Concept Note Submission

Nikhil Mahajan

TOPIC: Disease prediction from various symptoms

Concept of the project -

Accurate and on-time analysis of any health related problem is important for the prevention and treatment of the illness. The traditional way of diagnosis may not be sufficient in the case of a serious ailment. Developing a medical diagnosis system based on machine learning (ML) algorithms for prediction of any disease can help in a more accurate diagnosis than the conventional method. We have designed a disease prediction system using multiple ML algorithms. The dataset used had more than 230 diseases for processing. Based on the symptoms, age, and gender of an individual, the diagnosis system gives the output as the disease that the individual might be suffering from.

Problem Statement –

Build a Predictive Mode using Machine Learning so as to predict which disease to patient then proper medicine given by doctor.

The objective of the Project:

Medicine and healthcare are some of the most crucial parts of the economy and human life. There is a tremendous amount of change in the world we are living in now and the world that existed a few weeks back. Everything has turned gruesome and divergent.

Data sources used-

https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning

Data Analytics software used

Python & Jupyter Notebook Libraries used:

- Numpy- solve complex mathematical problems
- Pandas-use for dataframe manipulation
- Seaborn-to create data visualization
- Matplotlib- to create data visualization
- ipywidgets- Interactive analysis
- sklearn- Implement complex machine learning algorithm

Machine Learning Algorithms used

- Weighted KNN
- Fine KNN

Methodology –

From an open-source dataset, an excel sheet was created where we listed down all the symptoms for the respective diseases. After which depending on the diseases, age and gender were specified as a part of the dataset. We listed down around 230 diseases with more than 1000 unique symptoms in all. The symptoms, age, and gender of an individual were used as input to various machine learning algorithms. 3.1 K-nearest neighbours (KNN) The K-nearest neighbours (KNN) algorithm used is a type of supervised machine learning algorithm. It simply calculated the distance of a new data point to all other training data points. The distance can be of Euclidean or Manhattan type. After this, it selects the Nearest data points, where K can be any integer. Lastly, it assigns the data point to the class to which the majority of K data points belongs.
It is a modified version of KNN. In KNN we chose an integer parameter K and by using that parameter we found where the major predicted values lied. But if the value of K is too small the algorithm is much more sensitive to the points that are outliers. Also, if the value of K is too large then all the points that are almost very close to the K value are selected. To overcome this issue the weighted KNN gave more weight to the points that were nearest to the K value and the less weight to the points that were farther away. We were able to get the highest accuracy using this model. Also among all the KNN models, this model gave us the best results.

Probable Outcome

Different machine learning models were used to examine the prediction of disease for available input dataset. We used 11 different ML models for the prediction. Out of the 11 models we managed to get 50 % or above accuracy for 6 models. As shown in Figure 4, among all the models, we gained the highest accuracy for the Weighted KNN model of 93.5 %

Github link –

https://github.com/nikhilMahajan715/IBM_Intership.git