**Using Machine Learning Models to predict Most probable disease/infection from symptoms and Lab tests data.**

**Introduction**

The project aims to leverage machine learning models for predicting the most probable disease or infection based on input symptoms and laboratory test data. By harnessing the power of data-driven insights, this initiative seeks to enhance disease diagnosis accuracy, contributing to more effective and timely healthcare interventions. The integration of symptoms and laboratory test results will provide a holistic approach to disease prediction, catering to a diverse range of medical conditions.

**Objectives**

**Develop a Predictive Model**

Create a robust machine learning model capable of predicting diseases or infections with a high level of accuracy based on input symptoms and laboratory test data.

**Enhance Diagnostic Accuracy:**

Design the model to handle a wide array of medical conditions, ensuring its applicability across different disease categories and patient profiles.

**Integrate Symptoms and Lab Test Data**

Develop a seamless integration engine to incorporate both symptoms and laboratory test data into the machine learning model, providing a comprehensive and accurate basis for predictions via a GUI.

**Methodology**

**Data Collection:** The initial phase of this research project involves a meticulous approach to data collection, aggregation, and preparation. A comprehensive strategy was devised to source data from reputable and sanctioned health entities and public medical repositories, including PubMed, the U.S. National Library of Medicine (NLM), Public Health Organizations (predominantly WHO & CDC), health surveys conducted by the National Institute of Health (NIH), and research publications available in Electronic Health Records and open data platforms such as data.gov, Kaggle, and the Journal of Medicine.

The data collection process involved manual extraction of symptoms from webpages, downloading CSV files, and establishing connections via API links for platforms like data.gov and Kaggle. Upon collecting the data, the aggregation process ensues, where redundant disease prognoses are systematically reconciled through manual validation

The final dataset encompasses 133 columns, with 132 columns representing symptoms and the final column denoting the prognosis.

**Data Processing & Cleansing**: To guarantee optimal results, the dataset undergoes meticulous cleaning before its application in model training. All columns within the dataset are numeric, except for the target column (prognosis), which is of string type. Employing a label encoder, it is transformed into numerical form. Furthermore, thorough checks for data balance, outliers, missing values, and duplicates were also implemented during the dataset cleansing process.

**Training Model Development**: Following the completion of data collection and cleansing, the dataset is prepared for machine learning model training. In this context, the refined data is utilized to train three specific classifiers: Support Vector Classifier, Naive Bayes/Gaussian Classifier, and Random Forest Classifier. The assessment of model quality is conducted using a confusion matrix, providing a comprehensive evaluation of the classifiers' performance.

**Inference:** Following the training of the three models, disease prediction was executed for input symptoms by consolidating the predictions from all three models using the mode. This enhances the overall prediction's robustness and accuracy hence making the model more generalized while reducing bias.

**Finalization**: In conclusion, a REST APIs portal will be used to receive symptoms as input, predict the disease based on the symptoms using the trained models, and return predictions. This comprehensive approach ensures an effective disease prediction model leveraging various classifiers and a meticulous data preparation process.