Capstone project - 2 Proposal

Project Title: Disease Prediction Using Machine Learning
https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning

Student: Lakshika Jain

Context

The ability to accurately predict diseases using technological advancements such as machine learning can significantly improve the treatment and patient outcomes. I aim to apply the ML models to the dataset from Kaggle which includes 132 parameters across 42 diseases, to develop a robust system for predicting not only the presence of illness but also identifying the specific diseases.

Problem Statement

In today's healthcare industry, we face challenges in early diagnosis due to the complexity of symptoms and their correlation with multiple diseases. This project will focus on creating a ML model that can

1) Predict whether a person is well or not based on their medical parameters
2) Identify the specific disease from which a person may be suffering if they are predicted to be unwell.

Objectives

- Implement various machine learning algorithms to predict the wellness of an individual.
- Classify the specific type of disease an individual might have from a set of 42 possible diseases.
- Evaluate and compare the performance of different ML models based on accuracy, precision, F-1 Score and recall.

Dataset Description

This dataset from Kaggle includes data on 132 different health parameters collected from the individuals diagnosed with one of the 42 diseases. These parameters include, but are not limited to, demographic information, symptom presence, medical tests results and lifestyle information.

Scope of solution space

This model can help predict serious diseases by classifying illness into different categories for each disease. It can help physicians diagnose and treat serious, life-threatening diseases sooner thus improving the outcomes of the treatment. This tool could potentially assist healthcare professionals in personalized treatment planning.

Methodology

- Data Preprocessing: Explore augmentation techniques. Handling missing data, normalizing/scaling features, and encoding categorical variables.
- Exploratory Data Analysis (EDA): Statistical analysis and visualization of the data to understand distributions, correlations, and patterns.
- Model Selection: Exploring different models such as Logistic Regression, Decision Trees, Random Forest, and Neural Networks.
- Model Training and Validation: Using cross-validation techniques to train and tune the models.
- Performance Evaluation: Comparing models using metrics such as accuracy, precision, recall, and F1-score.

Criteria for Success:

- Accuracy of Predictions: The machine learning models should demonstrate high accuracy in predicting whether a person is well or not and in diagnosing specific diseases. Metrics like accuracy, precision, recall, and F1-score are typically used to evaluate performance.
- Model Robustness: The models should perform well across a variety of scenarios and be robust against overfitting. This can be assessed through cross-validation and external validation on separate data sets.
- Data Handling: Effective management and preprocessing of data, ensuring that all data quality issues are addressed, including handling of missing values, normalization, and encoding of categorical variables.
- Usability in Real-World Settings: The model should be practical and scalable for real-world applications, meaning it can be integrated into healthcare systems without requiring excessive computational resources.

Constraints

One of the constraints I see in this analysis is the size of the dataset, this is a small dataset and having more values could help train the model better.

Stakeholders

- Academic advisor: Vinit Koshti
- Data Science student/researcher: Lakshika Jain

Deliverables

- There will be code in Jupyter notebooks for each step in the methodology.
- And a slide deck explaining the scope of the step, methodology of each step, findings etc.
- All of the code and slide decks would be published on the Github repository for this capstone project.