IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

MULTIPLE DISEASES PREDICTION IN PYTHON

Empowering Healthcare Through Predictive Analytics

Mr. TAHSEEN ALI, Miss. SHIVKANYA ANNAPURVE, Miss. RUDRANI PULKANTAWAR, Miss. PRIYANKA GORE

Assistant Professor, Students
Computer Engineering
GRAMIN TECHNICAL NAD MANAGEMENT CAMPUS VISHNUPURI, NANDED.431606

Abstract: Disease Prediction is a Machine Learning based system which primarily works according to the symptoms given by a user. The disease is predicted using algorithms and comparison of the datasets with the symptoms provided by the user. Many of the existing machine learning models for health care analysis are concentrating on one disease per analysis. Like one analysis if for diabetes analysis, one for cancer analysis, one for skin diseases like that. There is no common system where one analysis can perform more than one disease prediction. In the proposed system, it provides machine learning algorithms for effective prediction of various disease occurrences in disease-frequent societies.

The analysis accuracy is reduced when the quality of medical data in incomplete. There are multiple techniques in machine learning that can in a variety of industries, do predictive analytics on large amounts of data. Predictive analytics in healthcare is a difficult end, but it can eventually assist practitioners in making timely decisions regarding patients' health and treatment based on massive data. Diseases like Breast cancer, diabetes, and heart related diseases are causing many deaths globally but most of these deaths are due to the lack of timely check-ups of the diseases. The above problem occurs due to a lack of medical infrastructure and a low ratio of doctors to the population.

***Index Terms* – Machine Learning Algorithm, python, SVM, KNN, Random Forest.**

INTRODUCTION:

Many times, we see that patients lose their life because of not getting treatment on time. Healthcare industries lack time and they cannot determine which patient they should treat first. But on the other hand, healthcare industries generate huge amount of data regarding patients' health. High level of insights can be drawn from this data. So, by using this data and advanced machine learning techniques we have decided to come up with project 'Multiple Disease Prediction System'.

NEED OF THE STUDY:

I. Data bias: One of the biggest concerns with machine learning systems is data bias. If the training data used to develop the system is biased or incomplete, it can lead to inaccurate predictions and misdiagnosis. This is especially problematic when it comes to underrepresented populations, as their data may not be well-represented in the training set.

II. Overfitting: Overfitting occurs when a machine learning model is trained too closely to a particular dataset and becomes overly specialized in predicting it. This can result in poor generalization to new data and lower accuracy.

III. Lack of interpretability: Many machine learning algorithms are "black boxes," meaning that it is difficult to understand how they arrive at their predictions.

Technical Feasible

During this study, the analyst identifies the existing computer systems of the concerned department and determines whether these technical resources are sufficient for the proposed system or not. If they are not sufficient, the analyst suggests the configuration of the computer systems that are required. The analyst generally pursues two or three different configurations which satisfy the key technical requirements but which represent different costs. During technical feasibility study, financial resources and budget is also considered. The main objective of technical feasibility is to determine whether the project is technically feasible or not, provided it is economically feasible

Data collection:

I. **Data collection:** The first component of the system involves collecting a large dataset of medical records containing patient information and various medical features related to multiple diseases. This dataset will be used to train the machine learning models.

II. **Data Preprocessing:** The collected data will be pre-processed to handle missing values, outliers, and to perform feature scaling. This component of the system involves cleaning and preparing the data for model training.

III. **Model Training:** This component involves training different machine learning algorithms such as decision trees, random forests, and artificial neural networks on the pre-processed data. The trained models will be used for disease prediction.

IV. **Model Selection:** The performance of different machine learning algorithms will be compared using metrics such as accuracy, precision, and recall, and the best- performing model will be selected for disease prediction.

V. **Model Evaluation:** The selected model will be evaluated on a separate test dataset to measure its accuracy and reliability in predicting multiple diseases. This component of the system involves testing the model and measuring its performance.

VI. **User Interface Development:** The final component of the system involves developing a user-friendly interface that allows healthcare professionals to input patient information and receive predictions for multiple diseases. The interface will be designed to provide an easy-to-use tool for disease prediction.

Methodology:

In this section consists of the methodology adopted by our proposed work. As stated earlier our work aims to develop a web application to detect diseases like breast cancer, diabetes, and heart diseases using machine learning models
then we collected data from various open data sources like Kaggle, UCI Machine Learning Repository, etc. Quality and quantity o f data is very important as it affect to our model. We use 80% of our data for training and 20% of data for testing. We decide to go with Random Forest Algorithm because it gives us maximum accuracy for all of our machine learning model. And it is easy to im plement as well. Here our all models were ready for the prediction of diseases.

It is feasible to anticipate more than one illness at once when using the multiple disease prediction method. Hence, the user does n ot have to visit several sites in order to anticipate the ailments. We're going to utilize Flask and machine learning methods to imple ment numerous illness analyses. The parameters of the disease must be sent together with the disease name when a user accesses t his API. Flask will execute the relevant model and return the patient's state.

Literature survey:

This literature survey conducted for this research project explores the existing body of knowledge regarding the application of machine learning techniques, specifically Support Vector Machines (SVM), for the prediction of multiple diseases, including cardiovascular disease, diabetes, and Parkinson's disease.

Their aim of this analysis was to invent a system that can help the patient to detect the diabetes disease of the patient with accurate results. Here they used mainly 4 main algorithms Decision Tree, Naïve Bayes, and SVM algorithms and compared their accuracy which is 85%,77%, 77.3% respectively.

Implementation:

Algorithms:

1.Random Forest algorithm
2.KNN Algorithm
3.Multinomial Naive Bayes (MNB).

i. Random Forest algorithm:

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

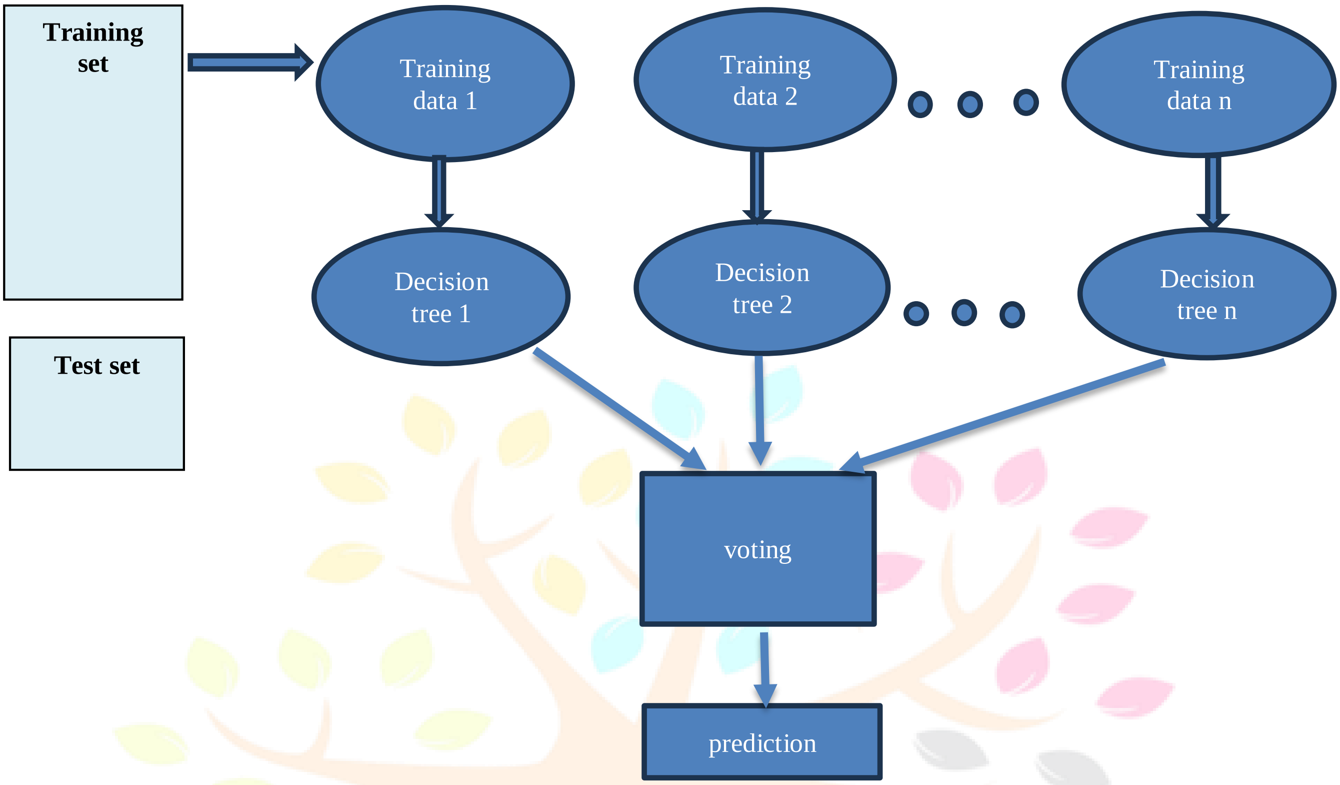


Fig: process diagram Random Forest Algorithm

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

ii. KNN algorithm:

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

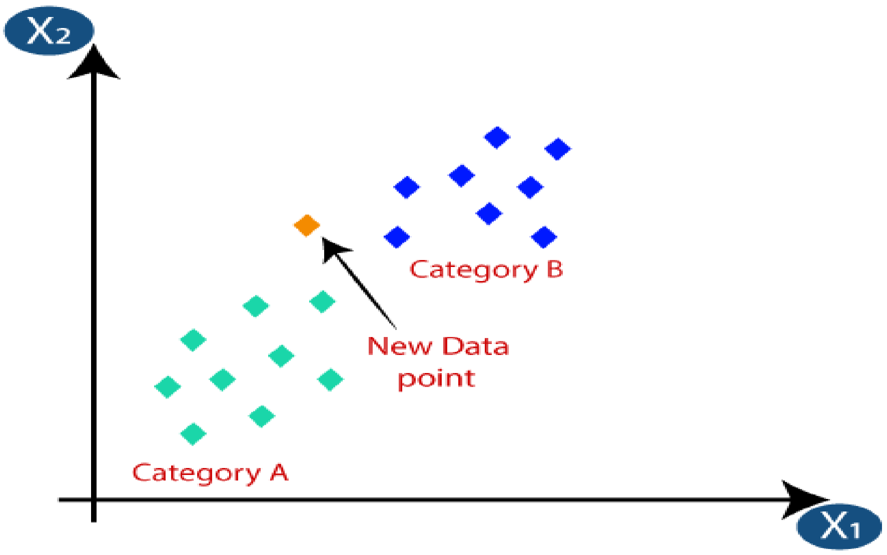
K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data. It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

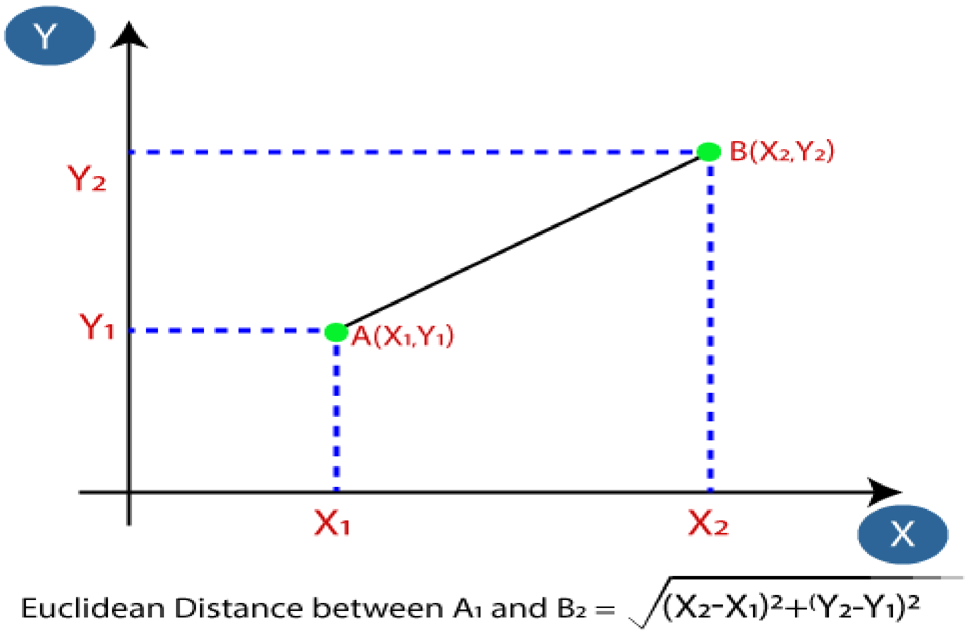
KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data

Suppose we have a new data point and we need to put it in the required category. Consider the below image:

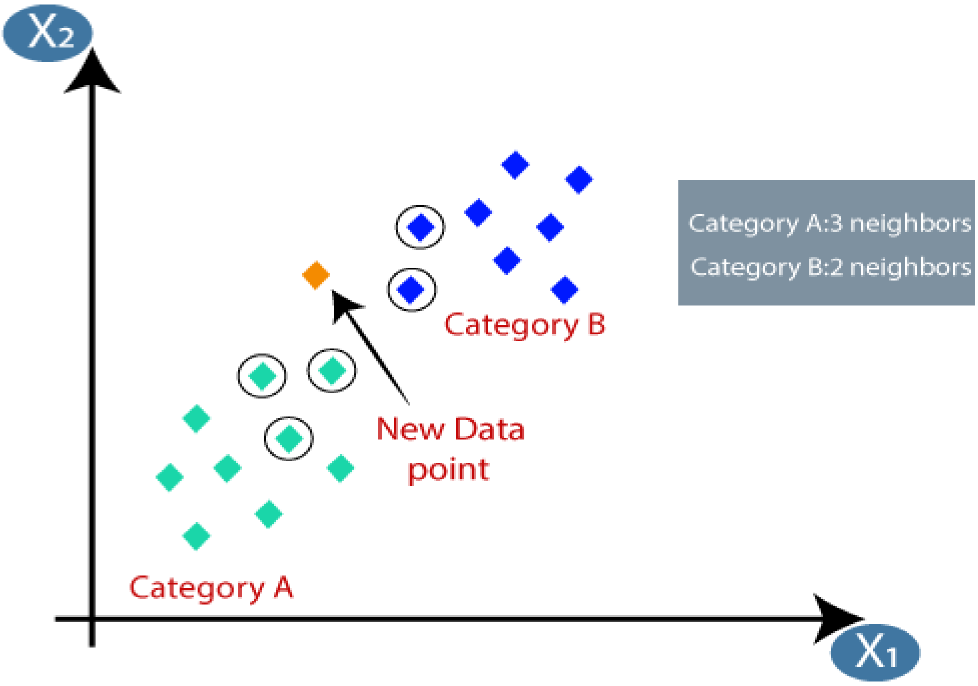


Firstly, we will choose the number of neighbors, so we will choose the k=5.

Next, we will calculate the **Euclidean distance** between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:



By calculating the Euclidean distance, we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image:



iii. **Multinomial Naive Bayes (MNB):**

Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.

Multinomial Naive Bayes works

Naive Bayes is a powerful algorithm that is used for text data analysis and with problems with multiple classes. To understand Naive Bayes theorem's working, it is important to understand the Bayes theorem concept first as it is based on the latter. Bayes theorem, formulated by Thomas Bayes, calculates the probability of an event occurring based on the prior knowledge of conditions related to an event. It is based on the following formula:

P(A|B) = P(A) * P(B|A)/P(B)

Future Scope:

i. This will be very useful to all medical industries to detect diseases at early stage of that patient.
ii. In the future we can add more diseases in it.
iii. We can try to improve the accuracy of prediction.
iv. We also try to make system user-friendly to user.

CONCLUSION:

In this paper the SVM model involved handling and filtering the data using libraries like pandas, performing model selection and training, comparison and fine-tuning the SVM model, evaluating its performance, and exporting the trained model for future use. We utilized the Support Vector Machines (SVM) model to develop a multi-disease prediction framework and achieved a high accuracy of 98.3%.

Accurate disease prediction using machine learning models has the potential to facilitate early interventions, personalized treatment plans, and targeted disease management strategies. It can assist healthcare providers in making informed decisions, enhance patient care, and improve resource allocation within healthcare systems.

References:

[1] Archana Singh, Rakesh Kumar, “Heart Disease Prediction Using Machine Learning Algorithms”, 2020 IEEE, International Conference on Electrical and Electronics Engineering (ICE3).

[2] Priyanka Sonar, Prof. K. Jaya Malini,” DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROACHES”, 2019 IEEE ,3rd International Conference on Computing Methodologies and Communication.

[3] Applying k-Nearest Neighbor in Diagnosing Heart Disease Patients Mai Shouman, Tim Turner, and Rob Stocker International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012.

[4] Arora S, Aggarwal P, Siva swamy J. Automated diagnosis of Parkinson's disease using ensemble machine learning. IEEE Trans Inf Technol Biomed. 2017;21(1):289-299**.**

[5] https://ieeexplore.ieee.org/document/10060903