Workflow :

1. Read Research Papers about the Topic

Explore the literature to have a complete and deep understanding of network theory and disease prediction

2. Exploratory Data Analysis

2a. Dataset

Choose a dataset between:

smaller: https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset?select=Symptom-severity.csv

larger, but artificially generated: https://www.kaggle.com/datasets/dhivyeshrk/diseases-and-symptoms-dataset?select=Final_Augmented_dataset_Diseases_and_Symptoms.csv

2b. Analysis

- Structure investigation
- One Hot Encoding (if needed)
- Check Missing Values
- Check number of distinct values
- Train, Test and Validation split
- Symptoms distribution for each disease

3. Network analysis

Decide the network structure (e.g. bipartite, weighted...). Define useful metrics, test their statistical significance and analyze their results. Retrieve communities.

3a. Structure:

1. Bipartite network with 2 type of nodes (symptoms and disease)
2. Weighted links (occurrence of a symptom in the disease: from 0 to 1 or just in absolute value)

3b. Node importance metrics:

1. **SS1**: Symptom Specificity. For each s, calculate the sum over d of all non-zero entries in the adjacency matrix, represented as $[\sum_{d} \text{nonzeroAdj}(s, d)]$. The lower the value, the higher the specificity.

2. **SO1**: Symptom Occurrence. For each s sum all the weights over d. Computed as $[\sum_{d} \text{nonzeroAdj}(s, d)]$.

3. **SC2**: Symptom Commonality: Measures if a symptom is present in diseases which are affected by many other symptoms or in disease which are affected by only few symptoms.

4. **DS1**: Disease Specificity. For each `d`, calculate the sum over `s` of all non-zero entries in the adjacency matrix, represented as [\sum_{s} \text{nonzeroAdj}(s, d)]. The lower the value, the higher the specificity.

5. **DO1**: Disease Occurrence. For each `d` sum all the weights over `s`. It tells how many times a disease occurs across the dataset. Computed as [\sum_{s} \text{nonzeroAdj}(s, d)].

6. **DC2**: Disease Commonality: Measures if a disease presents symptoms which affect many other diseases or symptoms which affect only few diseases.

   - Statistical Significance:

   Null Model with Random Network?

   - Plot the metrics distribution

     - Power Law distribution (Log-Log)
     - Beta coefficient
     - Z-score

3c. Community Detection

- Identify possible communities and similarities between diseases, this information could be useful in prediction explanation.
- Communities could have significant predictive properties.
- Modularity can be used to compare different partitions

4. Data cleaning

Remove outliers and fix invalid values.

5. Feature definition

Define which features will be used to make predictions. Network features

6. Model creation

Train different model with different parameters to find the best one

7. Comparison between models

Compare model with network features and the model without them.