Project Proposal

Team Members:

- MUHAMMAD SALMAN YOUNAS – **BSDSF22M001**
- USMAN AHMAD – **BSDSF22M014**

1. Project Title

Uncertain Symptom Clustering using Fuzzy Logic and Unsupervised Machine Learning

2. Problem Statement

In many medical cases, people describe their symptoms with terms like “moderate pain” or “occasional dizziness.” Traditional systems are not designed for situations like these and in most cases, machine learning assumes everything is structurally labeled. Yet, the data for studying real diseases is distorted, poorly labeled and can be very uncertain.

The project’s aim is to come up with a system that employs fuzzy logic to analyze uncertain symptoms and organizes them using unsupervised machine learning methods. The system is capable of spotting hidden connections and proposing groups of health conditions without relying on known outcomes for diseases.

3. Objectives

- To preprocess and structure vague or imprecise symptom data using fuzzy logic.
- To apply unsupervised learning methods (e.g., K-Means, Hierarchical Clustering, Fuzzy C-Means) to identify natural groupings in symptom data.
- To discover meaningful symptom clusters that can aid in early diagnosis or pattern recognition.
- To compare the clustering results with and without fuzzy preprocessing.

4. Proposed Methodology

Step 1: Dataset Collection & Preparation

Make use of freely available datasets featuring symptoms (e.g., symptom datasets by the WHO and Kaggle).

Convert labels such as low, moderate and high into numerical values.

Using membership functions, replace linguistic words with specific numerical values.

Step 2: Fuzzy Logic Integration

- Design fuzzy sets for key symptoms (e.g., temperature, fatigue, nausea).
- Implement Mamdani-type fuzzy inference to transform inputs into numerical vectors.

Step 3: Unsupervised Learning Techniques

Apply the following clustering models:

- **K-Means Clustering** (baseline)

- **Hierarchical Agglomerative Clustering (HAC)**
- **Fuzzy C-Means Clustering** (to handle overlapping clusters)

Step 4: Visualization & Evaluation

- Use dimensionality reduction (PCA or t-SNE) to visualize clusters.
- Evaluate cluster quality using internal metrics like:
    - **Silhouette Score**
    - **Dunn Index**
    - **Davies-Bouldin Index**

5. Dataset Description

The dataset will include symptom descriptions for a large number of patients, with or without final diagnoses. Possible sources

1. WHO or CDC reports
2. Kaggle Dataset
3. Augmented synthetic data using known symptom patterns

6. Expected Outcomes

- Grouping of patients with sharing similar but vague symptoms.
- Unclear inputs are made easier to understand through a fuzzy transformation process.
- Graphs illustrating similarities in the symptoms experienced by different groups of patients.
- Looking at how good the clusters are between fuzzy-based and non-fuzzy models

7. Timeline of Activities

Step	Task
1	Literature review, dataset search & cleaning
2	Define fuzzy sets and membership functions
3	Fuzzify symptom data
4	Apply K-Means and HAC clustering
5	Apply Fuzzy C-Means clustering
6	Evaluate and visualize clustering outcomes