



Health Disease Prediction Based On Symptoms Using Macchine Learning

Dr. A. Anjaiah

Associate Professor

Dept. of. CSE

St. peter's Engineering College(A)

Hyderabad, TS, India

anjaiah@stpetershyd.com

K. Vidya

IV B.Tech

Dept. of. CSE

St. peter's Engineering College(A)

Hyderabad, TS, India

kvidya1004@gmail.com

V. Sravya

IV B.Tech

Dept. of. CSE

St. peter's Engineering College(A)

Hyderabad, TS, India

vallapsravya.2000@gmail.com

B. V. R. Ankit

IV B.Tech

Dept. of. CSE

St. peter's Engineering College(A)

Hyderabad, TS, India

ankitbandaru@gmail.com

E. vikas Reddy

IV B. Tech

Dept. of. CSE

St. peter's Engineering College(A)

Hyderabad, TS, India

vikasreddyvikky2000@gmail.com

849

Abstract-

General health examination may be a part of healthcare in many countries. Identifying the patients in danger is very important in order to take emergency alert and preventative measures. In order to reduce or eliminate the risk, hospitals can take the required precautions and reducing the possibility of hospital admissions. The elemental challenge of learning a classification model for risk prediction lies within the unlabelled data that constitutes the bulk of the collected dataset. Additionally, because the recent Covid-19 pandemics has delivered to public attention worldwide, comorbidities have relevant structural and economic impact on healthcare .

A practical self-checkup with GUI that provides the public with an early disease prediction[2]. The GUI's event uses a tactic known as the Method Which is based for Healthcare. The majority of the information was acquired from a private clinic, and the doctors closely watched all of the research's findings after a series of tests. This work embeds the machine learning predictive technique that extracts information from existing data sets so as to see and predict future outcomes. during this manuscript, we consider a dataset of 106 symptoms as attributes, from this we are predicting the 43 diseases. After detecting the potential of predicting the disease. We are using KNN algorithm and Naïve bayes for our methodology[3]. Our objectives are precisely predicting from medical records, and supply useful rankings about the foremost important clinical variables available within the dataset.

Keywords—GUI, KNN, naïve bayes, Symptoms, Disease

DOI Number: 10.14704/nq.2022.20.7.NQ33107

NeuroQuantology 2022; 20(7):849-856

I. INTRODUCTION

Dealing with comorbidities is a major task in healthcare nowadays, with multi-morbidity being more the norm than the exception. Self-examination utilising online tools is now a widespread habit among the general public. Despite the abundance of health resources, some of them may not be trustworthy, and false information may harm someone's health and wellbeing, posing a risk to individuals looking for health information[4].

The healthcare sector is more in need of data mining today. when the data mining techniques

are used to extract data from the larger database this will be help to medical practioner to take early decision and improve the health condition. This is used to classify the data to assist the medical practioner[5]. Disease and health problems like malaria, diabetes, psoriasis etc., cause effect on one's health and sometimes it leads to death if ignored. The medical sector make an effective decision making by mining a huge amount of data they process by extraction the patterns and realationships in database. Machine learning algorithms can give a remedy to this problem. So According to the rule, we have created an automated service that can find and



extract secret information about diseases from a historical record.

This paper's major goal is to create a system based on machine learning for prediction and diagnosis based on symptoms. Applying various supervised learning algorithms is the initial step. Comparing and evaluating the algorithms that were used to build a prediction model based on different evaluation metrics serves as the second goal.

II. LITERATURE SURVEY

The healthcare sector is one area where machine learning algorithms are used. the desire exists to improve hospital services Such more obvious decisions regarding patient services and treatment are frequently made. Through clinical insights, computer science in healthcare enables individuals to analyse and interpret complex medical databases. Additionally, by providing patient care, doctors may grasp that more. When used in healthcare, machine learning will increase patient satisfaction. In this software, we are utilising healthcare machine learning capabilities.

When a classifier of disease is deployed utilising such predictive machine learning algorithms outside of treatment, healthcare is made knowledgeable. There may be many instances where a sickness is beyond the control of early identification. This section contains a summary of existing research on the evaluation of prediction of disease. Where the people can check their health condition only at home. Random forest and decision tree algorithms were employed in earlier studies in this area. It can assess disease prognosis, but its biggest flaw is that it takes longer because not all the elements are taken into account.

III. OVER VIEW

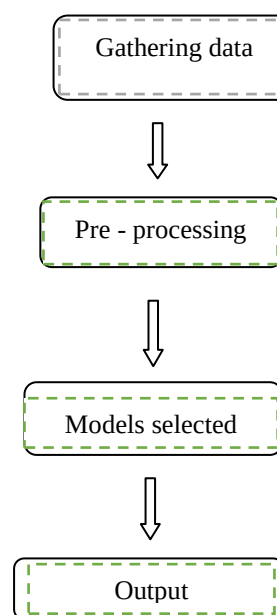
200 symptoms are included in the collection, and their combinations or permutations result in 60 disorders. We intend to develop a model to forecast the patient's disease based on their medical records by gathering their symptoms.

The considered some symptoms are diarrhea, Neck pain Fast heart rate, Mild fever, dizziness, Internal itching, Redness of eyes, Enlarged thyroid, Puffy face and eyes, Sinus pressure, Slurred speech, Hip joint pain, Runny nose Knee

pain, polyuria, congestion, Skin peeling, Family history.

The disease considered are Fungal Infection, Malaria, Varicose veins, Allergy, Chickenpox, Hypothyroidism, Gerd, Dengue, Vertigo, Chronic cholestasis, Peptic ulcer disease, acne, Drug reaction, Hepatitis A, Urinary tract infection.

The whole approach is depiced by following flowchart[6]



850

A. Input

We made the assumption that user is fully cognizant of their symptoms while designing a model. The prediction was created taking 95 symptoms into account in the middle. In this, the user inputs symptoms.

B. Data preprocessing

Data preprocessing turns unstructured data into a format data the computer can easily understand.

- Data cleaning: Data is cleaned using techniques including filling in missing variables and resolving discrepancies in the data.
- Data Reduction: When handling a large amount of data, analysis becomes challenging. Afterward, eliminate the noisy



data and eliminate the independent variables which might not impact on dependent variables.

C. Models Selected

Using KNN and the naive bayes algorithm, the system is trained to anticipate how the sickness would manifest itself. evaluating the effectiveness of the algorithm for the selected data base.

D. Output

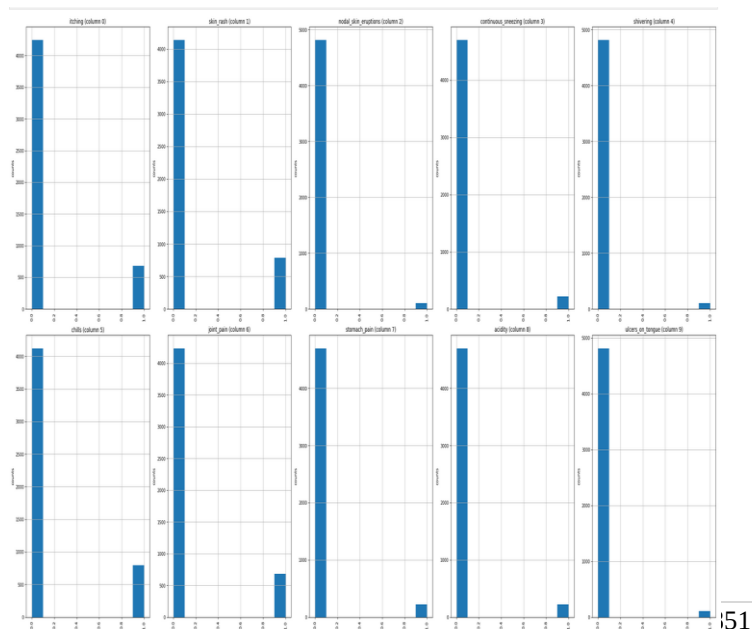
Once the system has been trained by using knn algorithm and the training data. When a user provides a model with symptoms as input, the model processes the symptoms and generates an output.

Data Visualization:

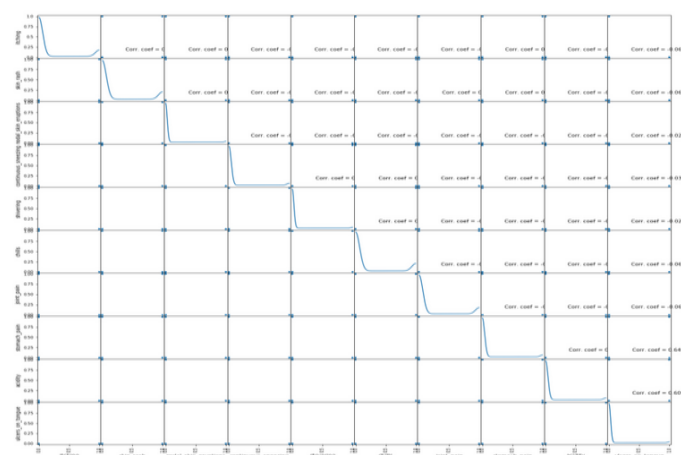
After that the data has been analysed, we will be able to identify the features and labels, therefore based on the In addition to the information provided above, the label must consider the elements that have a substantial impact on or influence disease prediction[7].

We can learn about specific characteristics that are more impacted by the visualisation. of distribution graph and scatter & density plots of the columns.

Distribution column of data:



Scatter and density plot column:



MODEL VALIDATION:

- Confusion matrix

Actual values vs. Predicted values

	0	1
0	TN	FP
1	FN	TP



IV. METHODOLOGY

The two data mining methods are used to create the illness prediction system.

The essential premise of Naïve Bayes is each characteristic:

- Independent
- Equal

Involvement in the result. Due to the fact that it uses less processing resources, it has the advantage of operating quickly even on large datasets.

i) Bayes theorem

The Nave Byes algorithm is based on the following Bayes theorem:

$$P(y/d) = (p(d/y) p(y)) / p(d)$$

Where

$P(y/d)$ = posterior probability

$P(d/y)$ = Likelihood

$P(d)$ = Class prior probability

$P(y)$ = Predictor prior probability

The letters "y" and "d" in the equations above stand for class and characteristics, respectively. $P(\text{denominator})$'s term is a factor of the information (attributes); it does not function of the category we are currently managing. It will therefore be the same for all classes. Traditionally, in naive Bayes Classification, we ignore this denominator since it has no impact on the predictions made by the classifier:

Key Terms:

- The proportion of disease in the dataset under consideration is the prior probability.
- The likelihood of classifying an illness in the presence of additional symptoms is known as likelihood.
- The percentage of symptoms in the dataset under consideration is the marginal probability[8].

Example:

Consider the identical case history,
For a given set of symptoms, we estimate the Naive Bayes outcomes, i.e.

- Chest pain= 1(Symptom is there)
- headache= 0 (Symptom is not there)

- loss of balance= 1
- dizziness=1
- features considered = 4 symptoms=chest pain, head ache, loss of balance, dizziness
- Classes = Hypertension, heart attack

Regarding the dataset we took into account:

1. Likelihood = $P(\text{Feature=symptoms} / \text{Class=Hypertension, Heart attack})$
2. Marginal Likelihood= $P(\text{Features=symptoms})$
3. Prior Likelihood= $P(\text{Class})$

As a result, the prediction is made after viewing the input symptoms by contrasting the posterior probability for each class.

Expression will be used to achieve this. Consequently, allow us to take into consideration the following notations in order to deflate our formula:

- S1 denotes "chest pain," S2 "headache," S3 "balance loss," S4 "dizziness," and H denotes "diseases(class)."

The probability of hypertension is first predicted. (For example, the class= Hypertension with the following input symptoms: "Chestpain = absent"; "headache=present"; "loss of balance=present"; "dizziness=Present")

As a result, the formula becomes:

$$\begin{aligned} P(y=\text{Hypertension} \mid S1=\text{Absent}, S2=\text{Present}, S3=\text{Present}, S4=\text{Present}) &= P(S1=\text{Absent}, S2=\text{Present}, S3=\text{Present}, S4=\text{Present} \mid Y=\text{Hypertension}) * P(Y=\text{Hypertension}) \\ &= P(S1=\text{Absent} \mid Y=\text{Hypertension}) * P(S2=\text{Present} \mid Y=\text{Hypertension}) * P(S3=\text{Present} \mid Y=\text{Hypertension}) * P(S4=\text{Present} \mid Y=\text{Hypertension}) * P(Y=\text{Hypertension}) \\ &= (4/12) * (4 / 12) * (5 / 12) * (5/12) * (8/12) \\ &= 0.01286 \end{aligned}$$

Second, the likelihood of having a heart attack is calculated (using the class=Heartattack and the same input symptoms as indicated in step one above).

$$\begin{aligned} P(Y=\text{Heartattack} \mid S1=\text{Present}, S2=\text{Absent}, S3=\text{Present}, S4=\text{Present}) &= P(S1=\text{Present}, S2=\text{Absent}, S3=\text{Present}, S4=\text{Present} \mid S=\text{Heartattack}) * P(Y=\text{Heartattack}) \\ &= P(S1=\text{Present} \mid S=\text{Heartattack}) * P(S2=\text{Absent} \mid S=\text{Heartattack}) * P(S3=\text{Present} \mid S=\text{Heartattack}) * P(S4=\text{Present} \mid S=\text{Heartattack}) * P(Y=\text{Heartattack}) \end{aligned}$$



$$\begin{aligned}
 &| Y= \text{Heartattack}) * P (S2=\text{Absent} | Y= \\
 &\text{Heartattack}) * P (S3=\text{Present} | Y= \text{Heartattack}) * \\
 &P (S4=\text{Present} | Y= \text{Heartattack}) * P(Y= \\
 &\text{Heartattack}) \\
 &= (3/12) * 0 * (2/12) * (2/12) * (4/12) \\
 &= 0.002348
 \end{aligned}$$

Using the mentioned calculations

$$0.0128 > 0.0023$$

$$--- > P(Y=\text{Hypertension}) > P(Y= \text{Heartattack})$$

Therefore, we shall assume that the data under consideration falls under the category of "Hypertension.", the patient with the symptoms hurting, head ache, loss of balance, dizziness is more likely to possess Hypertension than heartattack.

KNEAREST NEIGHBOR CLASSIFIER:

- Select the K wide variety of the associates
- Calculate the Euclidean distance of K
- Take the K nearest associates point as consistent with the calculated Euclidean distance.
- Among those ok associates, rely the wide variety of the records factors.
- Assign the brand new records factors to that class which the wide variety of the neighbor is maximum.
- Model is ready.
- Fitting K-NN classifier to the Training records:
- Now we healthy the K-NN classifier to the schooling records. To try this we are able to import the KNeighborsClassifier elegance of Sklearn Neighbors library. After uploading the elegance, we are able to create the Classifier item of the elegance. The Parameter of this elegance will be
- `n_neighbors`: To outline the specified associates of the algorithm. Usually, it takes 5.
- `metric='minkowski'`: This is the default parameter and it reveals the gap among the factors.
- `p=2`: It is equal to the usual Euclidean.

- Predicting the Test Result: To expect the check set result, we are able to create a `y_pred`.
- Creating the Confusion Matrix.
- Visualizing the Training set result[9].

V. IMPLEMENTATION AND RESULT:

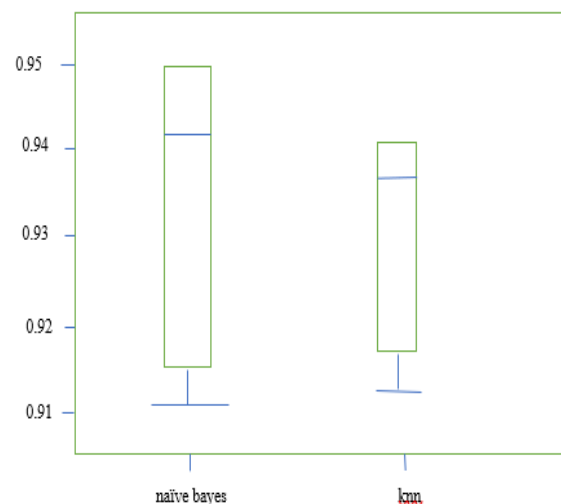
A) Algorithm performance on training data:

The system was trained on medical record of 4920 patients prone to 41 diseases which was due to the combination of various symptoms. We have considered 95 symptoms out of 132 symptoms to avoid overfitting.

We tested the effectiveness of each method on the dataset using the K fold cross validation technique (K=5).

We may conclude from these results that all two methods perform remarkably well on the dataset. However, compared to the other method, Nave Bayes may be performing somewhat better.

Comparision of algorithms:



B) Algorithm performance on test data:

The system was tested after training using the information of 41 medicare applicants



and 95 problems. Following is the accuracy score and confusion matrix:

Accuracy of the algorithm:

Algorithms used	Accuracy score	Confusion matrix correctly classified	Confusion matrix incorrectly classified
Naive bayes	0.9512	39	2
Knearestneighbor	0.9241	37	4

Table 1: Accuracy And Confusion Matrix

We can infer that the algorithms have comparable accuracy ratings. 95.12 percent of the predictions are correct.

A. GUI results:

Empty Disease Prediction GUI

854

Resulting GUI when the hn user given 5 symptoms he is suffering

Once the symptoms are given, the algorithms must be selected. The guidelines from the Process are applied as the algorithms are selected. section are used to process the symptoms and search for the disease based on those symptoms[13].

The symptoms are given by back_pain, chest_pain, dizziness, fast_heart_rate, lack of concentration and the disease prediction by the algorithm are

Knn: hypertension

Naïve bayes:hypertension

The majority of the data, which indicate that patients is much more likely to be have hypertension, can thus be relied upon by the physician.

The accuracy of the algorithms is 95.37%

Correctly classified data is: 39

Graph :



VII. REFERENCES

1. <http://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html> data set
2. Sayantan Saha, Argha Roy Chowdhuri et.al "Web Based Disease Detection System",IJERT, ISSN:22780181,Vol.2 Issue 4, April-2013
3. 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].
4. Min Chen, Yixue Hao et.al "Disease Prediction by Machine Learning over big data from Healthcare Communities", IEEE[Access 2017]
5. Mr Chintan Shah,Dr. Anjali Jivani, "Comparison Of Data Mining Classification Algorithms for Breast Cancer Prediction", IEEE-31661
6. Andrew Alikberov, Stephan Broadly et.al "The Learning Machine", Accessed on: March 26,2020. [Online]. Available: <https://www.thelearningmachine.ai>.
7. 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), Tirunelveli, India, 2019, pp. 292-297, doi: 10.1109/ICOEI.2019.8862604.
8. B. D. Kanchan and M. M. Kishor, "Study of machine learning algorithms for special disease prediction using principal of component analysis," 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), Jalgaon, 2016,pp.510,doi:10.1109/ICGTSPICC.2016.795526
9. 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.

855

VI. CONCLUSION:

The chronological growth of machine learning as well as its uses in the medical industry have regularly shown the rise of systems and approaches that have allowed for the simple and basic interpretation of complex data.

Integrating machine learning algorithms into practise. In this work, two algorithms' efficiency on a case history is thoroughly compared. Both methods produce results that are up to 95% accurate. Performance is assessed using the algorithms' accuracy score and confusion matrix.

We set out to create a system that could forecast sickness based on the symptoms provided. A system like this might lessen pressure on hospital OPDs and lighten the load on medical staff. To attempt the work, thus a system will be generally trustworthy. When developing this method, In order to preserve user input so that it can be utilised to improve the system in the future, we also incorporated a means to do so in the database. A nice user interface is also provided by our technology. It also includes several visual representations of the data gathered and outcomes obtained. Future Enhancement is implementing the model into a interaction of a doctor to administer medication for the actual disease in an exceedingly real situation.

Acknowledgement:

We are thankful to St.Peter's engineering college for helping us with laboratory and continuing support to prepare this paper in a bright manner.



- 1211-1215, doi: 10.1109/ICCMC.2019.8819782.
10. J.P. B. Jensen, L. J. Jensen and S. Brunak, "Mining electronic health records: Towards better research applications and clinical care", *Nature Rev. Genetics*, vol. 13, no. 6, pp. 395-405, 2012.
11. J. T. Chu, M. P. Wang, K. Viswanath, T. H. Lam, and S. S. C. Chan, "How, When and Why People Seek Health Information Online: Qualitative Study in Hong Kong," *Interact J Med Res*, vol. 6, pp. 1-10, 2017.
12. N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension," *IEEE Access: Special Section on Data-Enabled Intelligence for Digital Health*, vol. 7, pp. 144777-144789, 2019.
13. Analytics, J. Liebowitz, Ed., ed New York, 2013. [8] P. Zhang, X. Huang, and M. Li, "Disease Prediction and Early Intervention System Based on Symptom Similarity Analysis," *IEEE Access: Special Section on Data-Enabled Intelligence for Digital Health*, vol. 7, pp. 176484-176494, 2019

