Prediction of Diseases Using Different Machine Learning Approaches

Dr. Anish Gupta
Professor
Department of Computer Science & Engineering
Apex Institute of Technology
Chandigarh University
Mohali, Punjab, India
gupta.anish1979@gmail.com

Manish Kumar Gupta
Assistant Professor
Department of Computer Science & Engineering
Buddha Institute of Technology, GIDA
Gorakhpur, India
manish.testing09@gmail.com

2022 3rd International Conference on Intelligent Engineering and Management (ICIEM) | 978-1-6654-6756-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIEM54221.2022.9853132

***Abstract*— Every year, as the number of patients and diseases increases, the medical system becomes overburdened and, in many nations, expensive. The majority of the condition necessitates a visit with a doctor in order to be treated. With enough data, disease prediction using an algorithm can be simple and inexpensive. Predicting disease based on symptoms is an important element of treatment. In our project, we attempted to accurately forecast an illness based on the patient's symptoms. For this, we utilized four different algorithms and achieved 92-95 percent accuracy. A system like this has a lot of potential in future medical therapy. We've also created an interactive interface to help you interact with the data.**

Keywords*— Machine Learning, Disease Prediction, Disease Dataset, Decision Tree, Random Forest, KNN, Nave Bayes are some of the terms used in this paper.

I. INTRODUCTION

Health information needs are also influencing information seeking behaviour, as seen all across the world. Many people experience difficulties when looking for health information online on ailments, diagnostics, and treatments. It will save a lot of time if a recommendation system for doctors and medicine can be developed using review mining. Because the users are laypeople, they have difficulty grasping the diverse medical jargon in this type of system. Because there is so much medical information available on many channels, the user gets perplexed. The goal of the recommender system is to adapt to the unique requirements of the health sector in terms of users.

For the avoidance and action of illness, exact and appropriate examination of any health-related problem is serious. In the case of a serious disease, the typical method of identification may not be sufficient. The development of a medical diagnosis system based on machine learning (ML) algorithms for disease prediction can aid in a more accurate diagnosis than the current method. Using numerous machine learning techniques, we created a disease prediction system. There were about 230 diseases in the dataset that needed to be processed. The diagnosis system outputs the disease that an individual may be suffering from based on the individual's symptoms, age, and gender. In comparison to the other algorithms, the weighted KNN method produced the best results. The prediction accuracy of the weighted KNN method was 93.5 percent.

Our diagnostic model acts as a doctor for early detection of illness, allowing timely treatment and saving lives.

We have succeeded in creating such a system using four different algorithms. On average, we achieved an accuracy of about 94%. Such a system can reduce the rush to the OPD of the hospital and reduce the burden on the medical staff. I wanted to create a system that could predict illness based on given symptoms. We also tried to show and visualize the results of our research and this project. The idea behind the prediction system is to adapt to the specific needs of the medical sector from the user's perspective.

II. MACHINE LEARNING ALGORITHMS

A. Decision Tree: Decision Tree is a very efficient and multitalented classification technique. It is used for classification and pattern recognition for image. It is used for classification in very difficult problems due to its high compliance. It is also capable of appealing problems of higher dimensionality [2]. There are mainly three parts in decision tree- root, nodes and leaf. Decision Tree is used for both classification and regression tasks due to its property of nonparametric supervised learning method.

B. Random Forest: Random forests are a collection of various decision trees trained using bagging techniques. This is a supervised learning algorithm used for both regression and classification. Select a random data sample from the dataset. Create a decision tree for each selected sample dataset. In this step, each predicted result is assembled and collated. Finally, the prediction with the most votes is selected and displayed as the result of the classification. It creates a set of decision trees from a arbitrarily selected subset of training sets.

C. K Nearest Neighbours: KNN is an algorithm based on supervised learning technique, is one of the simplest machine learning algorithm. This algorithm also used for

978-1-6654-6756-8/22/$31.00 ©2022 IEEE

regression as well as classification, but mainly used for classification.

D. *Naïve Bayes Algorithm*: This is a simple "stochastic classifiers" based on the assumption of independence in Bayes' theorem. It is a supervised learning algorithm based on Bayes Theorem is mainly used for text classification. Naïve Bayes Algorithm predict on the basis of the probability of an entity.

Formula for the Naïve Bayes Algorithm

$$P(X|Y) = P(X|Y)P(X)/P(Y)$$

Where P(X|Y) means- probability of hypothesis X on the observation event Y.

Where P(Y|X) means- Probability of the confirmation given that the probability of a hypothesis is true.

III. PROPOSED METHOD

The steps proposed to build a predictive model of the disease are as follows:

- *Examine the dataset*: The datasets are inspected in the Python setting with the data vocabulary of the relevant attributes.

- *Data Alteration:* Estimating the missing value for a little variables, which is essential because almost all of the interpretations could not be performed with the absent data.

- *Attribute selection*: Improves model performance by removing all redundant features that are important for predictive modeling and are performed to account for multi co linearity and are highly correlated with each other. A backward selection method was used to eliminate features that are not important in diagnosing the disease. [6].

- *Model decent and Testing*: After feature selection, four classification algorithms are used on the selected features, including K- nearest neighbors, random forest, decision trees and naive bays, and their prediction accuracy comparisons are based on the train split method. Was done using. The trial size for assessment was set to 0.1. That is, almost 90% of the dataset trained the classifier and the remaining 10% were tested. Following figure explain the steps in the planned procedure.

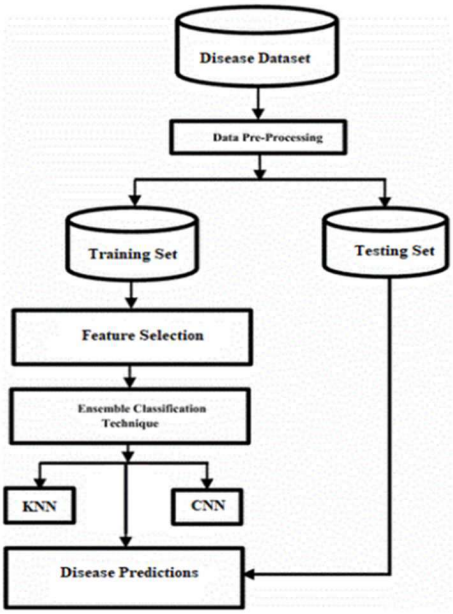

Figure 1. Projected Method

IV. EXPERIMENT

A. *Key (Symptoms)*: When developing the sculpt, presume that the user have a plain idea of the symptoms that he was facing. The developed predictions consider 95 symptoms, in which the user can provide the symptoms of his meting out as input [31].

B. *Data Preprocessing*: Conversion of raw data (encoded data) into a format that can be interpreted by algorithms, is known as data preprocessing and for this we use data mining technique [31]. The pretreatment techniques that are used in the work are:

- *Data clean-up*: The data is sanitized from side to side process such as entering missing values to resolve data inconsistencies.

- *Data decline*: Analyzing is difficult when dealing with large databases. Therefore, eliminate self-governing variables (symptoms) that may have little or no effect on the target variable (disease).

C. *Models Selected*: The system is trained to predict illness using four algorithms:

- Disease Tree Classifier
- Random forest Classifier
- K Nearest Neighbors
- Naïve Bayes Classifier

At the end of the work, a comparative study is presented that analyzes the performance of each algorithm in the database under consideration.

D. *Productivity (Diseases)*: When the system is trained in a training set using the above algorithm, a rule set is created, and when the user is given symptoms as input to the model, these symptoms are processed and classified according to the developed rule set. Is done and most are expected. They are probably ill.

V. LITERATURE REVIEW

Much research has been done on predicting illness based on human symptoms using machine learning algorithms. Montow et al. [4] designed a statistical model to predict whether a patient has the flu. They included 3744 unvaccinated adults and young flu patients who had fever and at least two other flu symptoms. Of the 330, 2470 confirmed influenza in the laboratory. Based on this data, their model provided 79% accuracy. Sreevallietal. [4] Used a random forest machine learning algorithm to predict illness from symptoms. With this system, there was little time and cost to predict the disease. The algorithm gave an accuracy of 84.2%.

Various tools have been developed by Langbehn et al. Developed. [20] Detect Alzheimer's disease. Data from 29 adults were used for training purposes in the ML algorithm. They used the Sumote BOOST and w RACOG algorithms to develop a classification model to reliably detect absolute changes in the score. Various ML methods such as ANN, BN, SVM, and DT are widely used and effective in cancer research for the development of predictive models. Results have been obtained. Accurate decision making [11]. Karaylan et al. [12] proposed a system for predicting heart disease using back propagation algorithms for artificial neural networks. Various machine learning algorithms have been proposed by Chen et al. Optimized for effective prediction of the onset of chronic disease[4] the data collected for training purposes was incomplete. To overcome this, a latent factor model was used.

Various machine learning models were used to study disease prediction in the available input datasets. We used 11 different ML models for the predictions. Of the 11 models, 6 models were able to achieve an accuracy of 50% or more. As shown in Figure 4, the 93.5% weighted ANN model achieved the highest accuracy of all models. Since the value of K changes in this model, the weighted ANN is high and the accuracy is high. This value changed depending on the dataset. In other words, it was a small value and a large value in the training set. Due to this variation, it turned out to be the most accurate model compared to other ML algorithms. Raw data was obtained and sorted by gender, age group and symptoms.

The manuscript presented techniques for predicting illness based on individual patient symptoms, age, and gender. The weighted ANN model provided the highest accuracy for predicting disease at 93.5% using the above factors. Almost all ML models provided excellent accuracy values. Some models were parameter-dependent, so disease could not be predicted and the percentage of accuracy was very low. When an illness is predicted, the medical resources needed for treatment can be easily managed. This model helps reduce the cost of treating the disease and also improves the recovery process.

VI. LIBRARY USED

Standard libraries for database analysis and model creation are used. The following are the libraries used:

1. Nkinter: It's a GUI based python library. It provides powerful object-oriented tool for creating GUI. It provides various widgets to create GUI some of the prominent ones being:

   - Button
   - Canvas
   - Label
   - Entry
   - Check Button
   - List box
   - Message
   - Text
   - Message box

   Some of these were used in this project to create our GUI namely message box, button, label, Option Menu, text and title. Using tkinter we were able to create an interactive GUI for our model.

2. Numpy: Numpy is Python's core scientific computing library. It provides a powerful tool for processing various multidimensional arrays in Python. It is a general purpose array processing package. Numpy's main purpose is to deal with multidimensional homogeneous array. It has tools ranging from array creation to its handling. It makes it easier to create a n dimensional array just by using np.zeros() or handle its contents using various other methods such as replace, arrange, random, save, load it also helps I array processing using methods like sum, mean, std, max, min, all, etc [7]. Array created with numpy also behave differently than arrays created normally when they are operated upon using operators such as +,-,*,/.

   All the above qualities and services offered by numpy array makes it highly suitable for our purpose of handling data. Data manipulation occurring in arrays while performing various operations need to give the desired results while predicting outputs require such high operational capabilities [24].

3. Pandas: For data analysis in Python, we use pandas. It provides highly optimized performance using backend source code written exclusively in C or Python. Data in python can be analyzed with 2 ways:

   - Series
   - Data Frames

Series is a 1-D array defined by a panda and used to store all types of data. Data frames are two-dimensional data structure used in python to store data consisting of rows and columns. Panda's data frame is used extensively in this project to use datasets required for training and testing the algorithms. Data frames make it easier to work with attributes and results. Several of its inbuilt functions such as replace were used in our project for data manipulation and preprocessing [9].

4. Sklearn: Sklearn is a python library with rigging a huge range of visualization algorithms, pre-processing, machine learning, and cross-validation. It features various regression, classification and clustering algorithm such as support vector machine, random forest classifier, decision tree, Gaussian naïve-Bayes, KNN to name a few [8].

   This project used sklearn to take advantage of built-in classification algorithms such as random forest classifiers, decision trees, KNNs, and naive bayes. We also used built-in cross-validation and visualization features such as classification reports, confusion matrices, and accuracy scores.

5. Labels are further used for different sections:

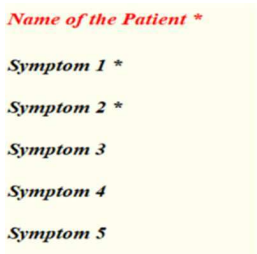

Figure 2. GUI of Inputs Needed for Prediction

Option Menu is used to create drop down menu:

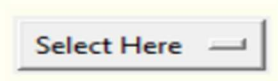

Figure 3. GUI of Select Here Icon

Buttons are used to give functionalities and predict the outcome of models also two utility buttons namely exit and rest are also created [22].

Figure 4. GUI of Different Buttons

Message box are used at three different places, one- to restrain then to enter name.

Figure 5. Message Box GUI

Ask for at least two symptoms,

Figure 6. At least first two Symptoms are needed to Predict GUI

VII. Implementation And Results

A. Presentation of Algorithms on Training data: The system was trained on the medical records of 41 illness-prone patients resulting from a combination of different symptoms. It accounted for 95 of 132 symptoms to avoid over fitting.

The accuracy score of each algorithm after training is:

TABLE I. Accuracy Table

Algorithm Used	Accuracy Score
Decision Tree	0.932927
Random Forest	0.932927
K Nearest Neighbour	0.942927
Naïve Bayes	0.936179

B. Graphical user Interface Result

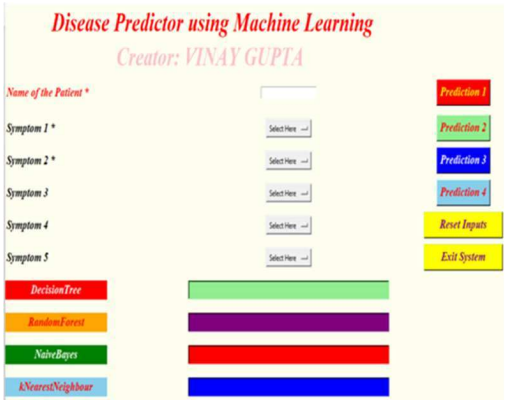

Figure 7. An vacant Disease prediction GUI

The created GUI picks up 5 symptoms from the user. Users can select a symptom from the list of symptoms that appear when they click the none option. Users can show up to 5 symptoms they are facing.

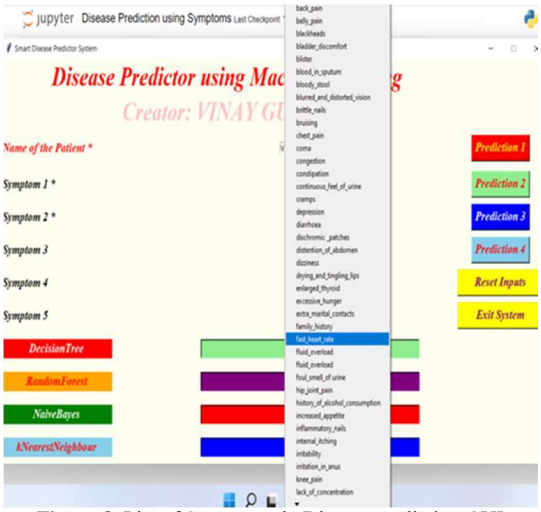

Figure 8. List of Symptoms in Disease prediction GUI

When you have symptoms, select an algorithm. When an algorithm is chosen, the symptoms are process and the infection is searched on the basic of rule which are defined.

The symptoms given by the patient VINAY were: *"depression"*, *"dizziness"*, *"fast heart rate"*, *"lack of concentration"*, *"mild fever"*.

Predictions through algorithms be:

- *Decision Tree:* Hypertension
- *Random Forest:* Hypertension
- *Naïve byes:* Hypertension
- *K Nearest Neighbors:* Hypertension

C. GUI Output

Figure 9. Result in Disease prediction GUI

VIII. CONCLUSION

We wanted to create a system that could predict illness based on given symptoms. Such a system can reduce the rush to the OPD of the hospital and reduce the burden on the medical staff. Such a system was successfully created using four different algorithms. On average, we achieved an accuracy of about 94%. With sufficient data, algorithmic disease prediction is very easy and cheap. Predicting illness based on symptom is an important part of treatment. In our project, we tried to accurately predict the disease based on the patient's symptoms. We used four different algorithms for this purpose and achieved 9295% accuracy. We have also developed an interactive interface to facilitate interaction with the system.

REFERENCES

[1] "UCI Machine Learning Repository." [Online]. Available: https://archive.ics.uci.edu/ml/index.php. [Accessed: 21-Apr- 2018].
[2] J. R. Quinlan, "Induction of Decision Trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106,1986.
[3] T. M. Mitchell, "Decision Tree Learning," *Machine Learning*. pp. 52–80,1997.
[4] P. U. Reesha, Jisha Jose Panackal. "Chapter 44 A Review on Using Predictive Analytics to Determine the Severity of Anaphylaxis" , Springer Science and Business Media LLC, 2022
[5] Sneha Grampurohit, Chetan Sagarnal. "Disease Prediction using Machine Learning Algorithms" , 2020 International Conference for Emerging Technology (INCET), 2020
[6] Pahulpreet Singh Kohli, Shriya Arora. "Application of Machine Learning in Disease Prediction" , 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018
[7] SayantanSaha, Argha Roy Chowdhuriet,al "Web BasedDisease
[8] Detection System",IJERT, ISSN:22780181,Vol.2 Issue 4, April-2013.
[9] Shadab Adam et.al "Prediction system for Heart Disease using Naïve Bayes", International Journal of advanced Computer and Mathematical Sciences, ISSN 2230- 9624, Vol 3,Issue 3,2012,pp 290- 294[Accepted-12/06/2012].
[10] Andrew Alikberov, Stephan Broadly et.al "The Learning Machine", Accessed on: March 26,2020. [Online]. Available: https://www.thelearningmachine.ai.
[11] Min Chen, Yixue Hao et.al "Disease Prediction by MachineLearning over big data from Healthcare Communities", *IEEE*[Access 2017]
[12] https://data-flair.training/
[13] https://docs.python.org/3/library/tkinter.html
[14] Qulan, J.R. 1986. "Induction of Decision Trees". Mach.Learn. 1,1 (Mar. 1986),81-10.
[15] Min Chen, Yixue Hao et.al "Disease Prediction by Machine Learning over big data from Healthcare Communities", IEEE[Access 2017].
[16] Marshland, S. (2009) Machine Learnings an Algorithmic Perspectives. CRC Press, New Zealand, 6-7.
[17] "P. B. Jensen, L. J. Jensen and S. Brunak. Mining electronic health records: Towards better research applications and clinical care".
[18] "Yulei wang1, Jun yang2, Viming.Big Health Application System based on Health Internet of Things and Big Data".
[19] Delen D, Walker G, (2005),"Predicting Dengue Disease survivability: a comparison of three data mining methods", Artificial Intelligence in Medicine, vol. 34, pp. 113-127.
[20] A.S. Monto, S. Gravenstein, M. Elliott, M. Colopy, J. Schweinle, Clinical signs and symptoms predicting influenza infection, Archives of internal medicine 160(21), 3243 (2000).
[21] R.D.H.D.P. Sreevalli, K.P.M. Asia, Prediction of diseases using random forest classification algorithm.
[22] D.R. Langbehn, R.R. Brinkman, D. Falush, J.S. Paulsen, M. Hayden, an International Huntington's Disease Collaborative Group, A new model for prediction of the age of onset and penetrance for huntington's disease based on cag length, Clinical genetics 65(4), 267 (2004).
[23] K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Computational and structural biotechnology journal 13, 8 (2015).

[24] M. Chen, Y. Hao, K. Hwang, L. Wang, L. Wang, Disease prediction by machine learning over big data from healthcare communities, Ieee Access 5, 8869 (2017).
[25] https://github.com/Vinay-gupta9/Disease-Prediction-using-Symptoms
[26] R. Priyadarshini, N. Dash, and R. Mishra, "A Novel approach to Predict Diabetes Mellitus using Modified Extreme Learning Machine," Int. Conf. Electron. Commun. Syst. (ICECS), IEEE, pp. 1–5, 2014.
[27] Shraddha Subhash Shirsath "Disease Prediction Using Machine Learning Over Big Data" International Journal of Innovative Research in Science, Vol. 7, Issue 6, June 2018.
[28] M. Maniruzzaman, M.J. Rahman, B. Ahammed, M.M. Abedin, Classification and prediction of diabetes disease using machine learning paradigm, Health Information Science and Systems 8(1), 7 (2020).
[29] Anish Gupta, Manish K. Gupta, HIVE- Processing Structured Data in HADOOP, International Journal of Scientific & Engineering Research Vol. 8, Issue 6, June2017
[30] Gupta A, Gupta M and Chaturvedi P 2019 Investing data with machine learning using python
[31] Manish Kumar Gupta, Amity Journal of Computational Sciences, Vol 1, Issue 1, pp. 22- 24, 2017
[32] Sneha Grampurohit, Chetan Sagarnal. "Disease Prediction using Machine Learning Algorithms" , 2020 International Conference for Emerging Technology (INCET), 2020.