PREDICTING A DISEASE FROM SYMPTOMS USING MACHINE LEARNING

Abstract

With the current world situation, environment and lifestyle affects an individual's health greatly and makes them prone to various diseases. In the rat race to provide for themselves and their families, individuals experience stress at both home and workplace (Kanakaraddi et al., 2021). This has an adverse effect on their health, leading to majority of untimely diseases. However, advancement in technology enables early detection of diseases such as arthritis, asthma. The early detection and prevention of diseases is of utmost importance with the ongoing stressful lives that people are experiencing and can help prevent many life threatening diseases (Kanakaraddi et al., 2021).

The main aim of this project is to solve health-related issues by supporting doctors to more accurately predict and diagnose diseases at an early stage using machine learning techniques, based on symptoms that patients may be experiencing, making timely treatment a possibility which benefits patient care (Grampurohit & Sagarnal, 2020). For example, an early diagnosis is essential for effective therapy in the case of cancer.

The dataset to be used for the purpose of this project is obtained from the following websites:

(1) https://impact.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html (The New York Presbyterian Hospital provides a database of health symptoms that individuals experienced along with the corresponding prognosis admitted during 2004. (*Disease*, n.d.))

(2) https://github.com/anujdutt9/Disease-Prediction-from-Symptoms

(3) https://www.kaggle.com/datasets/neelima98/disease-prediction-using-machine-learning

The size of the dataset is 133 columns and 4920 rows (Grampurohit & Sagarnal, 2020). The proposed model will take number of symptoms as input and give the probability of a disease as output (Gomathy, 2021). The aim is to provide good results through higher accuracy and precision. The quality of dataset plays an important role in conducting predictive analysis. The dataset has been bifurcated into training and testing data. The model will be trained through training dataset and the result will be tested on testing dataset. In this dataset, the 133 columns are split as follows:

1. Symptoms (132 columns e.g. stomach pain, vomiting, fatigue etc.)
2. Prognosis of diseases (1 column consisting of 41 diseases as categorical values) such as Dengue, Diabetes, Heart attack etc.

The rows consist of continuous variables i.e. dummy variable taking the value of 0 (false) and 1 (true). The diseases listed in this dataset range from communicable diseases (spread from one person to another) to non communicable diseases (do not spread to others).

The project will emphasize on the technique of Classification and Regression, Data Mining and Knowledge Discovery. Classification model will be used to predict the type of disease (as a dependent column) from the type of symptoms (as independent columns). For predictive modeling, the proposed algorithms to be used are (1) Random Forest, (2)
Naïve Bayes, (3) Decision Tree and (4) K-nearest neighbor. These algorithms will be implemented using Python programming language (numpy, pandas, matplotlib, seaborn).

Keywords: Early detection, Training and Testing dataset, Classification and Regression, Data Mining and Knowledge Discovery

References:

Kanakaraddi, S. G., Gull, K. C., Bali, J., Chikaraddi, A. K., & Giraddi, S. (2021b). Disease Prediction Using Data Mining and Machine Learning Techniques. In *Lecture notes on data engineering and communications technologies* (pp. 71–92). Springer International Publishing. https://doi.org/10.1007/978-981-16-0538-3_4

Grampurohit, S., & Sagarnal, C. (2020). Disease Prediction using Machine Learning Algorithms. In *2020 International Conference for Emerging Technology (INCET)*. https://doi.org/10.1109/incet49848.2020.9154130

Disease. (n.d.). https://impact.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html

Gomathy, C. K. (2021). THE PREDICTION OF DISEASE USING MACHINE LEARNING. *ResearchGate*.https://www.researchgate.net/publication/357449131_THE_PREDICTION_OF_DISEASE_USING_MACHINE_LEARNING