iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44408

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Disease Predictor Based on Symptoms Using Machine Learning

M Vamshi Krishna Reddy[1], G V P Sai Abhijith[2], K Sai Nath[3], Mangali Sathyanarayana[4]
[1, 2, 3]*B.Tech IV year Students,* [4]*Assistant Professor, Department of Computer Science and Engineering (CSE), Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India.*

Abstract: ***Given how important the health sector is in curing prescribers' problems, Disease Prediction based on Symptoms with Machine Learning is a system that predicts diseases based on the user's knowledge of clinical manifestations, assuring solid findings based on such facts. If the user simply has to know a little bit about the sickness and the patient isn't in any danger, this technique can be used to learn a little bit about minor ailments. It's a system that provides medical advice and tactics to clients, as well as a tool to help them figure out what ailment they have utilizing this forecast. It's also a big benefit for the healthcare sector as well as individuals who don't want to travel to a hospital or clinic for their initial diagnosis. The user can learn a lot about the condition that has been revealed to him or her by simply inputting the side effects and other critical information, and the health sector can benefit from this method by simply asking the patient for symptoms and giving them a diagnosis. To achieve Disease Prediction based on Symptoms, we used Machine Learning techniques, Python Programming with Tkinter Interface, and a dataset acquired from hospitals.***
The phrases Disease Predictor, Machine Learning, and Tkinter Interface are used in this research.
Keywords: Disease Predictor, Machine Learning, Tkinter Interface

I. INTRODUCTION

A well-functioning healthcare system is critical to the economy and the well-being of humanity. Between the world, we live in now and the world we lived in a few decades ago, there has been a substantial amount of change. Everything has gotten more disorderly and unattractive. In this situation, doctors and nurses are doing everything they can to save people's lives, even if it means putting their own lives in danger. Virtual doctors are board-certified doctors who choose to practice online using video and phone consultations rather than in-person consultations, albeit this is not always practicable in an emergency. In the absence of human error, machines are thought to be superior to humans because they can do jobs faster while maintaining a consistent level of precision. A disease predictor, often known as a virtual doctor, may accurately predict a patient's sickness without the need for human involvement. A disease predictor can save a person's life in extreme instances, such as COVID-19 and EBOLA, by recognizing their health without requiring physical touch. There are virtual doctors on the market now, but they lack the capacity to provide the kind of precision that is required. This Condition's Prognosis To forecast sickness, we'll use hospital data and Machine Learning methods based on the Python programming language and the Tkinter interface. Doctors may make errors when diagnosing a patient's disease, however, disease prediction systems with machine learning algorithms can help produce accurate results in these situations. For this project, we employed a mix of approaches, algorithms, and technologies to develop a system that can forecast a patient's status based on their symptoms. The symptoms are compared to the information previously saved in the system. We can accurately forecast the percentage of disease in a patient by combining those datasets with the patient's symptoms. The dataset and symptoms are uploaded to the system's prediction model, where the data is pre-processed for future references before the user picks the features and enters the symptoms. The data is then classified using a variety of algorithms and approaches, such as Decision Tree, KNN, and Naive Bayes, to mention a few.

II. LITERATURE SURVEY

Machine learning algorithms have been used in several studies to forecast diseases based on a person's symptoms. Monto et al. [6] created a statistical model that could predict whether or not a patient had influenza. The study comprised 3744 unvaccinated adult and adolescent influenza patients who had a fever and at least two additional flu symptoms. 2470 of the 3744 persons tested positive for influenza, according to the findings. Their model had a 79 percent accuracy rate based on this data.

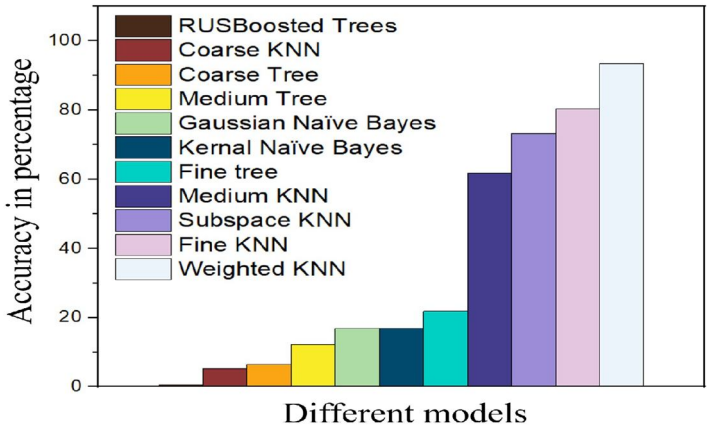


Fig. 1 Accuracy comparison graph

Fig.1 Comparison of the accuracy values of the different ML algorithms. The Weighted KNN model gave the highest accuracy as compared to the other ML algorithms. The RUSBoosted trees were the least accurate model. The Fine KNN performed better than the Subspace, Medium, and Coarse KNN models. The least efficient KNN model was coarse KNN. The Gaussian and the Kernel Naive Bayes algorithm had a comparable accuracy with each other though less than the KNN models. The Fine tree had a higher accuracy than the medium and the coarse decision tree models.

III. PROPOSED SYSTEM

In the suggested strategy, we use Machine Learning techniques to precisely forecast the ailment that the patient has been suffering from. When past healthcare records are used as a dataset, the results are more accurate. To train the model and predict user diseases based on the symptoms they enter, we use machine learning algorithms.

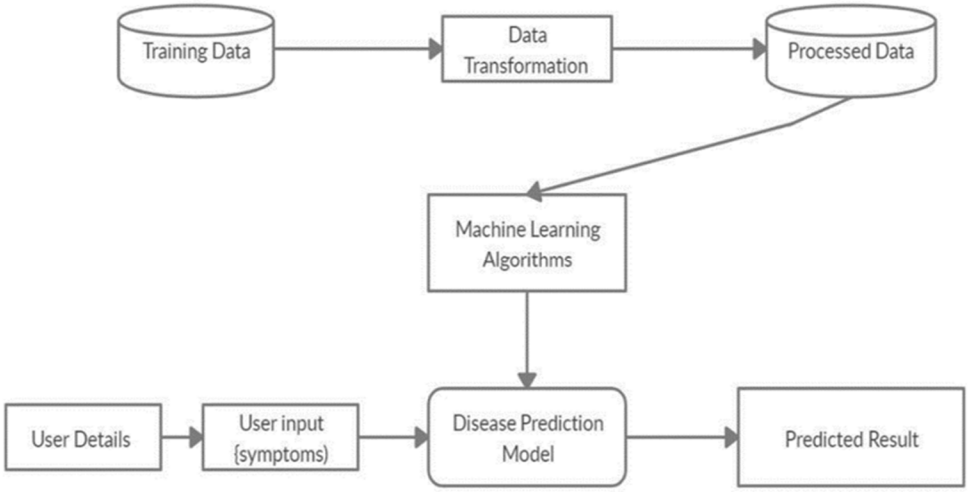


Fig. 2 System architecture

A. Advantages of Proposed System

The Proposed System's Advantages

1) First and foremost, seeing a doctor for modest treatment is unnecessary.
2) When compared to past treatments, you'll get more precise results.
3) Only a few risk variables are at play.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

IV. DATASET

This inquiry used data from a University of Columbia study done at New York-Presbyterian Hospital in 2004. The ailment is given in the first column, followed by the symptom in the second column. The strongest connections between the 150 most prevalent diseases have been identified, and symptoms have been categorized according to the strength of the connections. The technology created UMLS codes for diseases and symptoms using the MedLEE natural language processing system, which were then examined using statistical techniques based on frequency and co-occurrences.

A. Features of this Dataset

1) 'Migraine,' -'Cervical spondylosis,' -'Paralysis (brain hemorrhage),' -'Jaundice,' -'Malaria,' -'Chickenpox,' -'Dengue,' -'Cervical spondylosis,' -'Cervical spondylosis,' -' Cervical s

2) "Typhoid," "Hepatitis A," "Hepatitis B," "Hepatitis C," "Hepatitis D," "Hepatitis E," "Hepatitis F," "Hepatitis G," "Hepatitis H," "Hepatitis I," "Hepatitis J," "Hepatitis K," "Hepatitis L," "Hepatitis M," "Hepatitis N," "Hepatitis O," "

3) 'Alcoholic hepatitis,' 'Tuberculosis,' 'Alcoholic hepatitis,' 'Alcoholic hepatitis,' 'Alcoholic he

4) 'Pneumonia,' 'Common Cold,'

5) 'Heartattack,' 'Dimorphic hemorrhoids(piles),'

6) 'Varicose veins,' 'Hypothyroidism,' 'Varicose veins,' 'Varicose veins,' 'Varicose vein

7) 'Hyperthyroidism,' 'Hypoglycemia,' 'Hyperthyroidism,' 'Hyperthyroidism,' 'Hyperthyroidism,'

8) 'Osteoarthritis,' 'Arthritis,' 'Osteoarthritis,' 'Osteoarthritis,' 'Osteoarthritis,'

9) '(vertigo) Paroymsal Positional Vertigo', '(vertigo) Paroymsal Positional Vertigo', '(vertigo) Paroyms

10) 'Acne,' 'Urinary tract infection,' and so on.

11) 'Psoriasis' and 'Impetigo' are two terms for the same thing.

B. Images from Train and test Datasets

	itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	...	blackheads	scu
0	1	1	1	0	0	0	0	0	0	0	...	0	
1	0	1	1	0	0	0	0	0	0	0	...	0	
2	1	0	1	0	0	0	0	0	0	0	...	0	
3	1	1	0	0	0	0	0	0	0	0	...	0	
4	1	1	1	0	0	0	0	0	0	0	...	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
4915	0	0	0	0	0	0	0	0	0	0	...	0	
4916	0	1	0	0	0	0	0	0	0	0	...	1	
4917	0	0	0	0	0	0	0	0	0	0	...	0	
4918	0	1	0	0	0	0	1	0	0	0	...	0	
4919	0	1	0	0	0	0	0	0	0	0	...	0	

4920 rows × 133 columns

Fig. 3 Image of the training dataset

	itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	...	blackheads	scurrin
0	1	1	1	0	0	0	0	0	0	0	...	0	
1	0	0	0	1	1	1	0	0	0	0	...	0	
2	0	0	0	0	0	0	0	1	1	1	...	0	
3	1	0	0	0	0	0	0	0	0	0	...	0	
4	1	1	0	0	0	0	0	1	0	0	...	0	

5 rows × 133 columns

Fig. 4 Image of the test dataset

V. MODELS AND ALGORITHMS

To construct a disease prediction based on symptoms, we applied four machine learning algorithms: Decision Tree, Random Forest, KNN, and Naive Bayes. We can get an accurate forecast for our model using these tactics. The Prognosis of the Illness Currently, the effort is in full swing. Machine Learning is being used to diagnose and prevent disease in its infancy. As we all know, humanity has become so engrossed in the competitive environment of economic advancement that it has lost sight of its own well-being. Studies show that 40% of people ignore small symptoms, which might lead to more serious problems in the future. The project's interface is also built with Tkinter, a Python library interface. The user must first enter their name, then select symptoms from a drop-down menu; alternatively, the user must enter all symptoms, after which the system will return an exact result. Four machine learning approaches were used to create this forecast: Decision Tree, Random Forest, KNN, and Naive Bayes. When the user enters all of the symptoms and simply presses the Random Forest button, the result is computed using that method; similarly, we've utilized four ways to provide a more thorough perspective of the data, and the user must be satisfied with the anticipated conclusion.

A. Decision Tree

The most powerful and extensively used categorization and prediction tool is the decision tree. Each internal node represents an attribute test, each branch represents a test outcome, and each leaf node represents a class label in a decision tree that resembles a flowchart.

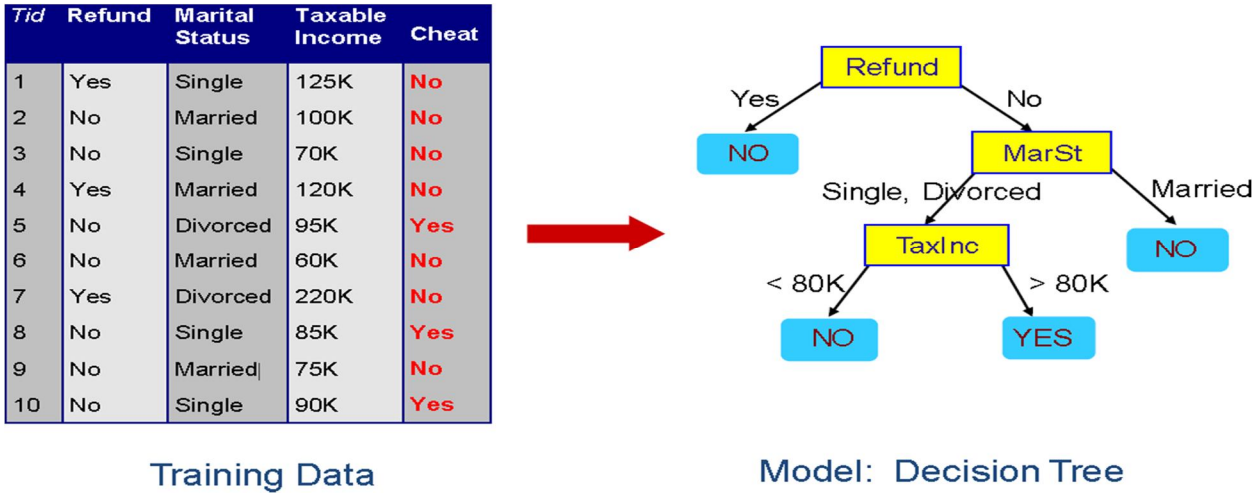


Fig 5. Decision Tree

B. Random Forest

Random Forest, a well-known machine learning algorithm, employs the supervised learning method. In machine learning, it can be utilized for both classification and regression issues. It is based on ensemble learning, which is a method for solving a complicated problem by merging numerous classifiers and improving the model's performance.

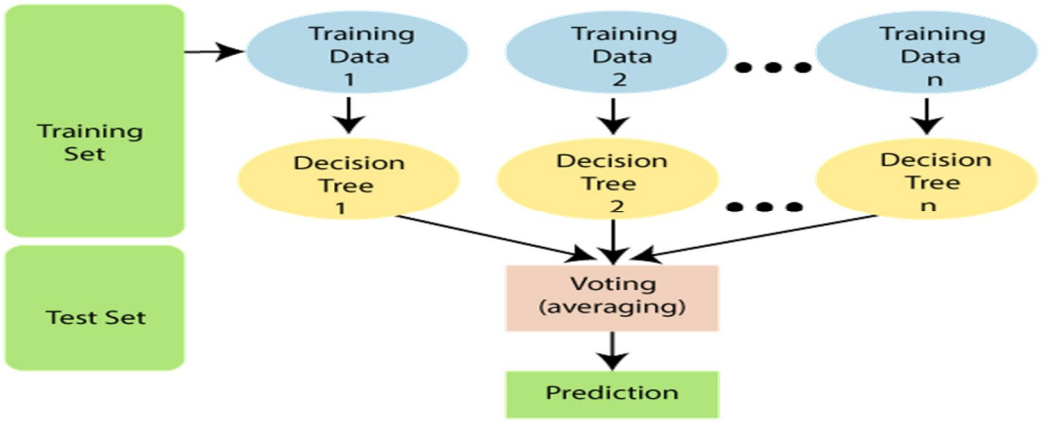


Fig. 6 Random Forest

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

C. KNN

One of the most fundamental Machine Learning algorithms is the K-Nearest Neighbour approach. It is based on the method of Supervised Learning. Because K-NN considers the new case/data and previous cases to be comparable, the new case is assigned to the category that is the most similar to the previous categories.

The K-NN method keeps track of all available data and categorizes new data points based on how similar they are to existing data. As fresh data arrives, the K-NN algorithm can quickly filter it into the appropriate suite category. Although this method can be used for both regression and Classification, classification is the most popular use.

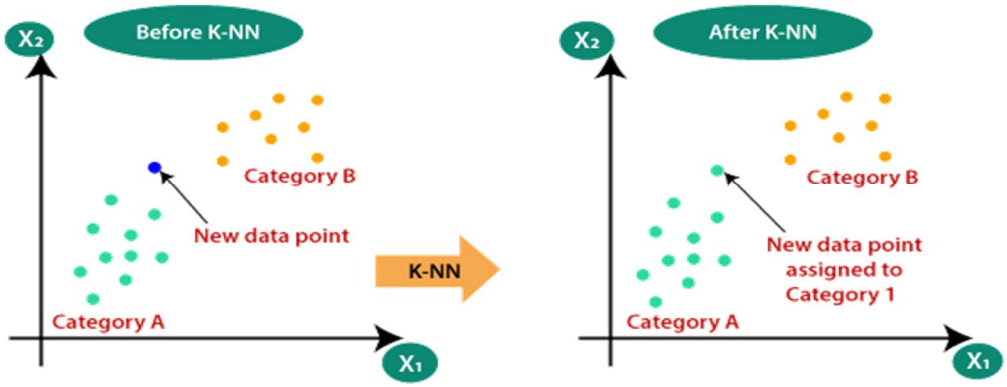


Fig. 7 K Nearest Neighbor Model

D. Naïve Bayes

The Naive Bayes algorithm is a supervised learning approach based on the Bayes theorem for classification tasks. It's commonly used in problems that require a large training dataset, such as text categorization. The Naive Bayes Classifier is a simple yet powerful classification method for quickly developing machine learning models that can make correct predictions. It's a probabilistic classifier, meaning it makes predictions based on the probability of an object. The Naive Bayes Algorithm can be used for spam filtration, sentiment analysis, and article classification, to name a few.

VI. EXPERIMENTS AND RESULTS

A. Experimentation

To conduct all of the experiments in the Jupyter notebook, we used the python3 programming language with the Tkinter interface, as well as NumPy and pandas.

B. Metrics for Assessment

We get accurate disease prediction because we supply symptoms as input to the system.

C. Disease Prediction Dataset

A CSV data file from New York-Presbyterian Hospital was provided by the University of Columbia. The training data file has 4920 rows and 133 columns, while the testing data file has 5 rows and 133 columns.

Itching, skin rash, shivering, chills, joint stiffness, and other symptoms are some of the most prevalent attributes.

D. Data Preprocessing

This step will remove any punctuation, HTML markups, hashtags, URLs, @names, and whitespace, as well as stop words, lemmatizing, and stemming text.

E. Training

The system will compare the user's symptoms to the dataset as they are entered, the dataset is made up of binary 0s and 1s, and once the model has assessed all of the user's symptoms, it will accurately forecast the disease associated with that manifestation.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

F. Results/ Outputs

To demonstrate our accuracy, we created a confusion matrix, and the patient's disease will be provided by the system

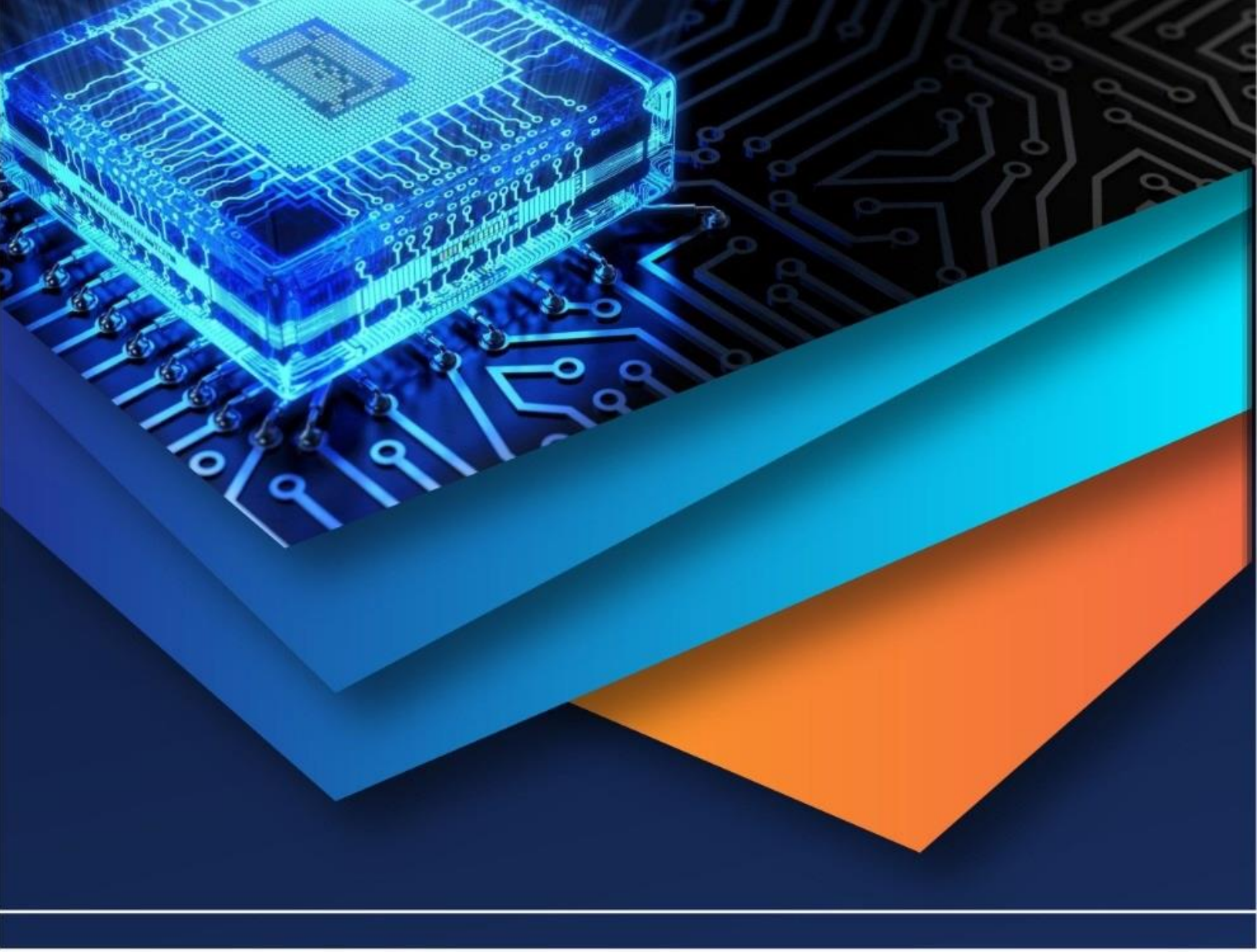
```
Decision Tree
Accuracy
0.9761904761904762
41
Confusion matrix
[[1 0 0 ... 0 0 0]
 [0 1 0 ... 0 0 0]
 [0 0 1 ... 0 0 0]
 ...
 [0 0 0 ... 1 0 0]
 [0 0 0 ... 0 1 0]
 [0 0 0 ... 0 0 1]]
Random Forest
Accuracy
0.9761904761904762
41
Confusion matrix
[[1 0 0 ... 0 0 1]
 [0 1 0 ... 0 0 0]
 [0 0 1 ... 0 0 0]
 ...
 [0 0 0 ... 1 0 0]
 [0 0 0 ... 0 1 0]
 [0 0 0 ... 0 0 1]]
Naive Bayes
Accuracy
0.9761904761904762
41
Confusion matrix
[[1 0 0 ... 0 0 0]
 [0 1 0 ... 0 0 0]
 [0 0 1 ... 0 0 0]
 ...
 [0 0 0 ... 1 0 0]
 [0 0 0 ... 0 1 0]
 [0 0 0 ... 0 0 1]]
KNN
Accuracy
0.9761904761904762
41
Confusion matrix
[[1 0 0 ... 0 0 0]
 [0 1 0 ... 0 0 0]
 [0 0 1 ... 0 0 0]
 ...
 [0 0 0 ... 1 0 0]
 [0 0 0 ... 0 1 0]
 [0 0 0 ... 0 0 1]]
```

VII. CONCLUSION

Finally, I'd like to emphasize how important this project, Disease prediction using machine learning, is to everyone's daily lives, but especially to those in the healthcare industry, who use these systems on a daily basis to predict patients' diseases based on their general information and symptoms. Because the health industry now plays such a large role in curing patients' diseases, this is often quite helpful for the health industry to inform the user, and it's also useful for the user if he or she doesn't want to travel to the hospital or other clinics, because the user can learn about the disease he or she is suffering from simply by entering the symptoms and any other relevant information, and the health industry can benefit from this system. Doctors' workload will be decreased if the healthcare industry embraces this notion, and they will be better qualified to foresee a patient's sickness. Disease prediction is a technique for foreseeing the onset of a range of common diseases that, if left untreated or ignored, can result in mortality and a slew of other problems for the patient and their family.

REFERENCES

[1] Disease Prediction and Doctor Recommendation System by www.irjet.net
[2] Disease Prediction Based on Prior Knowledge by www.hcup- us.ahrq.gov/nisoverview.jsp
[3] Kaveeshwar,S.A.,and Cornwall,J.,2014,"Thecurrentstateofdiseasemellitusin India". AMJ, 7(1), pp. 45-48.
[4] Dean, L., Mc Entyre, J., 2004, "The Genetic Landscape of Disease [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); Chapter 1, Introduction to Disease. 2004 Jul 7.
[5] MachineLearningMethods Used inDisease bywww.wikipedia.com
[6] https://www.researchgate.net/publication/325116774_disease_prediction_usin g_machine_learning_techniques
[7] https://ieeexplore.ieee.org/document/8819782/disease_prediction



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY