ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 8, August 2024

Implementation of Disease Prediction Algorithm Using Machine Learning

[1] Er. Parasdeep, [2] Dr. Lal Chand

[1] [2] Computer Science and Engineering, Punjabi University Patiala, Punjab, India
Corresponding Author Email: [1] Paras4important@gmail.com, [2] Lc.panwar5876@gmail.com

***Abstract*— The ability to predict human diseases from symptoms is a pivotal aspect of modern healthcare, revolutionizing the way we approach preventive medicine and patient care. The work presented on title "IMPLEMENTATION OF DISEASE PREDICTION ALGORITHM USING MACHINE LEARNING", delves into the importance and implications of disease prediction using symptoms. By harnessing the power of advanced data analytics, machine learning, and artificial intelligence, healthcare providers and researchers can now develop highly accurate predictive models that enable early disease detection. These models analyse an array of patient data, including symptoms, medical history, genetic information, and environmental factors, to identify potential health risks and predict disease onset. Such predictive capabilities have the potential to save countless lives, reduce healthcare costs, and improve overall quality of life for individuals by enabling timely interventions and personalized treatment plans.**

This work highlights the challenges and future prospects of disease prediction from symptoms. While significant strides have been made in this field, challenges remain in terms of data privacy, model interpretability, and the need for robust and diverse datasets. In this work, naive bayes has been implemented for disease prediction since it works better on the textual dataset according to best of my knowledge. Disease prediction from symptoms represents a transformative approach to healthcare, paving the way for a proactive and personalized healthcare system that not only treats diseases but also anticipates and prevents them, ultimately improving public health outcomes worldwide.

***Index Terms*— Python, Machine learning, Naïve bayes, decision tree, Data.**

I. INTRODUCTION

In the realm of healthcare, the race to detect diseases early is crucial for saving lives. Thanks to rapid advancements in technology, especially in data science and machine learning, predicting diseases from symptoms has become a game-changer. This shift doesn't just empower healthcare pros—it also gives people valuable insights into their health.

Several studies have delved into this exciting field. For example, Deepthi et al. (2020) focused on using machine learning to predict diseases based on symptoms, while Singh et al. (2017) explored how big data analytics could revolutionize healthcare decision-making. Other researchers, like Dahiwade et al. (2019) and Chen et al. (2017), have also investigated the potential of machine learning in disease prediction.

One particularly intriguing study looked at diagnosing Parkinson's disease early using explainable machine learning models. By making the models easier to understand, the hope is to improve diagnosis accuracy. Another study, by Wu et al. (2023), explored personalized medical recommendations using similar approaches.

In this Paper, we'll dive into the fascinating world of disease prediction from symptoms using Python and machine learning. We'll explore the methods, tools, and best practices involved in building predictive models. Ultimately, we aim to showcase how this innovative approach is reshaping healthcare, offering hope for earlier disease detection and prevention.

II. LITERATURE SURVEY

Automated medical diagnosis using machine learning has become a promising tool for improving the accuracy and efficiency of diagnosing diseases. This Paper explores several studies in this field.

One study focuses on predicting diseases based on symptoms using machine learning algorithms like Naive Bayes, Decision Tree, and Random Forest, implemented in Python. The research evaluates these algorithms' accuracy on a provided dataset to identify the best-performing one [1].

Another paper discusses utilizing big data analytics and Hadoop clusters to analyze real-time data for predicting severe emergency cases. It highlights how such insights can transform healthcare outcomes [2].

Magesh et al. (2020) explore using machine learning techniques, particularly LIME, for early Parkinson's disease detection through medical imaging analysis. They stress the importance of explainable AI models for trust in automated diagnostics [8].

Similarly, another study focuses on interpretable machine learning for personalized medical recommendations, aiming to provide transparent recommendations to patients [9].

The review also discusses the significance of machine learning in Computer-Aided Diagnosis (CAD), emphasizing its role in enhancing diagnostic accuracy [3].

Addressing the challenge of accurate disease prediction based on symptoms, another study proposes a general disease prediction model employing machine learning algorithms

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 8, August 2024

like K-Nearest Neighbour (KNN) and Convolutional Neural Network (CNN) [4].

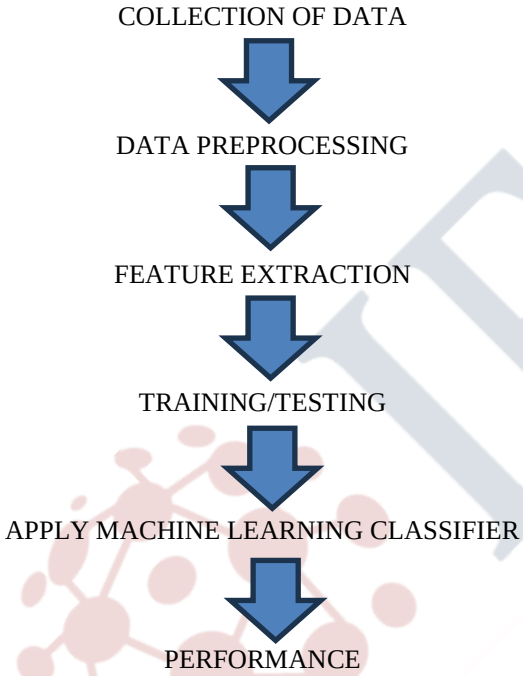
The incorporation of machine learning algorithms in healthcare and bioscience is emphasized for extracting valuable information from diverse datasets. This research aims to predict diseases early, supporting patient care and community services [5].

Additionally, the review touches upon the importance of accurate and timely prediction of heart disease, employing the Naïve Bayes algorithm for analysis [6].

Furthermore, the study discusses utilizing machine learning algorithms to analyze big data from healthcare communities for disease prediction, emphasizing the benefits of predictive analytics for early detection and prevention [7].

III. METHODOLOGY

To predict disease, we need follow the steps written bellow:



IV. DATASET

I gathered data on 132 symptoms and 41 diseases from patient records.

The considered symptoms are:

Table 1

Abdominal pain	Depression	Passage of gases
Back pain	Bloody stool	scurrying
Constipation	Irritation in anus	Weakness in limbs
Diarrhea	Neck pain	Fast heart rate
Mild fever	Dizziness	Internal itching
Yellow urine	Cramps	Toxic look
Yellow of eyes	Bruising	Palpitation
Acute liver failure	Obesity	Painful walking
Fluid overload	Swollen legs	Prominent veins on calf
Swelling of stomach	Irritability	Fluid overload
Swelled lymph nodes	Swollen blood vessels	Excessive hunger
Malaise	Muscle pain	Black heads
Blurred and distorted vision	Pain in anal region	Pain during bowel movements
Phlegm	Brittle nails	Rusty sputum
Throat irritation	Belly pain	Mucoid sputum
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
Chest pain	Extra marital contact	Swollen extremities
Yellow crust ooze	Swelling joints	Coma
Loss of smell	Stiff neck	Unsteadiness
Movement stiffness	Muscle weakness	Drying and tingling lips
Spinning movements	Red sore around nose	Weakness of one body side
Bladder discomfort	Foul smell of urine	Continuous feel of urine
Altered sensorium	Red spots over body	Abnormal menstruation
Dyschromic patches	Watering from eyes	Increases appetite
Lack of concentration	Visual disturbance	receiving blood transformation
Receiving unsterile injections	Blood in sputum	Stomach bleeding
Distention of abdomen	History of alcohol consumption	Puss filled pimples
Silver like dusting	Small dents in nails	Inflammatory nails
blister		

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 8, August 2024

The diseases that are contemplate:

Table 2

Fungal infection	Malaria	Varicose veins
Allergy	Chickenpox	Hypothyroidism
Gerd	Dengue	Vertigo
Chronic cholestasis	Peptic ulcer disease	Acne
Drug reaction	Hepatitis A	Urinary tract infection
Piles	Hepatitis B	Psoriasis
AIDS	Hepatitis C	Impetigo
Diabetes	Hepatitis D	Bronchial asthma
Alcoholic hepatitis	Cervical spondylosis	Hypertension
Tuberculosis	Arthritis	Migraine
Common cold	Osteoarthritis	Paralysis
Pneumonia	Typhoid	Jaundice
Heart attack	Gastroenteritis	Hepatitis E
Hypoglycemia	Impetigo	

Example:

Let's examine the medical history of 12 patients who had a history of malaria and dengue fever. They exhibit the following symptoms: a high temperature, shivering, vomiting, and pain muscles.

Sample

High Fever	Vomiting	Shivering	Muscle Paining	Disease
1	0	1	0	Dengue
1	1	0	0	Malaria
0	1	0	1	Malaria
0	0	1	1	Dengue
1	0	1	1	Dengue
1	1	0	1	Dengue
1	1	1	1	Malaria
0	1	1	1	Dengue
1	1	1	0	Malaria
0	1	1	0	Dengue
1	0	0	1	Dengue
0	1	0	0	Dengue

Dengue	Malaria	Total
8	4	12

A. Naïve Bayes Classifier:

Naïve: The term comes from the simplistic assumption that the presence of one feature is independent of the presence of other features.

Bayes: The reason it's named Bayes is that it relies on the Bayes' Theorem.

Naïve bayes works on the basis of bayes theorm:

$$P(x/y)= P\frac{P\left(\frac{y}{x}\right)P(x)}{P(y)}$$

Where

P(x/y) = Posterior probability

P(y/x) = Likelihood

P(x) = Class prior probability

P(y) = Predictor Prior Probability

"x" stands for class and "h" for features in the calculation above. The lone term in P(y)'s denominator is a function of the data (features); it is not a function of the class.

We are presently addressing. It will therefore be the same for every class. Typically, in naïve Bayes classification, we disregard this denominator since it has no bearing on the classifier's prediction-making outcome:

$P(x/y) \alpha P(y/x) P(x)$

Important Words:

The percentage of disease in the examined data set is known as the prior probability.

The likelihood of classifying an illness in the presence of additional symptoms is known as its likelihood.

Marginal Likelihood is the percentage of symptoms in the dataset under consideration.

Example:

- High fever= Present (denoted by vale'1')
- Vomiting=Absent (denoted by vale '0')
- Shivering=Present (denoted by value '1')
- Muscle wasting=Present (denoted by value '1')

From the Table (1):

- Four symptoms—high fever, vomiting, shivering, and muscle atrophy—were taken into consideration.
- Classes: Malaria, Dengue

As per the dataset we assume:

- Likelihood = P (Feature=symptoms /
- Class=Dengue, Malaria)
- Marginal Likelihood= P(Features=symptoms)
- Prior Likelihood= P(Class)

Thus, let us examine the following notations to deflate our formula: "F1" for "High fever," "F2" for "Vomiting," "F3" for "Shivering," "F4" for "Muscle Wasting," and "D" for "Diseases (class)."

Firstly, the probability for Dengue is estimated (i.e. the class=Dengue with input symptoms as follows: "High fever=Present"; "Vomiting=Absent"; "Shivering=Present"; "Muscle Wasting=Present").

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 8, August 2024

The formula thus modifies to:

P (X=Dengue | F1=Present, F2=Absent, F3=Present, F4=Present) = P (F1=Present, F2=Absent. F3=Present, F4=Present)

| X=Dengue) * P(X=Dengue)

= P (F1=Present | X=Dengue) * P (F2=Absent |

X=Dengue) * P (F3=Present | X=Dengue) * P (F4=Present|

X=Dengue) * P(X=Dengue)

$= \frac{4}{12} * \frac{4}{12} * \frac{5}{12} * \frac{5}{12} * \frac{8}{12}$

= 0.01286

Secondly, the probability for Malaria is estimated (i.e. the class=Malaria with the same input symptoms as mentioned in above step)

P (X=Malaria | F1=Present, F2=Absent, F3=Present,

F4=Present) = P (F1=Present, F2=Absent. F3=Present, F4=Present | X=Malaria) * P(X=Malaria) = P (F1=Present | X=Malaria) * P (F2=Absent | X=Malaria) * P (F3=Present | X=Malaria) * P (F4=Present | X=Malaria) * P(X=Malaria)

$= \frac{3}{12} * 0 * \frac{2}{12} * \frac{2}{12} * \frac{4}{12}$

= 0.002348

Result:

0.0128 > 0.0023 --- > P(X=Dengue) > P(X=Malaria)

Therefore, we can predict with the assumed dataset belongs to “Dengue”

B. Random Forest Classifier

The Random Forest Classifier is a versatile and user-friendly machine learning algorithm known for its consistently impressive performance, often without the need for extensive fine-tuning. Unlike Decision Trees, which can suffer from overfitting by essentially memorizing the training data, Random Forest offers a solution to this issue through ensemble learning. Ensemble learning involves using multiple instances of the same algorithm or different algorithms together.

In the case of Random Forest, it operates by creating a team of Decision Trees. The more Decision Trees in the forest, the better the model's ability to generalize to new data. Here's how it works:

1. Randomly selects a subset of symptoms (k) from the dataset (which may contain a larger set of symptoms, m) and constructs a Decision Tree based on this subset.
2. Repeats this process (bootstrap sampling) n times, generating n Decision Trees, each from a different random combination of symptoms.
3. Each Decision Tree is then used to predict the disease based on a random sample of the data. The predicted diseases from all the trees are recorded.
4. Finally, the algorithm tallies up the votes for each predicted disease and selects the one that appears most frequently (mode) as the final prediction.

C. Decision Tree Classifier

Naive Bayes often outperforms Decision Trees when it comes to textual data. Naive Bayes, a method rooted in Bayes' theorem, is effective for text categorization due to its simplicity and ability to process high-dimensional data. While it may be oversimplifying, the idea of feature independence is effective in texts where word frequency is an important factor. In spite of their flexibility, decision trees can overfit and struggle with the complex, high-dimensional spaces found in text data.

V. IMPLEMENTED SYSTEM

If we compare these two algorithms technically then naïve bayes proves to be a good and accurate because it works fine on the textual database.

Furthermore, researchers have investigated the use of natural language processing techniques for automated diagnosis from electronic health records. These techniques can help automate.

The diagnosis process by extracting relevant information from patient records and generating potential diagnoses based on this information.

The proposed system has the potential to significantly improve the accuracy and efficiency of medical diagnosis and can be used in both developed and developing countries to provide access to quality healthcare.

VI. CONCLUSION

Looking at the evolution of machine learning and its applications in healthcare, we see the emergence of systems and methods that make complex data analysis simple using machine learning algorithms. This study provides a thorough comparison of three algorithms' performance on medical records, with each achieving an impressive accuracy of up to 95 percent. The analysis includes examining confusion matrices and accuracy scores.

As technology continues to generate and store vast amounts of data, Artificial Intelligence is poised to play an even more significant role in data analysis. This is evident from the growing importance of AI in handling the immense data produced by modern technology.

REFERENCE

[1] Deepthi, Y., Kalyan, K.P., Vyas, M., Radhika, K., Babu, D.K. and Krishna Rao, N.V., 2020. Disease prediction based on symptoms using machine learning. In *Energy Systems, Drives and Automations: Proceedings of ESDA 2019* (pp. 561-569). Singapore: Springer Singapore.

[2] Singh, M., Bhatia, V. and Bhatia, R., 2017, December. Big data analytics: Solution to healthcare. In 2017 International conference on intelligent communication and computational

techniques (ICCT) (pp. 239-241). IEEE.

[3] Hamsagayathri, P. and Vigneshwaran, S., 2021, February. Symptoms based disease prediction using machine learning techniques. In 2021 Third international conference on intelligent communication technologies and virtual mobile networks (ICICV) (pp. 747-752). IEEE.

[4] Dahiwade, D., Patle, G. and Meshram, E., 2019, March. Designing disease prediction model using machine learning approach. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1211-1215). IEEE.

[5] Singh, P., Singh, N., Singh, K.K. and Singh, A., 2021. Diagnosing of disease using machine learning. In Machine learning and the internet of medical things in healthcare (pp. 89-111). Academic Press.

[6] Pattekari, S.A. and Parveen, A., 2012. Prediction system for heart disease using Naïve Bayes. International journal of advanced computer and mathematical sciences, 3(3), pp.290-294.

[7] Chen, M., Hao, Y., Hwang, K., Wang, L. and Wang, L., 2017. Disease prediction by machine learning over big data from healthcare communities. Ieee Access, 5, pp.8869-8879.