Disease Prediction Using Symptoms and Expert Recommendation System

Abstract

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 of an individual, the diagnosis system gives the output as the disease that the individual might be suffering from. The Random Forest algorithm gave the best results as compared to the other algorithms. The accuracy of the weighted Random Forest algorithm for the prediction was 98.3 %. Our diagnosis model can act as a doctor for the early diagnosis of a disease to ensure the treatment can take place on time and lives can be saved.

Keywords Disease prediction · machine learning · symptoms.

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 42 diseases with their unique symptoms in all. The symptoms, age, and gender of an individual were used as input to various machine learning algorithms.

Naive Bayes

It is a machine learning algorithm for classification prob lems and is based on Bayes' probability theorem. The primary use of this is to do text classification which in volves high dimensional training data sets. We used the Bayes theorem that can be defined as:

$P(h|d) = P(d|h) \cdot P(h) \; P(d)$

Where P(h|d) is the probability of hypothesis h

given the data d. This is called the posterior probability. P(d|h) is the probability of data d given that the hypothesis h was true. P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h. P(d) is the probability of the data (regardless of the hypothesis).

Support Vector Classifier

SVC, or Support Vector Classifier, is a supervised machine learning algorithm typically used for classification tasks. SVC works by mapping data points to a high-dimensional space and then finding the optimal hyperplane that divides the data into two classes. Sklearn SVC is the implementation of SVC provided by the popular machine learning library Scikit-learn.

The limitation of SVC is compensated by SVM non-linearly. And that's the difference between SVM and SVC. If the hyperplane classifies the dataset linearly then the algorithm we call it as SVC and the algorithm that separates the dataset by non-linear approach then we call it as SVM.

Random Forests

Random Forest developed by Leo Breiman is a group of un-pruned classification or regression trees made from the random selection of samples of the training data. Random features are selected in the induction process. Prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble. Each tree is grown as described in: By Sampling N randomly, If the number of cases in the training set is N but with replacement, from the original data. This sample will be used as the training set for growing the tree.

- For M number of input variables, the variable m is selected such that m<<M is specified at each node, m variables are selected at random out of the M and the best split on these m is used for splitting the node. During the forest growing, the value of m is held constant.

Logistic Regression

Logistic regression sometimes called the logistic model or logit model, analyzes the relationship between multiple independent variables and a categorical dependent variable, and estimates the probability of occurrence of an event by fitting data to a logistic curve.There are two models of logistic regression, binary logistic regression and multinomial logistic regression. Binary logistic regression is typically used when the dependent variable is dichotomous and the independent variables are either continuous or categorical. When the dependent variable is not dichotomous and is comprised of more than two categories, a multinomial logistic regression can be employed

Results and Discussion

Different machine learning models were used to examine the prediction of disease for available input dataset. We used 3 different ML models for the prediction. Among all the models, we gained the highest accuracy for the Random Forest Model of 98.2 %. The accuracy of Randomforest is high because the ensemble of decision trees has high accuracy because it uses randomness on two levels:

The algorithm randomly selects a subset of features, which can be used as candidates at each split. This prevents the multitude of decision trees from relying on the same set of features, which automatically solves Problem 2 above and decorrelates individual trees.

Each tree draws a random sample of data from the training dataset when generating its splits. This introduces a further element of randomness, which prevents the individual trees from overfitting the data. Since they cannot see all of the data, they cannot overfit it.

Model	**F1 Score %**	**Accuracy %**
Naïve Bayes	85.1	86.1
Logistic Regression	88.7	89.7
Random Forest	98.0	98.2
Support Vector Classifier	94.2	94.3

Team Members:

G. Vijith Pramod (160719733077)

M. Sathvika (160719733076)

K. Sreelekha (160719733067)

Under the Guidance of:

Dr. Diana Moses,

Assistant Professor