IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11418

Disease Prediction Application

Deepshikha[1], Charvi Singhal[2], Charu Tamar[3], Kumari Saloni[4], Garima Singh[5]

Department of Computer Science & Engineering, Inderprastha Engineering College, Uttar Pradesh, India[1-4]

Department of Data Science, Inderprastha Engineering College, Uttar Pradesh, India[5]

Abstract: The current medical system focuses on specific, well-known diseases and is unable to accurately diagnose and predict disease based on early symptoms. These models use a variety of patient characteristics to balance the probability of an outcome over some time and to harness the power to improve decision-making and personal care. Discovering hidden patterns and collaborations from a medical website and the growing testing of a predictable disease model is essential. This paper aims to design a model which can easily diagnose various diseases relying on their symptoms. The model evaluates the user's symptoms as input and returns the disease probability as an output[1].The disease probability is calculated by making use of the naive bayes classifier. Therefore this research paper will attempt to apply machine learning activities to health facilities in a particular program. The proposed web-based forecasting app uses the Naive Bayes Algorithm and Decision Tree, a machine learning method as a diagnostic separator based on real-life clinical information.

Keywords: Machine learning, Naive Bayes, Decision Tree, Disease Prediction.

I. INTRODUCTION

It is estimated that every 2 months, more than 70% of the population in India is prone to common physical ailments such as colds, coughs, and chills. While most people do not realize that the symptoms of these common illnesses may be symptoms of something very serious, 25% of people die from ignorance of the first symptoms. Therefore, the diagnosis of the disease in the early stages is important to prevent any unnecessary injuries. The current medical system focuses on specific, well-known diseases and is unable to accurately diagnose and predict disease based on early symptoms[2]. We propose an application where we will be diagnosing diseases and recommending leading physicians based on patient reviews[3]. Patient satisfaction is one of the best valid indicators of a physician when caring for quality and each patient's review is critical to providing the best possible outcome.

Many health care providers will be collecting patient inputs and analysing patient review data and will collect data from doctors' offices, clinics, hospitals and will record patient information to evaluate physician performance. The data set is analysed using the algorithm of Naive Bayes and the Decision Tree when approaching a problem with a specific question to analyse and find a solution between two or more independent variations and dependent variables. In this paper, we worked on multiple supervised machine learning models for each disease recognition task[4]. Because evaluating the effectiveness of a single method across multiple research settings introduces bias, resulting in imprecise results, this approach provides better comprehensiveness and precision.

The main contributions based on the research conducted are as follows:

Disease prediction: This paper provides a disease-predicting algorithm based on given indicators using modern class distinction algorithms.
Analyse ensemble results: Combine the results of the individual algorithms with voting mechanisms.

II. APPROACH

The system starts with a disease dataset as an input. Diabetes, day to day cold and cough, body aches, heart disease, and cancer databases were chosen for study because they contain a wealth of information about the individuals' health care and general statistics. For efficient data analysis, evaluate the missing values and correlation[5].

This helps to partition the training data into 80 percent original data and 20 percent testing data. To measure the system's performance against the input disease dataset, data- mining methods such as random forest and Naive Bayes are used. When the categorization results are compared to previous findings, it is clear that there has been a significant improvement.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

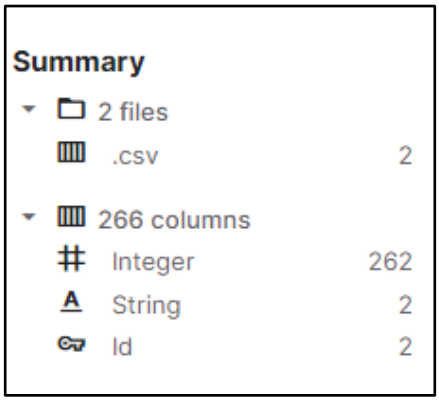
Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11418

A. Data Gathering:

Processing of the data is the foremost important step in any algorithm of machine learning techniques. In this step a database from the website Kaggle is taken [6]. The database contains two Comma-Separated Values files (CSV), in which one is for testing and the other is for training the dataset purposes. In this we have about 133 columns of the data from which 132 columns demonstrate symbols and the remaining column is for prediction in both test and training CSV files.

There are in total 4921 rows in the training csv file to train the dataset which consists of various disease names in columns and their symptoms in the rows.



Summary of Dataset

B. Data Cleaning:

To ensure quality of data and analysis outcomes, data cleaning will filter, detect, and treat unclean data. There may be noise in the form of impossible and extreme numbers, outliers, and missing values in this scenario. Inconsistent data and duplicated properties and data are examples of mistakes.

Therefore the data must be clean prior to placing it in the training sample. The data source used consists of columns which have integers, and the destination column is the only with string type and will be encrypted to numerical format with the help of a label encoding [7]. The initial stage will be to identify null values in the dataset and, if possible, replace them.

C. Building of Model:

Thereafter, collecting and cleansing the data, it is prepared and can be utilized for training purposes of machine learning samples. The split into two different sets - the training set and the test set - is represented by the two datasets. They have the same properties, but they don't have the same attribute values.

The categorization models are trained and constructed using the training set. Before evaluating the model's performance based on the performance metrics of accuracy, precision, recall, and F1-score of those classifications. The Confusion Matrix is used to evaluate the nature of the samples being used.

D. Prediction:

In succession of training the samples, the output is obtained which is the predicted disease on the basis of given symptoms. This makes our whole assumptions stronger and more accurate.

The block diagram of the basic stages followed by each machine learning model is shown in figure 3. To turn the raw data into a useful format, data cleaning is first undertaken.

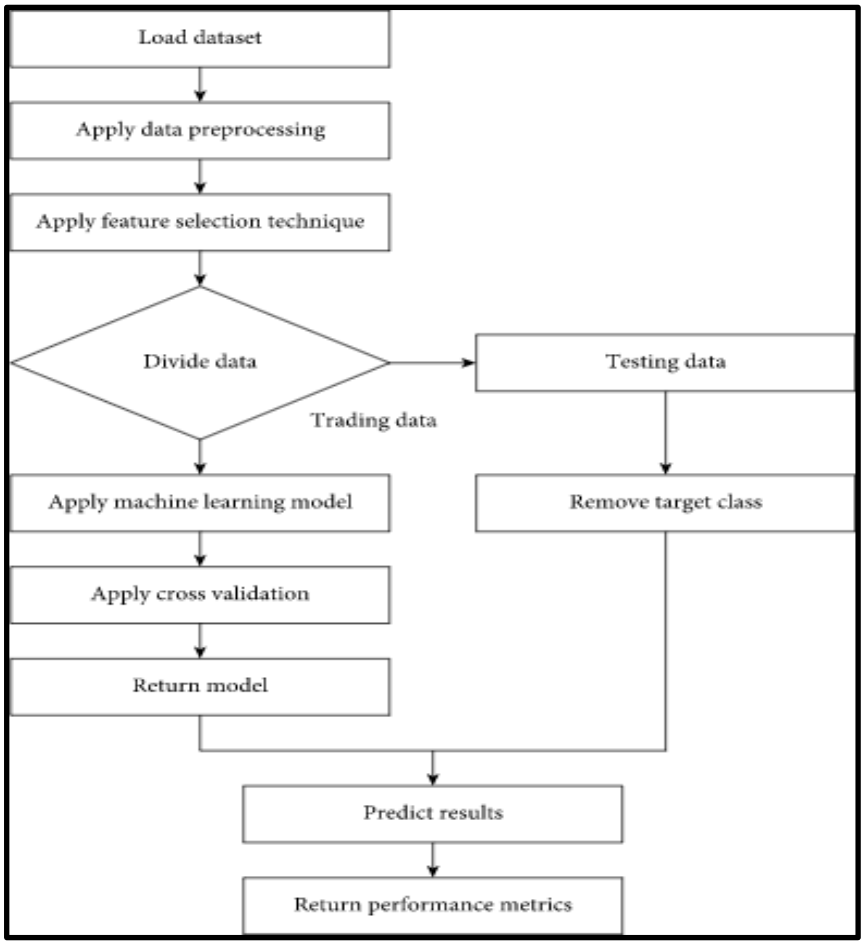
Information will be analyzed after data cleansing to assess the relevance of features. The features are identified via data analysis, and the data is translated into a format that machine learning methods may be applied to.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11418



Block Diag. of basic steps

III. CLASSIFICATION OF ALGORITHMS

A. Naive Bayes:

A machine learning algorithm for division problems, assign class labels to problem instances, and is based on the Bayes perspective. The main use of this is to create text classification that includes high-quality training data sets. We used Bayes theorem which can be defined as:

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

Where P (A | B) there is a possibility of hypothesis h if data is given d i.e. posterior probability.
P (B | A) data potential is hypothesis h which is true. P (A) probability that hypothesis h is true (nevertheless of data) i.e. prior probability. P (B) data potential (regardless of hypothesis) [8].

B. Gaussian Naive Bayes:

Algorithm is a sort of Naive Bayes algorithm that follows Gaussian normal distribution, which is continuous and unique. In Gaussian Naive Bayes we need a dataset where all the data values are numeric. It is as easy as computing the standard deviation and mean values of each input variable (x) for each class value. For instance, let training data hold a continuous attribute x. Firstly classify the data into categories, then compute the definition and variation of x for each class. Let μi be the meaning of values and σi be the difference of values associated with the ith category. Suppose we have a certain viewing xi. Then, the distribution of xi opportunities given to the class can be calculated by the following: [9]

$$p(x_i|y_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}}$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11418

C. Decision Tree:

The decision tree algorithm belongs to the supervised learning algorithms. Used for both classification and regression. The tree of decision prediction uses the tree diagram method at the top. The root node is then split into input features and then split again. These processes continue until all inputs are put back in place, the last extreme node containing weights on the bases of these weights separates the input. In the coarse tree, the maximum amount of split in the middle of each node is 4. Although in the Middle Tree, the maximum number of divisions in each area is 20.

D. K- Star:

It is a model-based separator, which is a test model-based class of those similar training conditions, as determined by a specific function of similarity. It is different from other algorithms because it operates on entropy-based distance functions. This algorithm uses an entropic scale formed on the likelihood of converting an event into random selection among all possible variations [10]. Using entropy as a distance measurement has many resources.

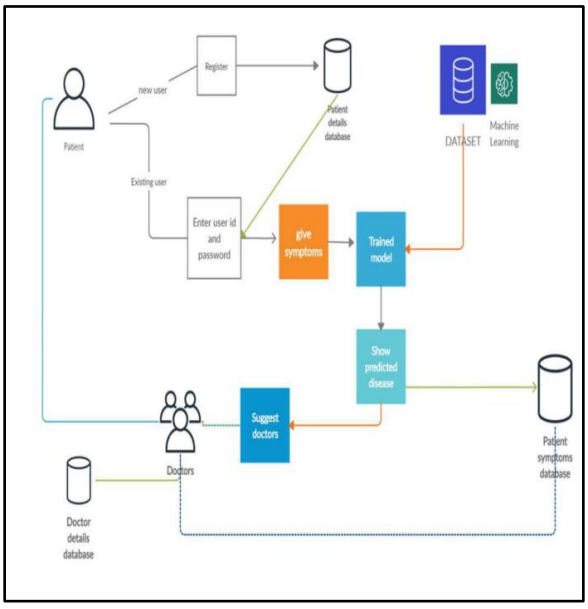
IV. PROPOSED METHOD

Two main components make up the proposed solution:

A) A web-based application and

B) A machine learning-based disease prediction system.

The web application acts as GUI for the user to interact with while the disease prediction prediction is to expect the illnesses primarily based totally at the given symptoms. The technology allows the patient to provide symptoms, and the algorithm will predict a condition with a high level of accuracy based on those symptoms. Following the diagnosis of the anticipated ailment, the system will recommend specialists that specialize in that disease, allowing the patient to visit with the doctor online. In the prediction system, the ensemble classifier technique using naive Bayes, Decision Tree, is utilized. The back-end of this project has been developed using the Django framework. HTML, CSS, JS, Bootstrap, and Jquery are used to connect the front-end and back-end, and PostgreSQL is utilized as the database.



Flow Chart of the proposed system

V. COMPARATIVE STUDY

To verify if the predicted and actual values are matching each other or not a cross-checking was implemented if they are the same both will print the same real disease else they show the wrong prediction.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11418

```
Pred: Acne Actual:Acne
Pred: Acne Actual:Acne
Pred: Hyperthyroidism Actual:Hyperthyroidism
Pred: AIDS Actual:AIDS
Pred: Chronic cholestasis Actual:Chronic cholestasis
Pred: Hypertension  Actual:Hypertension
Pred: Hypoglycemia Actual:Hypoglycemia
Pred: Arthritis Actual:Arthritis
Pred: Hepatitis B Actual:Hepatitis B
Pred: Migraine Actual:Migraine
Pred: Urinary tract infection Actual:Urinary tract infection
Pred: Diabetes  Actual:Diabetes
Pred: Hepatitis D Actual:Hepatitis D
```

Actual & Predicted Diseases Cross Checking

While comparing Naive Bayes with different algorithms of Machine Learning like Decision Tree, K* [11]. It was observed that the winning probability of NB was the highest when taken with a finite number of diseases.

Medical Problems	NB	LR	K*	DT	NN	ZeroR
Breast Cancer Wise	97.3	92.98	95.72	94.57	95.57	65.52
Breast Cancer	72.7	67.77	73.73	74.28	66.95	70.3
Dermatology	97.43	96.89	94.51	94.1	96.45	30.6
Echocardiogram	95.77	94.59	89.38	96.41	93.64	67.86
Liver Disorders	54.89	68.72	66.82	65.84	68.73	57.98
Pima Diabetes	75.75	77.47	70.19	74.49	74.75	65.11
Haeberman	75.36	74.41	73.73	72.16	70.32	73.53
Heart-c	83.34	83.7	75.18	77.13	80.99	54.45
Heart-statlog	84.85	84.04	73.89	75.59	81.78	55.56
Heart-b	83.95	84.23	77.83	80.22	80.07	63.95
Hepatitis	83.81	83.89	80.17	79.22	80.78	79.38
Lung Cancer	53.25	47.25	41.67	40.83	44.08	40
Lymphography	84.97	78.45	83.18	78.21	81.81	54.76
Postoperative Patient	68.11	61.11	61.67	69.78	58.54	71.11
Primary Tumor	49.71	41.62	38.02	41.39	40.38	24.78
Wins	**8/15**	**5/15**	**0/15**	**2/15**	**1/15**	**1/15**

Analysis of various Algorithms on different diseases

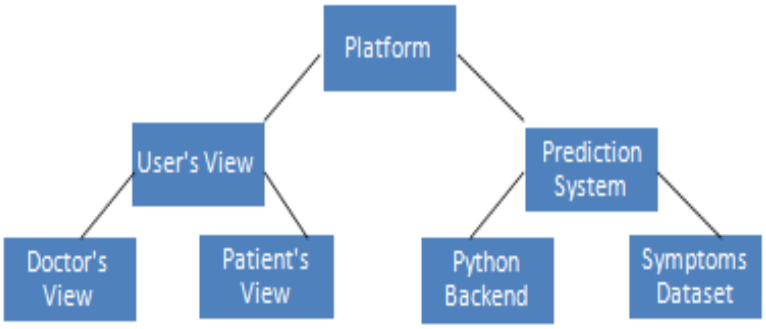
VI. RESULTS

Disease prediction applications are very helpful in day to day life. The recognition of disease is simple for doctors and disease can be managed on time. The phases for the various incoming diseases can be computed precisely and based on that, diseases can be treated. The system is set in such a way that the convenience of the user is at maximum priority and a user-friendly project that asks the user for the symptoms they are facing as input and provides the disease based on the input. Many health care providers will be collecting patient inputs and analyzing patient review data and will collect data from doctors' offices, clinics, hospitals and will record patient information to evaluate physician performance [12]. Finally, the accuracy of risk prediction in disease risk modeling is determined by the diversity of hospital data. This model not only helps in “predicting a disease” but also in curating medical data, enhancing research

DOI: 10.17148/IJARCCE.2022.11418

activities and controlling fraudulent activities. Our prediction approach can be useful and employed in the diagnosis of a disease in the current COVID-19 condition, where adequate facilities and resources are absent.

Layout



High level schematic representation of proposed framework

VII. CONCLUSION

The main motive of this research paper is to exploit a various Machine Learning algorithm to predict disease based on patient symptoms. In this study, we have used four Machine Learning algorithms for prediction, which showed outstanding rectification and higher accuracy when compared to previous work, as well as making this system more trustworthy than the present one for this purpose, and so giving higher user satisfaction. It also preserves the user's data, as well as the name of the condition that the patient is suffering from, in a database that can be used as a historical record and will aid in future treatment, making health management easier.

We've also created a user-friendly graphical user interface (GUI) to make it easier for users to interact with the system. This study shows how many indicators and models can be utilized to predict disease using a Machine Learning algorithm. Finally, we can state that our system has no user threshold because it is accessible to anybody and everybody.

REFERENCES

[1]Kedar Pingale, Sushant Surwase, Vaibhav Kulkarni, Saurabh Sarage, Prof. Abhijeet Karve. Disease Prediction using Machine Learning in International Research Journal of Engineering and Technology (IRJET), Volume: 06 Issue: 12 | Dec 2019.

[2]D. Dahiwade, G. Patle and E. Meshram, "Designing Disease Prediction Model Using Machine Learning Approach," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 1211-1215, doi: 10.1109/ICCMC.2019.8819782.

[3]A. N. Repaka, S. D. Ravikanti and R. G. Franklin, "Design And Implementing Heart Disease Prediction Using Naives Bayesian," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 292-297, doi: 10.1109/ICOEI.2019.8862604.

[4]Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN COMPUT. SCI. 1, 345 (2020).

[5]https://docs.microsoft.com/en-us/dotnet/machine-learning/how-to-guides/train-machine-learning-model-ml-net.

[6]https://www.kaggle.com/neelima98/disease-prediction-using-machine-learning.

[7]https://www.geeksforgeeks.org/disease-prediction-using-machine-learning/

[8]https://www.analyticsvidhya.com/blog/2021/11/implementation-of-gaussian-naive-bayes-in-python-sklearn/

[9]https://iq.opengenus.org/gaussian-naive-bayes/

[10]Painuli, Sanidhya & Elangovan, M. & Sugumaran, V.. (2014). Tool condition monitoring using K-star algorithm. Expert Systems with Applications. 41. 2638–2643. 10.1016/j.eswa.2013.11.005.

[11] Sayali D. Jadhav, H. P. Channe, Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611.

[12] Dash, S., Shakyawar, S.K., Sharma, M. et al. Big data in healthcare: management, analysis and future prospects. J Big Data 6, 54 (2019).

