DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES

1st ArpitaSheshrajWavare
arpitawaware@gmail.com
Department of Computer Science
And Engineering
Shri Sai College of Engineering and
Technology,Chandrapur ,India

2nd Prof Pushpa Tandekar
p.tandekar@yahoo.in
Department of Computer Science
And Engineering
Shri Sai College of Engineering and
Technology,Chandrapur ,India

3rd Prof Ashish Deharkar
ashish.deharkar@gmail.com
Department of Computer Science
And Engineering
Shri Sai College of Engineering and
Technology,Chandrapur,India

ABSTRACT: The "Disease Prediction" method, which is concentrated on predictive modeling, it predicts the user's disease based on the symptoms that the user provides as input. The method examines the user's symptoms as input and returns the disease's likelihood as an output. Disease prediction is accomplished using the random forest classifier.

Keywords — Random Forest, Chronic Disease

INTRODUCTION: Machine learning is a branch of Artificial intelligence, which is basically concerned with the development of algorithms .The data obtained is then processed by the algorithm is designed to identify complex relationships thought to be features of the underlying mechanism that generated the data, and employ these identified patterns to make predictions based on new data. It is the machine learning task of concluding a function from labeled training data .This function should predict the correct output value for any valid input object. The training data contains training examples or training values .Various types of Machine Learning Techniques used are as follows

Supervised machine learning algorithm

At its most basic sense, machine learning uses programmed algorithms that learn and optimise their operations by analysing input data to make predictions within an acceptable range.. These three categories are: supervised, unsupervised and semi-supervised.

In supervised machine learning algorithms, a labelled training dataset is used first to train the underlying algorithm. This trained algorithm is then fed on the unlabelled test dataset to categorise them into similar groups. Using an abstract dataset for three diabetic patients, Fig. shows an illustration about how supervised machine learning algorithms work to categorise diabetic and non-diabetic patients. Supervised learning algorithms suit well with two types of problems: classification problems; and regression problems. In classification problems, the underlying output variable is discrete. This variable is categorised into different groups or categories, such as ‘red’ or ‘black’, or it could be ‘diabetic’ and ‘non-diabetic’. The corresponding output variable is a real value in regression problems, such as the risk of developing cardiovascular disease for an individual. In the following subsections, we briefly describe the commonly used supervised machine learning algorithms for disease prediction.

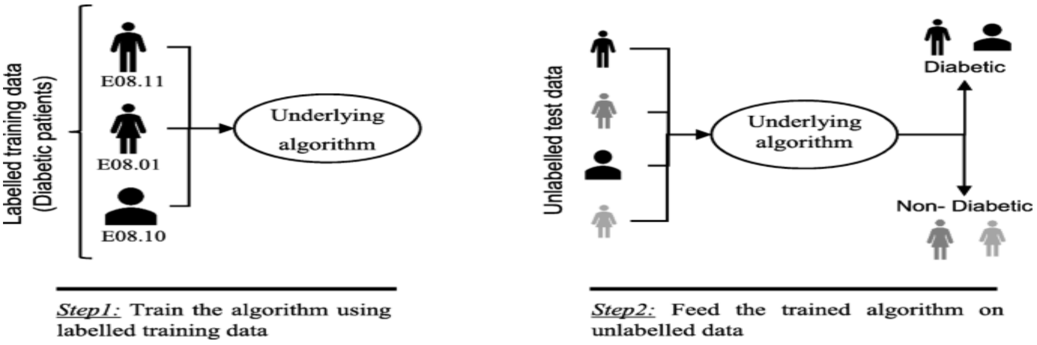
Labelled training data (Diabetic patients)
E08.11
E08.01
E08.10
Underlying algorithm
Step1: Train the algorithm using labelled training data
Unlabelled test data
Underlying algorithm
Diabetic
Non- Diabetic
Step2: Feed the trained algorithm on unlabelled data

Fig No.1 Supervised Machine Learning

Logistic regression

Logistic regression (LR) is a powerful and well-established method for supervised classification . LR helps in finding the probability that a new instance belongs to a certain class. Since it is a probability, the outcome lies between 0 and 1. Therefore, to use the LR as a binary classifier, a threshold needs to be assigned to differentiate two classes.

Support vector machine

Support vector machine (SVM) algorithm can classify both linear and non-linear data. It first maps each data item into an n-dimensional feature space where n is the number of features. It then identifies the hyperplane that separates the data items into two classes while maximising the marginal distance for both classes and minimising the classification errors.

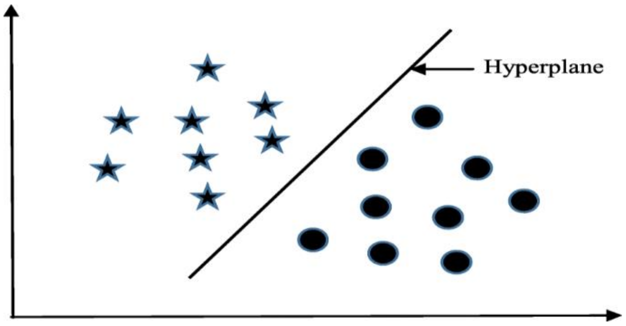

Fig No.2 Logistic Regression

Decision tree

Decision tree (DT) is one of the earliest and prominent machine learning algorithms. A decision tree models the decision logics i.e., tests and corresponds outcomes for classifying data items into a tree-like structure.

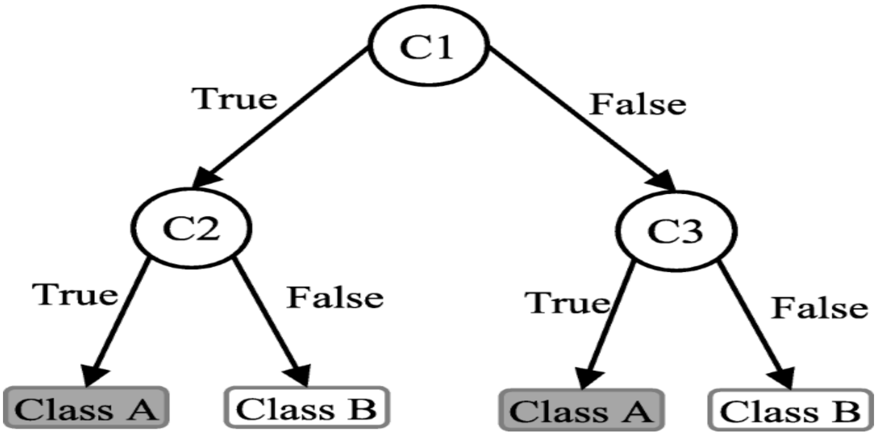

Fig no.3 Decision Tree

Random forest

A random forest (RF) is an ensemble classifier and consisting of many DTs similar to the way a forest is a collection of many trees

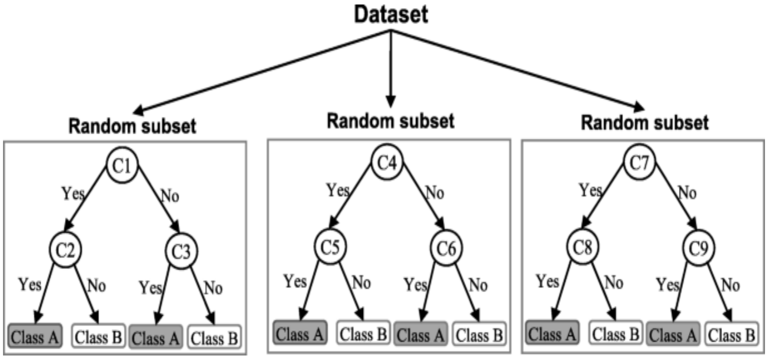

Fig No. 4 Random Forest

An illustration of a Random forest which consists of three different decision trees. Each of those three decision trees was trained using a random subset of the training data

Naïve Bayes

Naïve Bayes (NB) is a classification technique based on the Bayes' theorem This theorem can describe the probability of an event based on the prior knowledge of conditions related to that event.

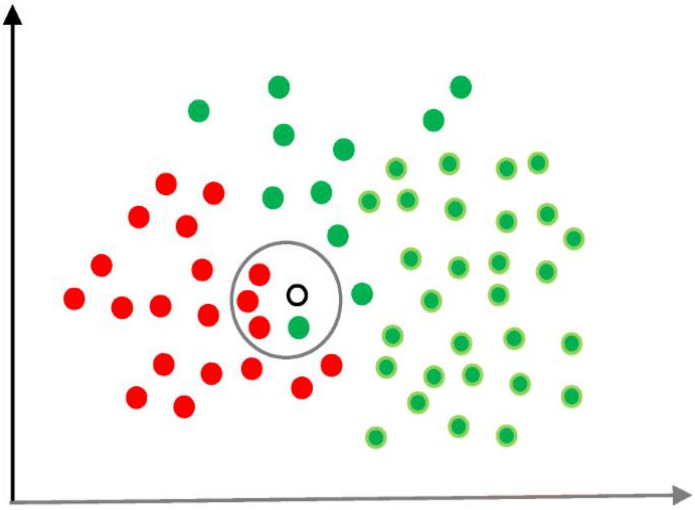

Fig No.5 Naïve Bayes

An illustration of the Naïve Bayes algorithm. The 'white' circle is the new sample instance which needs to be classified either to 'red' class or 'green' class

K-nearest neighbour

The K-nearest neighbour (KNN) algorithm is one of the simplest and earliest classification algorithms It can be thought a simpler version of an NB classifier. Unlike the NB technique, the KNN algorithm does not require to consider probability values.

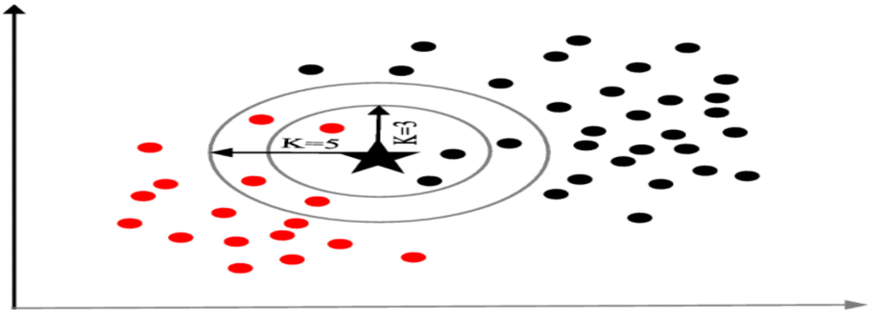

Fig No.6 KNN

A simplified illustration of the K-nearest neighbour algorithm. When K = 3, the sample object (‘star’) is classified as ‘black’ since it gets more ‘vote’ from the ‘black’ class. However, for K = 5 the same sample object is classified as ‘red’ since it now gets more ‘vote’ from the ‘red’ class

Artificial neural network

Artificial neural networks (ANNs) are a set of machine learning algorithms which are inspired by the functioning of the neural networks of human brain.. Likewise, ANN algorithms can be represented as an interconnected group of nodes. The output of one node goes as input to another node for subsequent processing according to the interconnection.

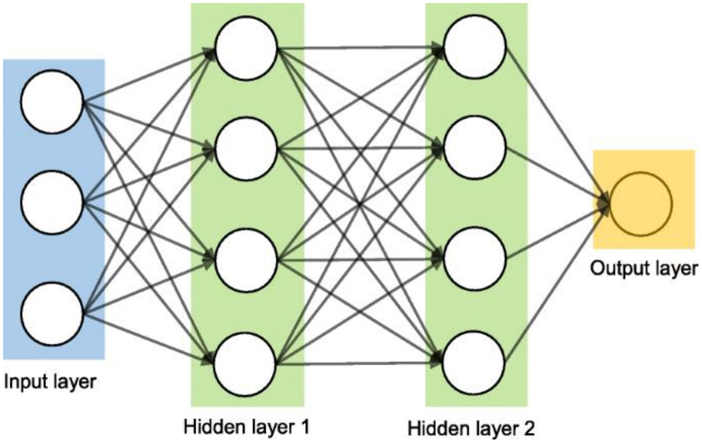

Fig No.7 Artificial Neural Network

An illustration of the artificial neural network structure with two hidden layers. The arrows connect the output of nodes from one layer to the input of nodes of another layer

IMPLEMENTATION

Dataset: The data is collected and this study of diseases and their corresponding symptoms is available on Kaggle.

Training Data: Training data is also known as training datasets, training sets, and training sets. It is an important aspect of the machine learning model which helps us to make accurate predictions and perform the tasks we want. Simply put, training data forms a machine learning model and tells you what the awaited result looks like. The model iteratively analyzes the dataset to understand its attributes precisely and make appropriate changes to enhance the performance.

Testing Data: The test dataset is a subset of the training dataset used to make an objective evaluation of the final model.

Balanced Data The observation of the dataset and its visualization leads us to the conclusion that the data is balanced and there's no imbalance in the data, which means that training and testing will give real exactness.

Correlation of Disease with Respect to their Corresponding Symptoms: Matrix data structures are used when there are multiple variables and the aim is to find the correlations between all these variables and store them using the applicable data structure. Thus this matrix is known as a correlation matrix.

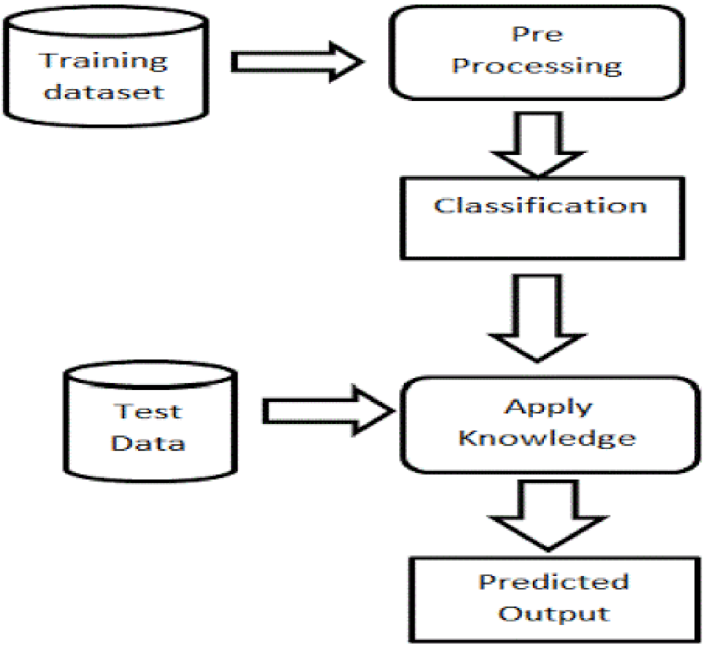

Fig No.8 Implementation

SOFTWARE DESCRIPTION

PYTHON:- Python is an open source programming language. It is a high-level language, which means a programmer can focus on what to do instead of how to do it. Writing programs in Python takes less time than in some other languages.

LIBRARY MODULES:-

1. NUMPY:- NumPy is a Python package. It stands for 'Numerical Python'. It is a library consisting of multidimensional array objects and a collection of routines for processing of array. Numeric, the ancestor of NumPy, was developed by Jim Hugunin.

2. PANDAS:- Pandas is an open-source Python library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

3.MATPLOITLIB:- It is a collection of command style functions that make matplotlib work like MATLAB.

4.SCIKIT-LEARN Scikit-learn is a machine learning library for Python. It is designed to work with Python Numpy and Scipy.

CONCLUSION:In this paper, algorithm used to predict the disease based on symptoms is discussed. Various symptoms are provided in the dropdown menu, out of which user selects any five of them and using algorithm the disease is predicted. The drugs that are commonly prescribed for a particular disease can also be suggested in this system. The main aim is to predict the disease at the early stage and lead to early diagnosis. This system can also be used by doctors to avoid confusion while predicting the disease. This system can provide assistance to doctors.

FUTURE SCOPE: In the future, the model can be used in various sectors and can enhance efficiency by considering moresymptoms to predict disease. The model can be used for providing an enhanced, more accurate framework that would lead to a better human disease prediction model.

REFERENCES:

[1] 1Palle Pramod Reddy, 2Dirisinala Madhu Babu, 3Hardeep Kumar and 4Dr.Shivi Sharma,” Disease Prediction using Machine Learning “International Journal of Creative Research and Thoughts Volume 9, Issue 5 May 2021 | ISSN: 2320-2882

[2] K. Gaurav, A. Kumar, P. Singh, A. Kumari, M. Kasar*, T. Suryawanshi,” Human Disease Prediction using Machine Learning Techniques and Real-life Parameters” IJE TRANSACTIONS C: Aspects Vol. 36 No. 06, (June 2023) 1092-1098

[3]KunalTakke, RameezBhaijee, Avanish Singh, Mr. Abhay PatilMedical Disease Prediction using Machine Learning Algorithms IJRASET **Publish Date:** 2022-05-02**ISSN:** 2321-9653

[4] PRIYANKA J. PANCHAL1 , SHAEEZAH A. MHASKAR2 , TEJAL S. ZIMAN3,” Disease Prediction Using Machine Learning” IRE Journals | Volume 3 Issue 10 | ISSN: 2456-8880 IRE 1702253 ICONIC RESEARCH AND ENGINEERING JOURNALS

[5] Akash C. Jamgade, Prof. S. D. Zade ,”Disease Prediction Using Machine Learning”, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-007

[6] Anjali Bhatt*1, ShrutiSingasane*2, NehaChaube*3,” DISEASE PREDICTION USING MACHINE LEARNING” International Research Journal of Modernization in Engineering Technology and Science (Peer-Reviewed, Open Access, Fully Refereed International Journal) Volume:04/Issue:01/January-2022 Impact Factor- 6.752

[7] Dr. K1 Kishore Raju S2 Hari Priya T2 Mohana Supraja R2 Lakshmi Sowjanya U2 Sai Pravallika,” MULTIPLE DISEASE PREDICTION USING MACHINE LEARNING JETIR April 2023, Volume 10, Issue 4 (ISSN-2349-5162)

[8] Alok Katiyar, Sajid Ali, Sameer Ray,” Multiple Disease Prediction Using ML” International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878 (Online), Volume-12 Issue-1, May 2023

[9] Md. Ehtisham Farooqui, Dr. Jameel Ahmad,” A Detailed Review on Disease Prediction Models that uses Machine Learning” International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN: 2347-5552, Volume- 8, Issue- 4, July-2020

[10] Priyanka Kadu, Amar Buchade ,”Non Communicable Disease Prediction System Using Machine Learning”, INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 8, ISSUE 09, SEPTEMBER 2019 ISSN 2277-8616 1307

[11] Lowlesh NandkishorYadav,”Predictive Acknowledgement using TRE system to reduce cost and Bandwidth”, IJRECE VOL7 ISSUE 1(January-march 2019) pg no 275-278