KDD Project Progress Report

Team: Data Lakers

Team Members:

Dinesh Nannapaneni – G02553811
Guna Sekhar Reddy. G – G02549096
Jaswanth Kanderi – G02553807
Harsha Bhuvuna Rikith Devearasetty-G02554327

Project Progress Overview

Completed Tasks:

- Successfully loaded and cleaned raw datasets from Kaggle and Disease Symptom KB.
- Processed disease and symptom names, filling missing values and handling categorical data.
- Generated a dictionary mapping diseases to symptoms and calculated occurrence counts.
- Visualized disease and symptom distributions using histograms, bar plots, count plots, and violin plots.

Datasets:

We have used the below dataset it is a raw dataset we have replicated all the values the below raw dataset into excel sheet and the using that excel sheet we made the data cleaning process and visualization process

Disease Symptom KB Dataset: Disease Symptom Knowledge Base

Challenges:

While doing the project we have experienced some difficulties and those are following:

- **Data Quality Issues:** Missing and inconsistent values in both datasets required extensive cleaning and preprocessing.
  - *Solution:* Used forward fill and imputation techniques.
- **High Dimensionality of Symptoms:** Some diseases had numerous symptoms, making visualization and feature encoding complex.
  - *Solution:* Focused on top 10 symptoms and diseases for detailed analysis.
- **Initial Visualization Overlap:** Some visualizations were unclear due to overlapping data points.
  - *Solution:* Adjusted plot types (e.g., switching to violin plots) and refined axes for better readability.

Unresolved Challenges

- Need further exploration of class imbalance and scaling issues.
  - *Proposed Approach:* Implement SMOTE for oversampling in the next phase.

Collaboration:

Meeting Frequency: We are meeting in-person weekly for in-depth discussions and integration

and also we made a whatsapp group for team where we discuss about different problems we encountered while executing the project. And also we have been discussing about the project updates on completion of tasks.

Next Steps:

- The main agenda of this project is to train the data with different algorithms like Random Forest, K- Nearest Neighbor, and Support Vector Machine.
- Hyperparameter tuning and cross-validation.
- Final model evaluation and comparison.

Plan for Completion

- Follow the timeline outlined in the proposal, dedicating specific weeks to each task.

- Weekly integration and testing sessions to ensure alignment and progress tracking.

Potential Challenges

- Hyperparameter tuning may take longer than anticipated.
  - *Mitigation:* Automate using grid search or randomized search techniques.
- Possible underfitting/overfitting of models.
  - *Mitigation:* Perform cross-validation and regularization.