Research Paper on Diseases Prediction Using Machine Learning

1) Rakshita Harde
2) Khushi Raut
3) Nikita Nakade
4) Sanika Bhoyar
5) Tikesh Kannake

Guided By : Prof. Chandrapal Chauhan

Bachelor Of Computer Science & Engineering Government College of Engineering, Chandrapur, Maharashtra, India.

ABSTRACT

There may be drawbacks to using traditional methods for both diagnosis and treatment, especially in the case of serious disorders. Consequently, it's critical to have accurate and analysis of health issues in a timely manner for efficient treatment and prevention. When developing a medical diagnosis system, machine learning (ML) algorithms can be a useful tool as they can yield more accurate disease predictions than traditional approaches. Our team has effectively created a disease prediction system by utilizing multiple machine learning techniques, such as Decision trees, Naive Bayes, and KNN. We have established a framework that facilitates the creation and utilization of prediction models through the implementation of a rule-based methodology. Because of the system's exceptional accuracy, doctors are better equipped to predict and diagnose illnesses early on and resolve health-related issues more skill fully. Index Terms: Prediction, Symptoms of Disease, Machine Learning, Classification

INTRODUCTION

The healthcare industry collects a lot of data, which can help make better decisions. To do this, we need to use advanced techniques to find hidden insights in the data. Diagnosing certain illnesses can be tricky because they have similar symptoms. Our project aims to create a system that predicts diseases using different algorithms like Decision trees, Naïve Bayes, and KNN. This system helps doctors and patients by giving accurate predictions based on symptoms. It can help doctors treat illnesses earlier, potentially saving lives. Our system acts like a virtual doctor, predicting diseases accurately without human involvement. This is important because healthcare can be expensive and sometimes it's hard to see a doctor in person. This system can quickly and cheaply help people understand their symptoms and what might be wrong with them. Doctors and nurses are working hard, especially now with more virtual healthcare. This system can help them by providing accurate predictions for patients. Machines are reliable and fast, so this system can be trusted to give good results. Our project used different machine learning techniques to analyze data and make predictions about diseases. We tested these techniques to see which ones were most accurate in predicting diseases based on symptoms. This helps us create a system that can be relied upon to give good predictions.

PROBLEM STATEMENT

While machine learning applications for disease prediction are developing at a rapid pace, there is still a critical need to improve the predictive models' interpretability, accuracy, and generalizability in a variety of healthcare settings. The inadequate integration of multi-modal data sources, the opaque nature of machine learning models in clinical decision-making, and the possible biases present in predictive algorithms are some of the current issues. Moreover, there is still work to be done in order to effectively translate study findings into workable, morally sound solutions for healthcare professionals. In order to create machine learning models for disease prediction that are more reliable, understandable, and broadly applicable, our research aims to methodically explore and solve these issues . This problem statement identifies important issues and provides a basis for a research project that attempts to improve and provide solutions in the field of machine learning-based disease prediction. These issues include data integration, model interpretability, biases, and practical implementation. The problem statement's specifics can be adjusted to fit the characteristics of the targeted diseases or healthcare settings, as well as the research aims.

OVERVIEW OF SYSTEM ARCHITECTURE

The goal of the system architecture for improving illness prediction using symptoms-based machine learning is to provide a reliable and effective framework for precisely forecasting illnesses based on the symptoms that people showed. With the use of machine learning techniques, this architecture is able to evaluate massive amounts of data, extract pertinent symptoms, and develop prediction models that can help with a variety of medical diagnosis and treatments. The first step in creating the architecture is gathering a large dataset of symptoms and the associated disease designations. The machine learning models are trained and assessed using this dataset as their basis. Preprocessing methods are used to handle missing values, guarantee data quality, and standardize the data in preparation for additional analysis.

The second phase in the procedure entails using feature extraction techniques to change the gathered symptoms into a format that is appropriate for machine learning algorithms. similar to embedding or one-hot encoding. The final objective is to implement these trained models in a real-world setting where users may enter symptoms and receive precise disease predictions. User-friendly interfaces or applications for seamless interaction and prediction might be included in this deployment phase. Moreover, the architecture of the system facilitates ongoing enhancement and optimization. Periodically updating and retraining the models with new data ensures that The system maintains great forecast accuracy and keeps up with the most recent data.

In conclusion, the system architecture offers a thorough framework that integrates data gathering, symptoms-based machine learning, and machine learning to improve disease prediction. preprocessing, feature extraction, model building, testing, implementation, and ongoing enhancement. Using algorithms for machine learning, Problem Formulation

The problem formulation involves optimizing a disease prediction model by selecting relevant features, training it with high-quality data, and evaluating its performance. The goal is to achieve accurate disease predictions that can contribute to improved healthcare decisions and patient outcomes. The task is to develop a machine learning-based disease prediction model using a dataset of symptoms and corresponding disease labels. The aim is to enhance disease prediction through symptoms-based machine learning, enabling early diagnosis and more effective treatment.

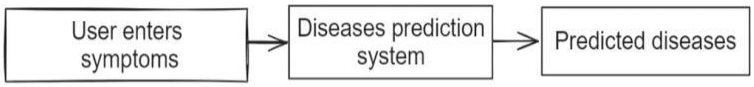


fig: System for predicting Diseases

METHODOLOGY:

The methodology section focuses on the methods and strategies used to use machine learning to accurately predict disease. Data collection, preprocessing, feature extraction, model building, training, evaluation, and deployment are the steps that are involved. Utilizing machine learning algorithms to forecast illnesses, and support early detection and efficient treatment is the aim. The approach seeks to improve healthcare outcomes by increasing the accuracy of disease prediction. An open-source dataset was used to create an Excel sheet with a detailed list of symptoms related to each illness. A person's symptoms can be entered into a machine learning model to enable the system to process the data and generate predictions for possible illnesses. After digesting the data, the model iteratively updates and improves its predictions by training and testing the algorithm using the most recent input.

The goal of the disease prediction system is to determine the probability of a given illness from the symptoms that have been recorded. In order to guarantee reliable findings, it includes a data refinement procedure that enhances the quality of the original raw dataset. In order to provide precise disease predictions, the system analyse and trains the model on user-input data using data mining techniques.

Based on reported symptoms, the system improves accuracy and provides insightful information about suspected diseases by utilizing machine learning and data mining approaches. A dataset is gathered and split into training and testing subsets as part of the illness prediction process. K-fold cross-validation is used for model selection. Well-known algorithms like the Naive Bayes classifier, Decision Trees and K-NN. During the training phase Naïve Bayes, is employed, and metrics are calculated on the test data for each algorithm. The same algorithms are used to test the models' prediction power using a different validation dataset. Combining the forecasts from various models yields the final prediction result.

A. Patient Symptoms as Inputs

It was believed during the algorithm's creation that the client fully comprehends the symptoms they are dealing with.

B. Preparing the data

A variety of methods are employed in data pre-processing to get raw data ready for analysis. These methods consist of feature selection, data normalization, transformation, and cleaning. Enhancing the quality of the data, eliminating errors, and raising the precision and efficacy of ensuing analysis are the principal objectives of data pre-processing.

C. Constructing Models

Three machine learning methods are used by the illness prediction system: Decision Tree, Naive Bayes, and KNN . These algorithms are crucial for assessing the correctness and

performance of the system since they perform predictive analysis on the input dataset. Throughout the trials, the subsequent models were constructed:

Decision tree:

For tasks involving regression or classification, the decision tree method is a supervised learning technique. It uses a tree diagram and a top-down methodology to generate predictions. Starting with a root node, the algorithm divides it recursively according to the most significant input attribute. Recursively splitting the input data continues until all inputs are assigned and the final nodes have the weights required for either classification or regression.

Naive Bayes:

Based on the Bayes theorem, the Na¨ve Bayes algorithm is a supervised learning technique for classification issues. It is one of the simplest and most efficient classification algorithms known today, and is mostly used for text categorization. With the use of the Naive Bayes Classifier, effective machine learning models with precise prediction capabilities can be developed. Predictions are made by this probabilistic classifier based on the possibility that an event will occur. The Naïve Bayes theorem, which is described as follows, was applied by the team throughout this investigation.

$$P(X|Y) = \frac{P(X|Y)\cdot P(Y)}{P(X)}$$

P(Y—X)" denotes the probability that, in the event that hypothesis 'X' were correct, data 'Y' would exist. P(X)" stands for the hypothesis "X's" prior probability. The probability is shown by P(Y)“of the information separate from the theory. P(X), often known as the prior probability, represents the likelihood that hypothesis "a" is true given all the information at hand P(Y) represents the likelihood that the data will not change regardless of the hypothesis.

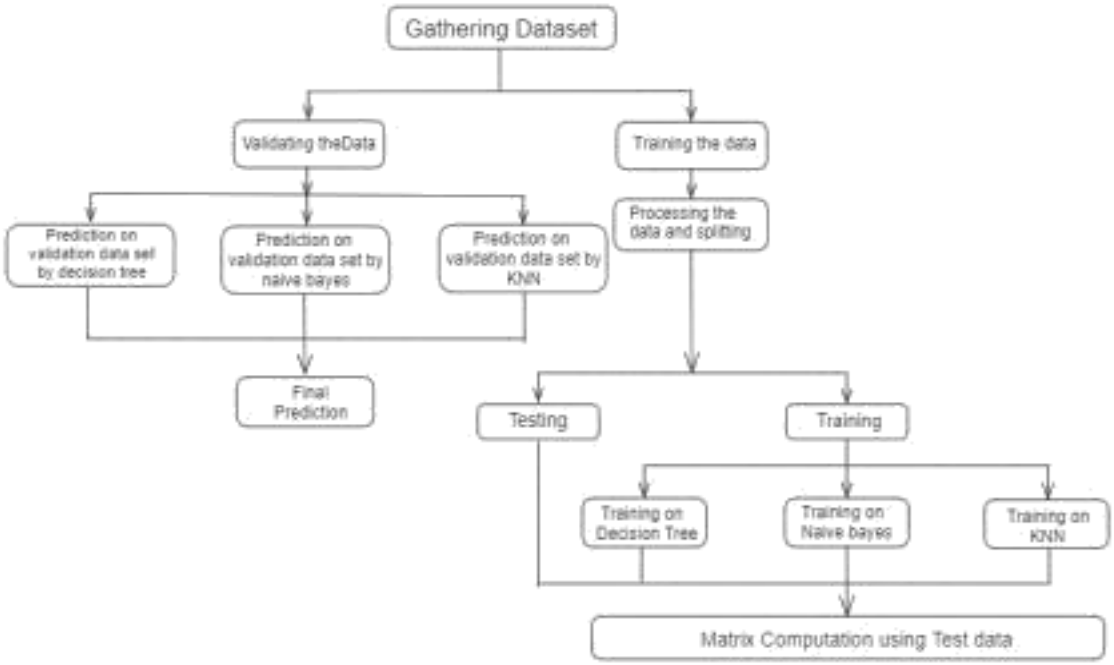
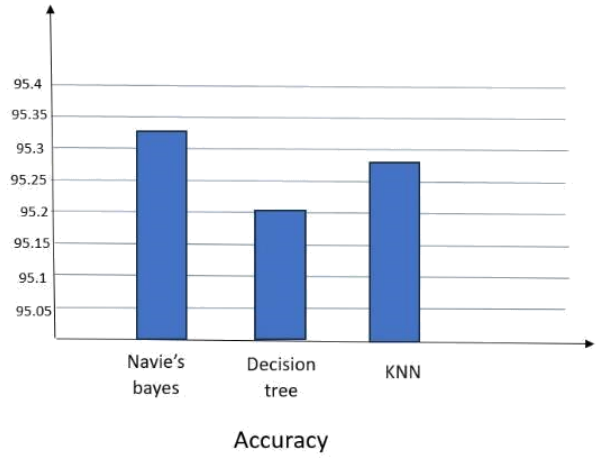


Fig : Workflow of System



KNN (K Nearest Neighbour):

The dataset has attributes pertaining to the illness (such as test results, patient characteristics) along with labels designating whether or not those attributes apply to specific individuals .Determine the distance between any given instance in the testing set and every instance in the training set. Depending on the type of data, common distance metrics include Euclidean distance, and other measures. Determine which of the training set's 'k' examples are closest to the testing instance in terms of distance. Ascertain which of these 'k' closest neighbour is the majority class. To do this, a simple majority vote might be used.
Based on the majority class from the neighbour, assign the testing instance to the expected class (presence or absence of the disease). Make predictions for each occurrence in the testing set by iteratively going through each one and repeating the procedure. By comparing the model's predictions with the genuine labels in the testing set, you can evaluate its accuracy and performance using measures like accuracy, precision, recall, and F1 score. Adjust variables such as the 'k' value or distance measure to maximize the predictive power of the model. The premise behind KNN's prediction is that cases with comparable feature values are probably members of the same class.

RESULT AND DISCUSSION:

The research paper's results and Discussion part provides the findings and analysis of the machine learning-based disease prediction models. The purpose of this section is to explain the findings, go over their ramifications, and offer some insight into how accurate and successful the suggested strategy was . We provide an overview of the main conclusions and emphasize their importance in relation to improving disease prediction using symptoms-based machine learning. We give a synopsis of the goals of the study and reiterate the key techniques that were applied in creating the prediction models.
Based on an input dataset, many machine learning models were utilized in this work to predict the occurrence of diseases, including Decision Tree, Naive Bayes, and 95% accuracy was attained overall. This prediction method is helpful for disease diagnosis in scenarios when access to physicians and medical resources may be restricted.

CONCLUSION :

The study offers a fresh method for estimating a patient's prognosis for their illness by using symptom analysis.
There are various advantages to this approach, such as its effectiveness.
distribution and administration of health care resources for diseases that are anticipated. It speeds up the healing process and helps to reduce treatment costs by correctly predicting disorders.
It also encourages the preservation of wellbeing by offering free health examinations on a regular basis. The model's user-friendliness allows users to choose three diseases of interest, which allows for excellent forecasts. Users get insightful information about their health after obtaining the prediction, enabling them to seek the necessary medical care when necessary.

REFERANCE:

[1]Rinkal Keniya · Aman Khakharia · Vruddhi Shah · Vrushabh Gada · Ruchi Manjalkar · Tirth Thaker · Mahesh
Warang · Ninad Mehendale , "Disease prediction from various symptoms using machine learning"

[2] D. Dahiwade, E. Meshram , and G. Patle, "Designing disease prediction model using machine learning approach",in 2019 Proceedings of the 3rd International Conference on Computing Methodologies

[3] Y. Amirgaliyev, A. Serek and S. Shamiluulu, "Analysis of chronic kidney disease dataset by applying machine learning methods", in 2018 IEEE 12th International Conference on Application of Information and Communication
Technologies (AICT).

[4] M. Marimuthu, M. Abinaya, K. S., K. Madhankumar, and V. Pavithra, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach," International Journal of Computer Applications, vol. 181, no. 18, pp.
20–25, 2018.

[6] Stepheny Lucas, Mitali Desai, "Smart Care: A Symptoms Based Disease Prediction Model Using Machine Learning Approach",in 2022, International Journal for Research in Applied Science & Engineering Technology

[7] Nidhi Kosarkar; Pallavi Basuri, "Disease Prediction using Machine Learning",in 2022, IEEE, 10th International
Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-22).

[8] Yohanes Gutema Robi and Tilahun Melak Sitote ,"Neonatal Disease Prediction Using Machine Learning Techniques", IN 2023, Hindawi Journal of Healthcare Engineering.

[10] S. Uddin, A. Khan, M. A. Moni and M. E. Hossain, "Comparing different supervised machine learning algorithms for disease prediction", in 2019, BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1–16, 2019.