the health recommendation system, we chose the k-Nearest Neighbors (k-NN) algorithm due to its effectiveness in identifying similar user profiles based on multi-dimensional features such as age, gender, medical conditions, and test results. k-NN operates on the principle of proximity, finding the nearest neighbours in the feature space by calculating distances . This allows the system to recommend health tips based on the preferences and characteristics of users with similar profiles.

Additionally, k-NN is inherently intuitive and straightforward, making it easy to implement and interpret. It doesn’t require extensive training since it is a lazy learner, meaning it stores all the training data and makes decisions based on the proximity of new data points to existing ones. This feature is particularly beneficial in the context of health recommendations, where user profiles can be dynamic, and updating the model with new user data is seamless.

Furthermore, using k-NN allows for the consideration of multiple dimensions in user profiles, enhancing the accuracy of recommendations by providing a more personalised approach. Overall, the combination of these characteristics makes k-NN a suitable choice for our health recommendation system, ensuring relevant and targeted advice for users based on their unique profiles.

4. Performance Metrics

The evaluation metrics for this recommendation system were primarily qualitative due to the nature of the task.

Health Tip Accuracy: One approach to evaluating the system's performance was to compare the top 3 health tips suggested by the model with established health guidelines. For example, for users with diabetes, common health advice like "monitor blood sugar" or "maintain a balanced diet" should appear frequently in the recommendations.

5. Theoretical Explanation of the Model

k-Nearest Neighbors is a straightforward algorithm used for classifying or predicting values based on the characteristics of nearby data points. It works by looking at how similar different data points are.

How Does k-NN Work?

1. Distance Measurement: k-NN measures how far apart data points are. The most common way to do this is by using Euclidean distance, which is like measuring the straight-line distance between two points.
2. Choosing 'k': The letter kkk stands for the number of nearest neighbours to consider. For example, if k=3k=3k=3, the algorithm looks at the three closest data points to make a decision.
3. Making Predictions:
    - For Classification: If you want to determine what category a new data point belongs to, k-NN checks the kkk closest points and sees which category is the most common among them. That's the category it assigns to the new point.
    - For Regression: If you want to predict a numerical value k-NN averages the values of the kkk closest points to make a prediction.

6. Suggested Improvements

- Collaborative Filtering allows the model to learn from the behaviours and interactions of other users, leading to more nuanced recommendations.
- More Data increases the model's ability to capture complex relationships between various factors that affect health outcomes, enhancing the personalization of tips.

- Hybrid Models combine the strengths of both content-based and collaborative filtering approaches, ensuring that the system accounts for both user profile similarity and patterns in collective user behaviour.

These improvements would not only make the system more robust but also ensure that users receive health tips that are both relevant and backed by broader user experiences.