International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 6 ǁ 2022 ǁ PP. 173-176

Predicto - Symptomatic Disease Predictor using Machine Learning

Janhavi Deshmukh[1], Pallavi Bherde[2], Pranjal Gujar[3], Himani Mandhan[4]
(Department of Information Technology, Government College of Engineering, Amravati-444601 Maharashtra, India)

Abstract *- Nowadays, people are suffering from various diseases due to several factors such as environmental conditions, their living habits, sometimes genetic issues and many more. Thus, predicting the disease at the earliest and at an initial stage is very crucial. In such scenarios, the concept of Machine learning plays a very vital role. With the help of disease data, machine learning is capable of predicting the disease that the patient is suffering from. Predicto is a healthcare sector project. The wide adaptation of computer-based technology in the healthcare industry resulted in the accumulation of electronic data. Due to the increasing amount of data, medical doctors are facing challenges to analyze symptoms accurately and identify diseases at an early stage. Supervised machine learning algorithms can be used for disease diagnosis and aiding medical experts in the early detection of high-risk diseases.*

Keywords: *Machine Learning, Disease Prediction, Symptoms, Algorithms, Diseases.*

Date of Submission: 22-05-2022 Date of acceptance: 04-06-2022

I. INTRODUCTION

Evolution and development in medical AI applications have been increasing in today's world due to the substantial technical enhancement and the availability of large amounts of digital data. AI is gradually uplifting medical practice with the help of a number of algorithms available.

We are witnessing a tremendous amount of change in today's world and the world that existed before COVID times. As almost everything has turned virtual nowadays, board certified doctors who wish to practice online can eventually help patients via predicto. This disease prediction model aims to work as a doctor for the early diagnosis of a disease that could ensure timely treatment and simultaneously save lives. There may be situations where the doctor may not always be available, here the symptoms of an individual along with the age and gender can be given to the ML model for further prediction. This disease prediction system is extremely relevant and can be a boon to predict a disease without any physical or human contact.

Predicto can act as a doctor for the early diagnosis of a disease to ensure timely treatment and ultimately save lives. In this project, we aim to design a general disease prediction system based on symptoms of the patient. The users simply need to enter the symptoms that they have been observing as an input and the system in turn, will predict the disease that the user might be suffering from, along with some details of the predicted disease.

II. SYSTEM ARCHITECTURE

A. Dataset

Every machine learning model requires a dataset for training. In our project, we have used the dataset provided by Kaggle, a community of data scientist and machine learning practitioners. Dataset size is one of the parameter on which accuracy of the ML model depends. So, we tried to make sure the size of dataset should be as large as possible.

B. Architecture Overview

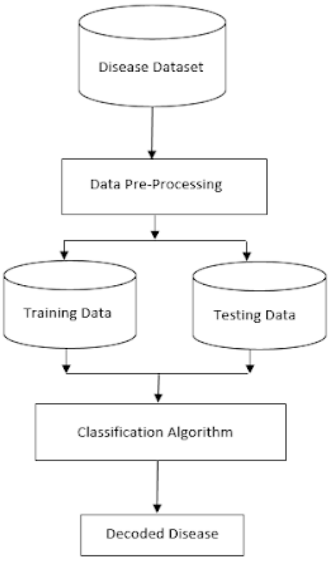
From an open-source dataset, an excel sheet was created where we listed down all the symptoms for the respective diseases. The proposed method for building the predictive model of disease prediction using symptoms proceeds as follows:

Dataset Exploration: The python environment can be used to explore the dataset consisting of symptoms and associated diseases along with data dictionary of the attributes involved.

Data Munging and Pre-processing: Data Munging or Wrangling refers to the estimation of missing values in some variables and is necessary as most of the interpretations cannot be done with missing data. The missing values are then substituted with the mean value in case of a continuous variable or it is substituted with the mode value in case of a categorical variable. Data preprocessing is working on data and manipulating the data so that it can go through the appropriate machine learning technique without any problem.

Training and Testing Data: The proposed model needs to be trained and tested under various conditions so that the model's correctness can be obtained. In addition, we consider that the model's accuracy is maximum. From the collected data, the 67:33 ratio will be used to train and test the model, respectively.

Fig 1 Architecture



Whenever necessary, there must be provisions to change or improvise on the algorithm that is being used. Furthermore, the model must adapt to new changes made in the dataset as dataset size will increase constantly. Data preprocessing would be needed to be performed again to the newly added data and include it to previously collected results. The model can then be retrained and checked for efficiency.

C. Methodology

The diagram shown below gives an insight of data flow in the system. The module is connected with the database PostgreSQL. Initially, the entity patient will visit our web application and if the user is new to the system then they need to register. If the user is an existing user of the system then the user can access the system by entering User ID and Password. The steps discussed till now are inescapable for all users. After that if the users wants to know whether they are suffering from any serious disease or not and if they want to know what type of disease it is, so they will simply enter the symptoms that they might be observing. These symptoms are fed to the train model. Trained model will also be getting input from data set and after processing the inputs, the system will show predicted disease as output. Then according to the output, the user can consult the specialist doctor.

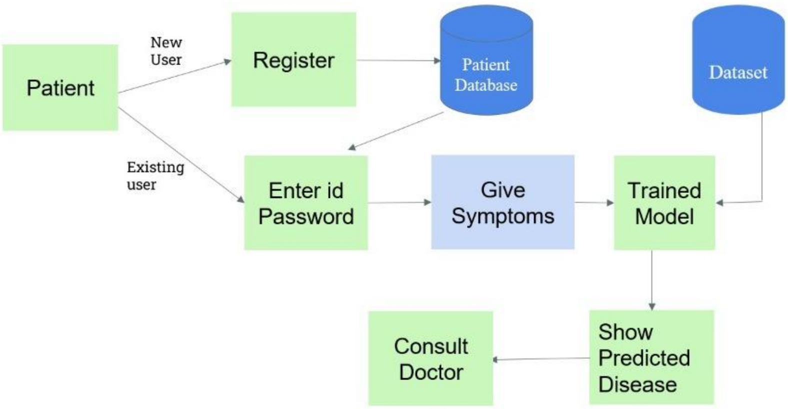


FIG 2 BLOCK DIAGRAM OF PROPOSED SYSTEM

III. MACHINE LEARNING ALGORITHMS

A. Naive Bayes: The Naïve Bayes algorithm is basically based on Bayes theorem and is a supervised learning algorithm used in a wide variety of classification purposes. It usually has a high-dimensional training dataset for dealing with classification problems. Naïve Bayes Classifier is an operational Classification algorithm that helps in the construction of effective machine learning models that can make quick and reliable predictions.

This algorithm predicts mainly based on the probability of an object and hence it is a probabilistic classifier. Bayes' Theorem is a mathematical formula that helps in calculating conditional probabilities. Conditional probability is a measure of the probability of an event occurring given that another an event has occurred.

The formula for Bayes Theorem is -

$$P\ (A|B) = \frac{P\ (B|A).P\ (A)}{P\ (B)}$$

This formula tells us how often A happens given that B happens, and is written as P(A|B) also called posterior probability, when we already know that how often B happens given that A happens, written as P(B|A) and how likely A is on its own, written as P(A) and how likely B is on its own, written as P(B).

Naive Bayes classifier is a term that refers to conditional independence of each of the features in the model, whereas Multinomial Naive Bayes classifier (a special type of Naive Bayes classifier) is a specific instance of a Naive Bayes classifier which uses a multinomial distribution for each of the features in the model. The multinomial Naive Bayes classifier is suitable for classification with a variety of features (e.g., word counts for text classification).

B. Decision Tree: Decision Tree is one of the type of Supervised Machine learning algorithm that can be used in classification and regression. The name "Decision Tree" comes from its structure which is almost similar to the tree, like it starts with a root node, which further expands to branches, contributing to a structure like tree. This splitting can be done using algorithms like Reduction in variance, Gini Index, Chi-Square and Information Gain.

The motive behind using a Decision tree is creating a training model which can be used for predicting the class or value of the target variable by learning simple decision rules concluded from training data.

$$\text{Entropy}(S) = -P(Y)\log P(Y) - P(N)\log P(N)$$

Where,

S stands for the total number of samples,
Y stands for the probability of Yes (Y)
N stands for the is probability of No (N)

C. KFold: KFold is a cross validation method, used to split a dataset into k equal folds or groups and evaluate different algorithms at each value of k.

In this method, for every k value, one group will be treated as test data and the remaining group will be used as training data. The model will be trained on a training dataset. On each iteration of cross-validation, a new model is trained independently of the model from the previous iteration.This can be helpful to avoid overfitting.

K value should be greater than or equal to 2. The most used k value is 10. If k is a very large value then it may result in less variance across the training dataset.

IV. RESULTS

The algorithm used in this model is Decision Tree as all other algorithms were found to be overfit. The accuracy of the model is 87.6% and the k value is 2.

This healthcare sector project is a system where users simply need to initially login with their details in order to predict the disease that the user might be suffering from, along with some details of the predicted disease.

The system consists of an input field where the users insert all the symptoms and the system will display the Patient name, Age of patient and Predicted Disease along with the Confidence Score as output and also provide an option to consult the respective doctor.

The users can also provide their valuable feedback for the upgradation of the system ensuring user friendly service to the users through the feedback forms and also consult board certified doctors through the "Consult a doctor" feature available in the system.

V. CONCLUSION

Predicto ensures early detection of high-risk diseases and mitigates suffering from various diseases. To train, validate and test the model, we used the data set, which consists of more than 40 diseases. These images are from Kaggle. Thus, a system like Predicto implemented using supervised machine learning algorithms, can be used for disease diagnosis and help in early detection of high-risk diseases.

Along with the prediction of the disease this system helps us to bridge the gap between patients and doctors with the help of "Consult a doctor" service available where the patient can online consult any of the listed doctors and book offline appointments in severe conditions.

ACKNOWLEDGEMENT

We would like to express our deep sense of gratitude and sincere thanks to **Prof. S. A. Lohi,** Department of Information Technology**,** Government College of Engineering Amravati, for his exemplary guidance and constant supervision. We are highly indebted and like to express our special gratitude to **Prof A. W. Bhade,** Head of the Information Technology Department, Government College of Engineering Amravati, for constant support and motivation.

REFERENCES

[1]. D. Dwivedi, "Data Science, Artificial Intelligence, Deep Learning, Computer Vision, Machine Learning, Data Visualization and Coffee".

[2]. Geeksforgeeks, "Disease Prediction Using Machine Learning"

[3]. Mrunmayi Patil , Vivian Brian Lobo , Pranav Puranik , Aditi Pawaskar , Adarsh Pai , Rupesh Mishra, " A Proposed Model for Lifestyle Disease Prediction "

[4]. Dhiraj Dahiwade, Prof. Gajanan Patle, Prof. Ektaa Meshram, "Designing Disease Prediction Model Using Machine Learning Approach"

[5]. Rahtara Ferdousi,M. Anwar Hossain, (Senior Member, IEEE),And AABDUL MOTALEB EK SADDIK, " Early-Stage Risk Prediction of Non-Communicable Disease Using Machine Learning in Health"

[6]. Mohammad Jamshidi, Ali Lalbakhsh , Jakub Talla, Zdenek Peroutka, Farimah Hadjilooei, Pedram Lalbakhsh, Morteza Jamshidi Luigui La Spada, Mirhamed Mirmozafari, Mojgan Dehgani, Asal Sabet, Saeed Roshani, Sobhan Roshani, Nima Bayat Makou, Bahare Mohamadzade, Zahra Malek, Alireza Jamshidi, Sarah Kiani, Hamed Hashemi-Dezaki and WahabMohyuddin,"Artificial Intelligence and COVID-19: Deep Learning Approaches for Diagnosis and Treatment"

[7]. Pahulpreet Singh Kohli, Shriya Arora, "Application of Machine Learning in Disease Prediction"