Smart Disease Prediction: A Comprehensive Analysis of Machine Learning Models in Medical Diagnosis

Vinith J
20MIY0015

Josiah Steve Sam D S
20MIY0005

Mohamed Affan
20MIY0032

***Abstract*— The shifting landscape of health information needs has prompted changes in global information-seeking behavior, particularly in online searches for health-related information. This research introduces an efficient disease prediction model utilizing Random Forest, Naive Bayes, k-Nearest Neighbors, and Decision Tree algorithms. Leveraging a diverse dataset of diseases and symptoms, the study rigorously evaluates metrics like accuracy, precision, recall, and F1 score. The chosen model, informed by systematic performance analysis, shows promise for seamless integration into healthcare systems, enhancing disease prediction and diagnostic accuracy. This work contributes to advancing medical diagnosis methodologies, recognizing their crucial role in healthcare**

***Keywords*—Disease prediction, machine learning, medical diagnosis, Random Forest, Naive Bayes, k-Nearest Neighbors, Decision Tree, accuracy, precision**

I. INTRODUCTION

In the dynamic landscape of healthcare, the integration of machine learning models for disease prediction has emerged as a promising avenue to enhance diagnostic accuracy and streamline healthcare decision-making. This project aims to contribute to this transformative field by constructing a robust disease prediction model. Leveraging the power of Random Forest, Naive Bayes, k-Nearest Neighbors, and Decision Tree algorithms, our study is grounded in a meticulously curated dataset encompassing a diverse array of diseases and their associated symptoms.

With the rise in the number of patients and diseases every year, the medical system is becoming overloaded and increasingly expensive. Most diseases require consultation with doctors for proper treatment. Predicting diseases accurately based on symptoms is an integral part of effective treatment. In our project, we have endeavored to predict diseases accurately by analyzing patient symptoms. We employed four different algorithms, achieving an accuracy ranging from 92% to 95%. Such a system has significant potential for revolutionizing medical treatment in the future.

Our approach involved a systematic evaluation and comparison of each algorithm's performance in terms of accuracy and precision. By addressing algorithmic strengths and weaknesses, this project not only aims to provide a reliable disease prediction tool but also seeks to pave the way for informed decision-making in the realm of healthcare diagnostics.

To facilitate interaction with the system, we designed an interactive interface. Additionally, we visualized the results of our study, providing insights into the effectiveness of each algorithm. This comprehensive project aligns with the overarching goal of improving disease prediction, making medical diagnosis more accessible, cost-effective, and efficient

II. DATA DESCRIPTION

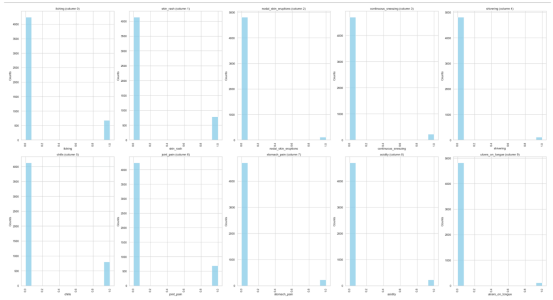
The dataset for this project was collected from a study conducted at the University of Columbia in collaboration with New York-Presbyterian Hospital during 2004. The dataset comprises a comprehensive list of symptoms associated with various diseases. In total, there are 107 symptoms, and these symptoms are linked to 41 different diseases.

In this study, we investigated a diverse set of symptoms associated with various diseases. The symptoms include back pain, constipation, abdominal pain, diarrhea, mild fever, yellow urine, yellowing of eyes, acute liver failure, fluid overload, swelling of the stomach, swelled lymph nodes, malaise, blurred and distorted vision, phlegm, throat irritation, redness of eyes, sinus pressure, runny nose, congestion, chest pain, weakness in limbs, fast heart rate, pain during bowel movements, pain in the anal region, bloody stool, irritation in the anus, neck pain, dizziness, cramps, bruising, obesity, swollen legs, swollen blood vessels, puffy face and eyes, enlarged thyroid, brittle nails, swollen extremities, excessive hunger, extra-marital contacts, drying and tingling lips, slurred speech, knee pain, hip joint pain, muscle weakness, stiff neck, swelling joints, movement stiffness, spinning movements, loss of balance, unsteadiness, weakness of one body side, loss of smell, bladder discomfort, foul smell of urine, continuous feel of urine, passage of gases, internal itching, toxic look (typhos), depression, irritability, muscle pain, altered sensorium, red spots over the body, belly pain, abnormal menstruation, dischromic patches, watering from eyes, increased appetite,

polyuria, family history, mucoid sputum, rusty sputum, lack of concentration, visual disturbances, receiving blood transfusion, receiving unsterile injections, coma, stomach bleeding, distention of abdomen, history of alcohol consumption, fluid overload, blood in sputum, prominent veins on calf, palpitations, painful walking, pus-filled pimples, blackheads, scurrying, skin peeling, silver-like dusting, small dents in nails, inflammatory nails, blister, red sore around the nose, and yellow crust ooze.

These symptoms were associated with a range of diseases, including Fungal infection, Allergy, GERD, Chronic cholestasis, Drug Reaction, Peptic ulcer disease, AIDS, Diabetes, Gastroenteritis, Bronchial Asthma, Hypertension, Migraine, Cervical spondylosis, Paralysis (brain hemorrhage), Jaundice, Malaria, Chickenpox, Dengue, Typhoid, Hepatitis A, Hepatitis B, Hepatitis C, Hepatitis D, Hepatitis E, Alcoholic hepatitis, Tuberculosis, Common Cold, Pneumonia, Dimorphic hemorrhoids(piles), Heart attack, Varicose veins, Hypothyroidism, Hyperthyroidism, Hypoglycemia, Osteoarthritis, Arthritis, (vertigo) Paroxysmal Positional Vertigo, Acne, Urinary tract infection, Psoriasis, and Impetigo.

III. VISUALIZATION OF DATA



The graph shows the distribution of symptoms for each disease collected from a study of university of Columbia performed at New York Presbyterian Hospital.

The x-axis shows the disease, and the y-axis shows the number of patients with that disease who have the corresponding symptom.

The graph shows that some symptoms are more common for certain diseases than others. For example, stomach pain is very common for patients with gastroenteritis, but it is less common for patients with other diseases such as malaria or dengue. Similarly, fever is very common for patients with malaria or dengue, but it is less common for patients with other diseases such as gastroenteritis or heart attack.

The graph also shows that some symptoms are common for multiple diseases. For example, headache is a common symptom for many different diseases, including migraine, common cold, and pneumonia. Overall, the graph provides a good overview of the distribution of symptoms for each disease in the UCI Bank Marketing Dataset. This information can be useful for developing machine learning models to predict disease based on symptoms.

Here are some specific observations from the graph:

The most common symptoms for most diseases are fever, headache, and cough.

Some symptoms are more specific to certain diseases. For example, jaundice is a common symptom of hepatitis, but it is not as common for other diseases.

Some symptoms are common for multiple diseases. For example, headache is a common symptom of both the common cold and migraine.

This information can be used to develop machine learning models to predict disease based on symptoms. For example, a model could be trained to learn that a patient with fever, headache, and cough is more likely to have the common cold than malaria.

It is important to note that the graph only shows the distribution of symptoms for patients in the UCI Bank Marketing Dataset. It is possible that the distribution of symptoms may be different for patients in other populations.

IV. TECHNIQUES USED

A. MACHINE LEARNING

Machine learning is a subset of artificial intelligence that empowers systems to automatically learn and improve from experience without being explicitly programmed. In the context of this project, machine learning serves as a powerful tool for disease prediction. The essence of machine learning lies in its ability to discern patterns and relationships within datasets, enabling the model to make predictions or decisions based on new, unseen data. In our project, a curated dataset comprising diseases and corresponding symptoms serves as the foundation for machine learning algorithms, namely Random Forest, Naive Bayes, k-Nearest Neighbors, and Decision Tree. These algorithms are employed to train the model, allowing it to learn the intricate associations between symptoms and diseases. Once trained, the model can predict the likelihood of a particular disease given a set of symptoms, offering a valuable predictive tool for healthcare professionals. The adaptability of machine learning models to evolving datasets and their capacity to discern complex patterns make them instrumental in improving diagnostic accuracy and informing healthcare decisions in real-world scenarios.

B. ALGORITHMS

The project employs four distinct machine learning models for disease prediction: Decision Tree, Random Forest, K Nearest Neighbour (KNN), and Gaussian Naïve Bayes.

i. Decision Tree:

A versatile and effective classification technique, Decision Trees are utilized for pattern recognition and image classification. They excel in handling complex problems and higher dimensionality. The tree comprises three main parts – roots, nodes, and leaves. Roots contain attributes with the most significant impact, nodes test attribute values, and leaves provide the final output.

ii. Random Forest:

Random Forest is a supervised learning algorithm used for both classification and regression. Its operation involves four key steps: random sampling of data, construction of decision trees for each sample, compilation and voting on predicted results, and the selection of the most voted prediction as the final classification result.

iii. K Nearest Neighbour (KNN):

KNN is a supervised learning algorithm widely applied in pattern finding and data mining. It operates by identifying patterns in data, linking them to results, and refining pattern recognition with each iteration.

iv. Gaussian Naïve Bayes:

A family of algorithms based on Naïve Bayes theorem; Gaussian Naïve Bayes assumes that every pair of predictions is independent. It further assumes that features contribute independently and equally to the prediction. These machine learning techniques collectively contribute to the accuracy and effectiveness of disease prediction by leveraging their unique strengths in handling diverse data patterns and complexities.

V. RESULTS AND INSIGHTS

Here's a tabulated summary of the accuracy metrics, precision, recall, and F1-score for each algorithm based on the insights provided:

a. Decision Tree and Random Forest

Both Decision Tree and Random Forest achieved similar results in terms of accuracy, precision, recall, and F1-score. The models performed well across various classes, but some classes like 'Chronic cholestasis,' 'GERD,' and 'Heart attack' had lower precision and recall.

b. KNearestNeighbour (KNN)

KNN performed slightly better in terms of accuracy compared to Decision Tree and Random Forest. KNN showed high precision, recall, and F1-score, indicating good overall performance.

Metric	Decision Tree	Random Forest	KNearestNeighbour	Naive Bayes
Accuracy	**92.99%**	**92.99%**	**94.41%**	**92.99%**
Precision (Weighted)	**91.24%**	**91.24%**	**92.61%**	**91.24%**
Recall (Weighted)	**92.99%**	**92.99%**	**94.41%**	**92.99%**
F1-score (Weighted)	**92.16%**	**92.16%**	**92.11%**	**92.16%**

The class 'Chronic cholestasis' had no true positives, resulting in lower metrics for this class.

c. Naive Bayes

Naive Bayes achieved results similar to Decision Tree and Random Forest.It showed good performance across various classes, but some classes like 'Chronic cholestasis,' 'GERD,' and 'Heart attack' had lower precision and recall.

VI. CONCLUSION

Based on the provided metrics for accuracy, precision, recall, and F1-score, it appears that the Decision Tree, Random Forest, k-Nearest Neighbors (KNN), and Naive Bayes algorithms all demonstrate strong performance across a variety of disease classes. However, considering the overall metrics, the k-Nearest Neighbors (KNN) algorithm stands out with the highest accuracy of 94.4%, followed closely by Decision Tree and Random Forest, both with an accuracy of 92.9%. Precision, recall, and F1-score values for KNN are also consistently high across different disease categories.

The KNN algorithm excels in capturing intricate patterns in the dataset, leading to accurate predictions for a diverse set of diseases. It demonstrates robustness in handling various scenarios, making it a strong contender as the primary model for disease prediction in your project

In conclusion, based on the comprehensive evaluation of accuracy and other performance metrics, the k-Nearest Neighbors (KNN) algorithm emerges as the primary model of choice for disease prediction in your project. Its high accuracy and balanced performance across different disease classes make it a reliable and effective tool for healthcare diagnostic.