**Automated Disease Prediction Using Machine Learning Ensembles**

**Phase 1: Problem Definition and Data Understanding**

* 1. **Project Overview [Shreya Kulkarni]**

There are a number of project objectives which include enabling, the implementation of a machine-based system that will give predictions of given diseases like diabetes for prediction ensembles bridging the gap in healthcare towards predictive analytics. The term predictive analytics refers to the usage of data and statistical algorithms to identify the likelihood of a future outcome based on historical data. Given that modern society has become more dogmatic in providing care post discovery of diseases, it is important to adapt the mindset of preventing the onset of a disease at all, and for that, predicting disease occurrences will be crucial. One of the ways to improve intermediate outcome such as the possibility to provide extended range coverage of care is to improve prediction accuracy. Specific populations of individuals or the entire population of a particular geographical unit can be classified into “Low risk” or “High risk” categories in order to achieve the goal of encompassing in order to maximize the coverage of care provided.

The aim of the system is to enable health professionals and other individuals to avoid high-risk patients prior to taking action towards the disease enabling better decision-making and care not only for every individual but for society as well.

* 1. **Purpose of The Project [Vasavi C Kulkarni]**

**-Purpose:**

This project aims at developing a disease prediction system using machine learning that will identify patients who may develop diabetes among other diseases. The system is to enhance prediction quality and reliability through the application of ensemble techniques. The ensemble learning technique is the fundamental strategy of this project because it increases the accuracy of the prediction system by incorporating multiple models in making a decision, hence lessening the chance of overfitting or bias.

**- Target Users:**

This project is aimed at health practitioners, researchers and other individuals who seek to assess the risks of getting diseases. It also serves clinically for data scientists who are applying ensemble learning techniques in the area of medicine.

**- Possible Applications:**

* Early Diagnosis of Diseases: Help forecast diseases enabling the sick to get treatment on time.
* Tailoring Recommendations on Health: Assist people in initiating optimal health life practices.
* Clinical Decision Support: Aid the clinician to prioritize and channel the efforts to patients with higher chances of the prognosis.
* Research and Public Health: Enables studies to be done on common conditions and the application of preventive oriented approaches as well.

**1.3 Dataset Overview and Data Requirements [Radha Sangshetty]**

Most accurate predictions regarding a disease would be possible, if the dataset includes important clinical, demographic and behavioural risk factors for that disease.

**- Features:**

* Demographics: Age, sex, culture, area of residence.
* Clinical Features: BMI, glucose levels, blood pressures, cholesterol levels and other vital signs.
* Lifestyle Features: Levels of physical activity, eating habits, smoking/alcohol, sleeping habits.
* Medical Background Data: Previous medical conditions, familial diseases, and drugs taken before.

**- Labels:** The dataset will contain the binary labels: for instance, “0” to Low risk and “1” to High risk. This will be regarded as the target variable in the supervised tasks in modelling.

**- Dataset Format:** The dataset must be in a non-graph form such as a CSV, an Excel file or a database in SQL.

- Each row represents a single unit of observation while every column represents a unit of measurement for a different feature. The dataset can have datetime or categorical values which have to be encoded prior to being passed into the model.

* 1. **Data Sources [Sumeet U Pattan]**

Public repositories and proprietary systems can be utilized to get data in this project. Some of the potential sources include:

**- Public Datasets:**

- UCI Machine Learning Repository: Datasets pertaining to health care Systems in predicting diseases.

- Kaggle Datasets: Diabetes and hypertensive and different cardiovascular risk factors datasets.

- Government Databases: Nationally sponsored health researches and datasets for health.

**- Private Data:**

- Electronic Health Record (EHR): Medical records of patients from the clinics and hospitals.

- Wearable Technology: Devices like smart watches helps in collecting information.

- Insurance Data: Information obtained provides an insight into previous visits and treatments availed from different health practitioners.

**- Web Scraping:**

Health forums and portals: Any additional detail regarding ailments that can be found on the web.

Research Articles: Data that was sourced from the scholarly works on the subject.

* 1. **Initial Data Understanding [Shreya Kulkarni]**

Let us put together the dataset before we apply any methods. This is the phase in which one would determine the quality and structure of the data. Activities involve:

**- Missing Data:** Identify and apply appropriate imputation methods completing missing values (e.g., mean for numerical data, mode for categorical data).

**- Outliers:** There is a need to counter and take care of outliers for prediction accuracy.

**- Data Distribution:** Check important attributes' distributions to determine normalization or scaling is needed.

**- Correlation:** Examine correlations between features to look for relationships among them, and redundant variables.

**- Exploratory Visualizations:** Scatter plots, histograms, and heat maps are good approaches to getting the initial patterns and trends of the data.

* 1. **Preprocessing Objectives [Sumeet U Pattan]**

The main role of pre-processing is to change the raw data into a format suitable for the required machine learning applications. These include among others the key primary preprocessing steps:

- **Feature Scaling:** Where necessary use Min-Max scaling or Z-score to normalize numerical features.

- **Categorical Encoding:** Use one-hot encoders or other methods to encode categorical variables into numerical values.

- **Feature Selection:** This is a nominal filter used to eliminate non-useful or correlation intensive features thereby saving on dimensions and enhancing output of the model.

- **Data Transformation:** For skewed features, transformations like log transformations can be used to normalize the feature distribution.

* 1. **Conclusion of Phase 1 [Vasavi C Kulkarni]**

Phase 1 is an introductory phase of the project where the main purpose of the project, the end users, the data inputs and the location of the data are defined. This dataset will then be prepared by cleaning and analysing the clinical, demographic and lifestyle features together with the fusion machine learning methods that will be used later on. This will be the foundation of the prediction engine in which risk assessments for diseases will be thorough and accurate. Subsequent phases will include the construction, training, and assessment of the models developed.