

Automated Disease Prediction System Using Machine Learning

Abit Mon Rajan
Master of Computer Applications
Amal Jyothi College of Engineering
Kottayam, Kerala, India
abitmonrajan@mca.ajce.in

Dr.Bijimol T K
Master of Computer Applications
Amal Jyothi College of Engineering
Kottayam, Kerala, India
tkbijimol@amaljyothi.ac.in

Abstract—Due to poor dietary habits, inadequate sleep, and a lack of daily activity, the majority of individuals in the modern world are disease-prone. Making a precise prognosis based on symptoms is too difficult for a doctor, though. Making an accurate diagnosis of a condition is the most difficult undertaking. The amount of data in the medical sciences increases significantly every year. The expansion of data in the medical and healthcare industries has led to an increase in the accurate analysis of medical data, which has profited from early patient treatment. Using machine learning, created a method to forecast diseases. These systems will output the potential disease that an individual may have based on their symptoms.

Keywords — Machine Learning, Random Forest, Decision Tree, K Nearest Neighbor (KNN), Naive Bayes.

I. INTRODUCTION

The most prevalent health conditions frequently have certain fundamental signs that people typically display. For instance, a person with a headache may also display a number of other diseases' symptoms. We rely on doctors in these situations where we demand an immediate diagnosis based on the symptoms. The type of the disease can be predicted using a machine learning model. The predictions of the model may enable early disease detection and faster diagnosis.

We used an online dataset for this research's purposes in order to make predictions based on machine learning about some test data. Building a framework and computational model for disease prognosis using a variety of signs is the study's main goal. We employ a dataset that details the symptoms of the 42 most prevalent ailments as well as information about their incidence rates for this aim. The used dataset required extensive cleaning and preprocessing. Transposing the datasets into a form where diseases were the target column and each symptom was a dummy variable for the diseases on which models were to be trained was the first step. Four separate sets of algorithms were implemented in the study to map symptoms to diseases. In order to assess the generality of the algorithms to be trained, we first divided the dataset into training and testing sets. We used two-thirds of the dataset for learning and the remaining third for testing. An overfitted model for the study was found to be more appropriate for later phases, therefore testing was carried out on the complete dataset.

II. LITERATURE REVIEW

Machine Learning technologies are developing increasingly beneficial within the medical field. To forecast diseases, numerous Various machine learning and deep learning techniques and algorithms have been used by researchers. There are many different algorithms, including the Gradient Boost classifier algorithm, the Ada Boost algorithm, the perceptron, the Linear Discriminant Analysis algorithm, the Logistic Regression algorithm, the Gaussian Nave Bayes algorithm, the Decision Tree Classifier, the Extra Tree Classifier, Random Forest Classifier, and the Linear Discriminant Analysis algorithm.

In this study, Random Forest, Decision Tree, Naive Bayes, and KNN are four machine learning algorithms that are used. Decision Tree, Random Forest, Naive Bayes, and KNN all provide accuracy results of 90%, 95%, 92%, and 90% respectively. The dataset comes from Kaggle and contains about 4920 rows of information. Disease diagnosis is regarded as a difficult subject for quantitative research because it depends on numerous elements, each of which varies accordingly.

Sayali Ambekar, Rashmi Phalnikar, [1] recommended Disease Risk Prediction and performed the task use a neural network for convolution. In this study, machine learning techniques such the CNN-UDRP algorithm, Naive Bayes, and KNN algorithm are used. The system employs structured data for training and can achieve an accuracy of 82% by using Naive Bayes.

Pahulpreet Singh Kohli and Shriya Arora, [2] suggested disease prediction by techniques including Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, and Adaptive Boosting were used using machine learning applications and methods. The prognosis of diabetes, breast cancer, and heart disease is the main topic of this essay. The best accuracy rates are produced by logistic regression, and they are 95.71% for breast cancer, 84.42% for diabetes, and 87.12% for heart disease.

Rati Shukla, Vikash Yadav, Parashu Ram Pal and Pankaj Pathak [3] suggested using machine learning methods such as Decision Tree, Support Vector Machine, Random Forest, Naive Bayes, Neural Network, and KNN, we predict and detect breast cancer. The Support Vector Machine provides

outcomes in this system that are more accurate than any other algorithm. Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava, [4] suggested using machine learning methods such as Decision Tree, Support Vector Machine, Random Forest, Naive Bayes, Neural Network, and KNN, we predict and identify cardiac disease. This system has an accuracy rate of 88.47%.

Deeraj Shetty, Kishor Rit, Sohail Shaikh and Nikita Patil, [5] used Naive Bayes and KNN algorithms to research the applications of data mining for predicting the development of diabetic disease. This system can predict diabetes, and KNN's accuracy is higher than Naive Bayes.

III. MOTIVATION

It is clear that the desire for health information is changing how people look for information on a global scale. For many people, it can be challenging to find health information online about ailments, diagnoses, and treatments. If a recommendation system for doctors and pharmaceuticals can be developed using review mining, it will save a lot of time. The fact that these users of technology are laypeople makes it difficult for them to comprehend the various medical jargon. The user becomes perplexed because there is a lot of medical information available in a variety of formats. The recommender system's objective is to adjust to the particular user-related requirements of the health sector.

IV. METHODOLOGY

The goal of this study is to develop a machine learning model that accurately predicts diseases from the provided dataset. Two training and testing data sets are created from the dataset. For the model to learn accurately, more data should be used during training. A disease prediction model can be made using the model's output. To predict the disease, the Naive Bayes, Decision Tree, Random Forest and K Nearest Neighbor (KNN) algorithms will be employed. The organisation is:

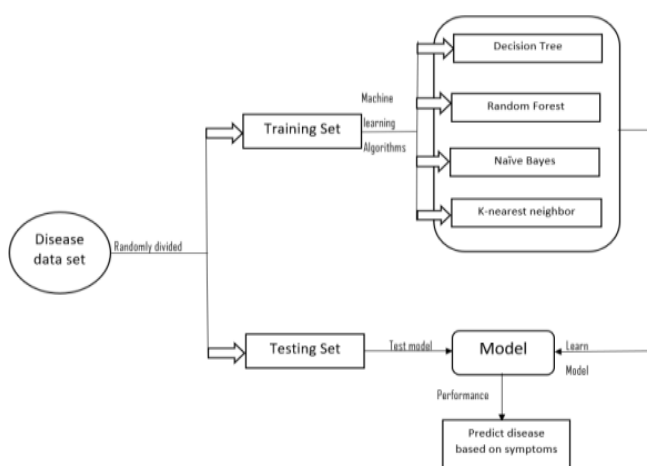


Figure 1: Disease Prediction System Methodology

The tasks that must be taken are:

- A. Data Collection
- B. Data Pre-processing
- C. Model Building
- D. Analyzing
- E. Result

A. Data Collection:

For this system, we have used a structured format for the dataset. The dataset being used contains each disease's name and the symptoms that go with it. Because this method relies on supervised machine learning, the dataset is labelled with 0 or 1. After then, the dataset was divided into a training dataset and a testing dataset. Our models were built using training datasets, which we subsequently utilised to train all of our machine learning algorithms on. In order to assess this trained model's accuracy, we lastly provided a testing dataset.

B. Data Pre-processing:

Every machine learning technique starts with this phase. Data transformation, data cleaning, and data minimization are all part of this process. To increase the effectiveness of the data, certain actions are taken. We can process the data to improve the precision of our model to guarantee that the classification is correct.

C. Machine Learning:

To forecast disease, this is employed. In order to check the news provided, machine learning techniques are employed to detect sickness. To find the disease, various classifiers are available. The accuracy of the classifier depends on how it was trained. High accuracy should be required of classifiers.

1. Decision Tree:

This particular supervised machine learning technique focuses primarily on categorization issues. Making a practical training framework for forecast the class or values using decision trees is mostly to achieve the required value, which are used to learn simple decision procedures using data already available (training data). In the decision tree algorithm, we begin by predicting the class at the tree's root. We combine the root characteristic's values with the trait of the data. We advance to the next node using the branch that is parallel to that value and go ground on the basis of differentiation. Decision Tree divides the symptoms in this system according to its classification, which reduces the complexity of the dataset. Decision trees are best described graphically using machine learning algorithms. Without the need of several parametric structures, it manages large and complex datasets. A decision tree model is chosen with the aid of training datasets, and a validation data gathering determines the right tree size to provide the best final model.

2. Random Forest:

The supervised machine learning algorithm Random Forest falls into this group. Although it is employed for both classification and regression, classification problems are its main focus. Random Forest is incredibly simple to use and very simple to implement. If necessary to create a model quickly, Random Forest is a fantastic replacement. During the training phase, a large number of decision trees are constructed using the Random Forest ensemble learning technique. Through voting, the best option is chosen. Multiple decision trees make up Simply Random Forest. A forest of trees is produced.

The accuracy rate is precisely related to the number of trees in the forest, which also solves the overfitting issue. Because Random Forest is not dependent on overfitting and is insensitive to dataset noise, it performs well when applied to real-world issues. It performs admirably and outperforms other tree-based algorithms in terms of execution. Bootstrap

aggregation or bagging are the main techniques used for tree learning.

With responses $Y = m_1, m_2, m_3, \dots, m_n$, the bagging from $b=1$ to B is repeated for the given data, $X = m_1, m_2, m_3, \dots, m_n$.

3. Naive Bayes:

Naive Bayes is a form of probabilistic algorithm that uses the Bayes Theorem and probability theory to estimate the likelihood of diseases. The performance of a Naive Bayes algorithm is comparable to that of decision trees and other chosen classifiers. It is possible to drastically reduce the cost of computing. It is easy to construct and beneficial for big datasets. By computing the probability of an independent variable, Naive Bayes classifies the data. Complete transactions are placed in the high probability class after the probabilities for each class are computed. Naive Bayes performs superbly in a variety of challenging real-world problems. The benefit of using Naive Bayes is that it needs very minimal training data to assess the parameters relevant to categorization.

The decision-maker can benefit from Bayesian reasoning. Naive Bayes is portrayed as being based on probabilities. It is based on the Bayes Theorem and probability theory and attempts to predict the class value of an unexplored dataset. In a report, a learned Naive Bayes model records the likelihood.

4. KNN:

Using metrics of similarity, the supervised learning algorithm KNN categorises new data points. Since it is non-parametric, no assumptions are made on the underlying data. KNN is another example of a lazy learner strategy because it only stores the training dataset without attempting to extract any discriminative functions from it.

KNN is based on the concept of feature similarity approach, which assumes that homogenous objects exist nearby. It is also an instance-based learning method with a roughly local function. This approach is employed when there are non-linear decision-making divisions between classes and it can handle vast amounts of data. KNN is capable of handling the problem of function estimation and is resilient to noisy training data. KNN uses the distance function to determine how far a new data point is from each training point.

Assume that (y_j, d_j) contains data points with $I = 1, 2, \dots, m$.

Characteristic values are denoted by Y_j , while each j 's tag is designated by d_j . assuming d classes are present.

For all values of j , $d_j = 1, 2, 3, \dots, d$

Consider y to be a fact that hasn't yet been given a name. KNN technique will be used to find the tag class.

V. BUILD MODEL

1. Examining the criteria and the problem statement

Consider the challenge in terms of the predictions we wish to make and the kinds of observational data we have at our disposal to support those predictions. Predictions typically consist of a label or a target response; it could be a binary classification with a yes/no label, a associated with equipment classification with a category, or a real number (regression).

2. Assemble and purge the data

The next stage is to get the data from datasets or from any other data sources after determining what kind of historical data we have for prediction modelling.

3. Prepare the data for the ML application.

Put the data through a transformation so that the machine learning system can interpret it.

4. Create the model's Graphical User Interface (GUI).

The Graphical User Interface (GUI) is made to display output and accept input. There are five input text boxes with dropdown menus of symptoms in each of which the user can individually choose. The GUI is created using the Python Tkinter library. The disease is predicted in the output box when you click "Result." The medications are also described in the relevant field.

5. Train the model

The data must be divided into training and evaluation sets prior to model training in order to track how effectively a model generalises to new data. A pattern and mapping between the feature and the label will now be learned by the algorithm.

6. Analyze and raise model accuracy

An indicator of a model's performance on an unobserved validation set is accuracy. Evaluate the model using the validation sets in light of the present learning.

7. Test the model

Test the model with hypothetical data. The model is finished once the system begins operating correctly.

TABLE 1. TEST RESULT

Algorithm	Accuracy
Decision Tree	90.24
Random Forest	95.12
Naive Bayes	92.68
KNN	90.24

VI. RESULT

The findings of the suggested system, which outperforms the current method in terms of speed, accuracy, and dependability, are shown in this section. Implementing several machine learning techniques yields the outcomes. The machine learning classification methods decision tree, random forest, naive bayes, and KNN are developed using Python programming.

VII. CONCLUSION

The main objective of this study is to forecast the disease using a suitable machine learning algorithm and patient-provided symptoms. We were able to reach a mean accuracy of more than 95% using four machine learning algorithms in this study, which displays outstanding rectification and high accuracy as well as boosting the system's dependability for this task. With different models and parameters, this study shows how a machine learning algorithm can be used to predict disease. The major objective is to identify the disease early on and anticipate it.

REFERENCES

- [1] Sayali Ambekar, Rashmi Phalnikar, "Disease Risk Prediction by Using Convolutional Neural Network" IEEE, 978-1-5386-5257-2/18, 2018.
- [2] Pahulpreet Singh Kohli and Shriya Arora. "Application of Machine Learning in Disease Prediction" IEEE, 978-1-5386-6947-1/18, pp. 1-4, 2018.
- [3] Rati Shukla, Vikash Yadav, Parashu Ram Pal and Pankaj Pathak, "Machine Learning Techniques for Detecting and Predicting Breast Cancer" IJITEE, ISSN: 2278-3075, Volume-8, pp. 2658-2662, 2019.
- [4] Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" IEEE Access, DOI 10.1109/ACCESS.2019.2923707, pp. 81542-81554, 2019.
- [5] Deeraj Shetty, Kishor Rit, Sohail Shaikh and Nikita Patil, "Diabetes Disease Prediction Using Data Mining" IEEE, 978-1-5090-3294-5/17, 2017.