iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.60512

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 12 Issue IV Apr 2024- Available at www.ijraset.com

Human Disease Prediction Using Machine Learning

Prof. H.T. Gwalani[1], Unnati Katole[2], Ankit Rathod[3], Vishwanath Ingle[4], Akash Pawar[5], Dipansh Gore[6]
[1]*Assistant Professor,* [2, 3, 4, 5, 6]*Student, Computer Science and Engineering, SIPNA College of Engineering and Technology, Amravati, India*

Abstract: ***The Human Disease Prediction A machine learning system is based on predictive modeling and uses the symptoms the user enters into the system to predict the disease of the patient or user. The application has three login methods: user or patient login, doctor login, and administrator login. The tool evaluates the symptoms provided by the user or patient as input and provides the results of the disease as output based on predictive algorithms. A smart health assessment is done using a Naive Bayes classifier. The Naive Bayes classifier evaluates the probability of the disease, taking into account all subjects during the study period. Accurate interpretation of disease data facilitates patient/user disease prediction and provides users with a clear view of the disease. After the prediction, the user or patient can consult the specialist using the interactive window. It uses machine-learning algorithms and database management technology to extract new patterns from historical data. By using machine learning algorithms, prediction accuracy will be improved, and users or patients will have easy and convenient access to the application.***

Index Terms: Machine Learning, Prediction, supervised learning, accuracy, Naive Bayes

I. INTRODUCTION

Machine learning is a method that uses a series of examples to build predictive models. It is a branch of artificial intelligence that supports the idea that machines can learn from data, recognize patterns, and make decisions with minimal human intervention. Machine learning is a programming algorithm that uses sample data or previously collected data to optimize results with high precision. Machine learning algorithms have two phases: planning and research. User or patient records of signs and symptoms are used to predict disease. Machine learning provides a powerful platform to solve the problem of customer/patient experience-based disease prediction in healthcare. We use machine learning to track all signs and symptoms. Machine learning helps predictive models analyze data faster and produce results faster. With the help of technology, users and patients can make decisions about seeking care for their specific symptoms, thus improving healthcare for people's pain. The Naive Bayes classifier technique was used to analyze the data. For each subdomain of disease prediction, we also show that collecting symptom data along with classification data can aid in management, treatment, education, and studies of predicting disease from symptoms.

II. LITERATURE REVIEW

Machine learning is a field of artificial intelligence that uses data analysis and is considered a useful field that can help classify patient s or make predictions about a patient's diabetes based on information provided by people with diabetes. The main advantage of this process is that the algorithm learns from the data and uses this information to make predictions and subsequent decisions. So far, t here are many machine learning and statistical models involved in solving various problems. Although other products perform well, S VM outperforms other classifiers in terms of accuracy. According to the analysis of the article, diabetes is a dangerous disease in the world that can cause many diseases, including blindness. In this paper, they used machine learning techniques to recognize diabetes based on whether the patient has the disease, and these techniques can be done easily and adaptably. The purpose of their analysis is to create a system that can help accurately identify diabetes in patients. Here,they use only four main algorithms: decision trees, naive bayes, and the SVM algorithm, and compare their accuracy rates, which are 85%, 77%, and 77.3%, respectively.

They also used the ANN algorithm to see the response of the network, indicating whether the virus was isolated after the training process. Here they compare actual returns, F1 support scores, and accuracy across all models.

In fact, machine learning and artificial intelligence have become indispensable for many industries, including the healthcare industry. Predictive models based on machine learning algorithms can help detect diseases accurately and quickly.

Allowing doctors to provide better treatment and care to patients. Your plan to use machine learning algorithms to diagnose various di seases, such as heart disease, liver disease, and diabetes, is a good start. Using algorithms such as Random Forest and K-Nearest Neighbors (KNN) can help you achieve maximum accuracy and maximize the overall value of your prediction model.

©IJRASET: All Rights are Reserved

However, it is worth noting that machine learning models are not always perfect and may have limitations. It is important for doctors t o confirm the accuracy of the model using real data and prove the results to ensure the safety and health of the patient. Overall, the us e of machine learning and artificial intelligence in the healthcare industry has great potential to make major advances in healthcare.
The use of computer technology in the health sector has led to the production of electronic devices. This makes it difficult for medical personnel to identify symptoms and diagnose the disease early. However, supervised machine learning algorithms show significant promise in complementing existing diagnostic procedures and supporting physicians in the early detection of high-risk diseases. This literature review attempts to highlight differences in the use of observational study models in disease identification through analysis of performance measures. Naive Bayes, decision trees, and Knearest neighbors are the most popular machine learning algorithms. According to the results, support vector machines are the best method for diagnosing Parkinson's disease and kidney disease. Predicting cardiovascular disease using logistic regression. Finally, higher-level estimates of breast cancer and disease prevalence were made by random forest and central clustering based on..

III. PROBLEM ANALYSIS

The smart health prediction system focuses on optimally reducing healthcare costs. There are several functionalities that remain untouched in the health prediction system. To tackle this, research is carried out in the health prediction system. In a symptom-primarily-based ailment prediction model, the use of a machine learning method does not focus on the prediction of a selected sickness; instead, it predicts disorder based totally on the signs and symptoms given by the user. Traditional symptom-based diagnostic methods may lead to ambiguity in identifying complex medical conditions, generic medication recommendations, and inefficient appointments. Scheduling processes contribute to longer wait times, lack the ability to generate detailed and easily accessible medical records, and adherence to recommended treatments can be challenging. To solve the above challenges, we carried out the health prediction system**.**

IV. SYSTEM IMPLEMENTATION (METHODOLOGY)

If a person is actually diagnosed with some sort of disease, they need to see a doctor or physician, which is both time-consuming and expensive. It can also be difficult for the user to reach doctors and hospitals, so the disease cannot be detected. Because if the above procedure can be done with an electronic software application that saves time and resources, it could be better for the patient if the process runs smoothly. Smart health prediction is a web-based program that predicts a user's illness based on the symptoms that the user or patient can feel. Data sets for the Smart Health Prediction Framework have been compiled from various health-related websites. The consumer will be able to assess the likelihood of a disease on the basis of the symptoms presented in the web application.

A. Architecture

The aim of the project is to create a forum of web applications that can predict disease occurrence based on different symptoms and conditions. The user will select different symptoms and use the necessary information from the collected data to find the virus.

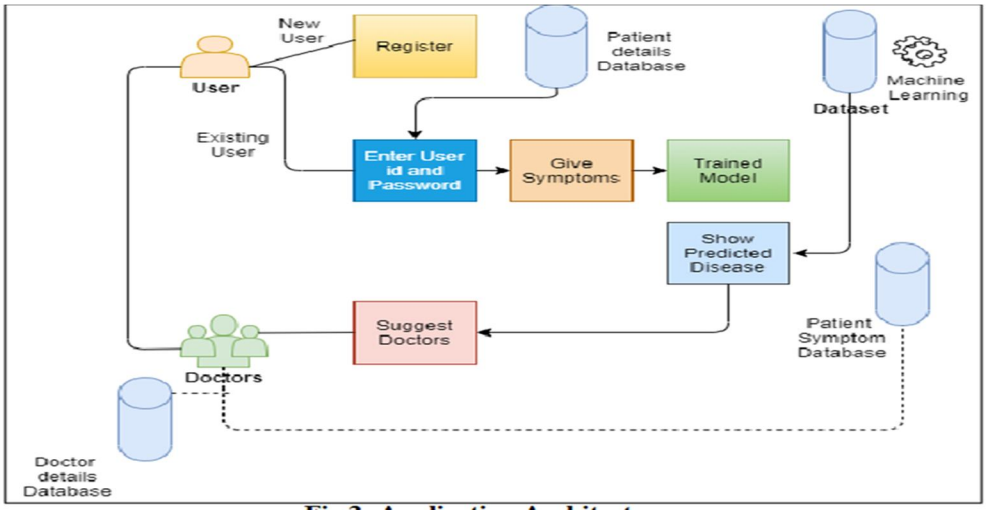


Fig.1 Application Architecture

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

B. Data Flow Diagram (DFD)

DFD has an organization consisting of patients, a manager, and a doctor. The system includes six processes, such as logging in, managing patients, managing doctors, managing appointments, generating reports, and managing information. Five stores have been created in the system: user information, patient information, appointment information, and management information. The system process is as follows:

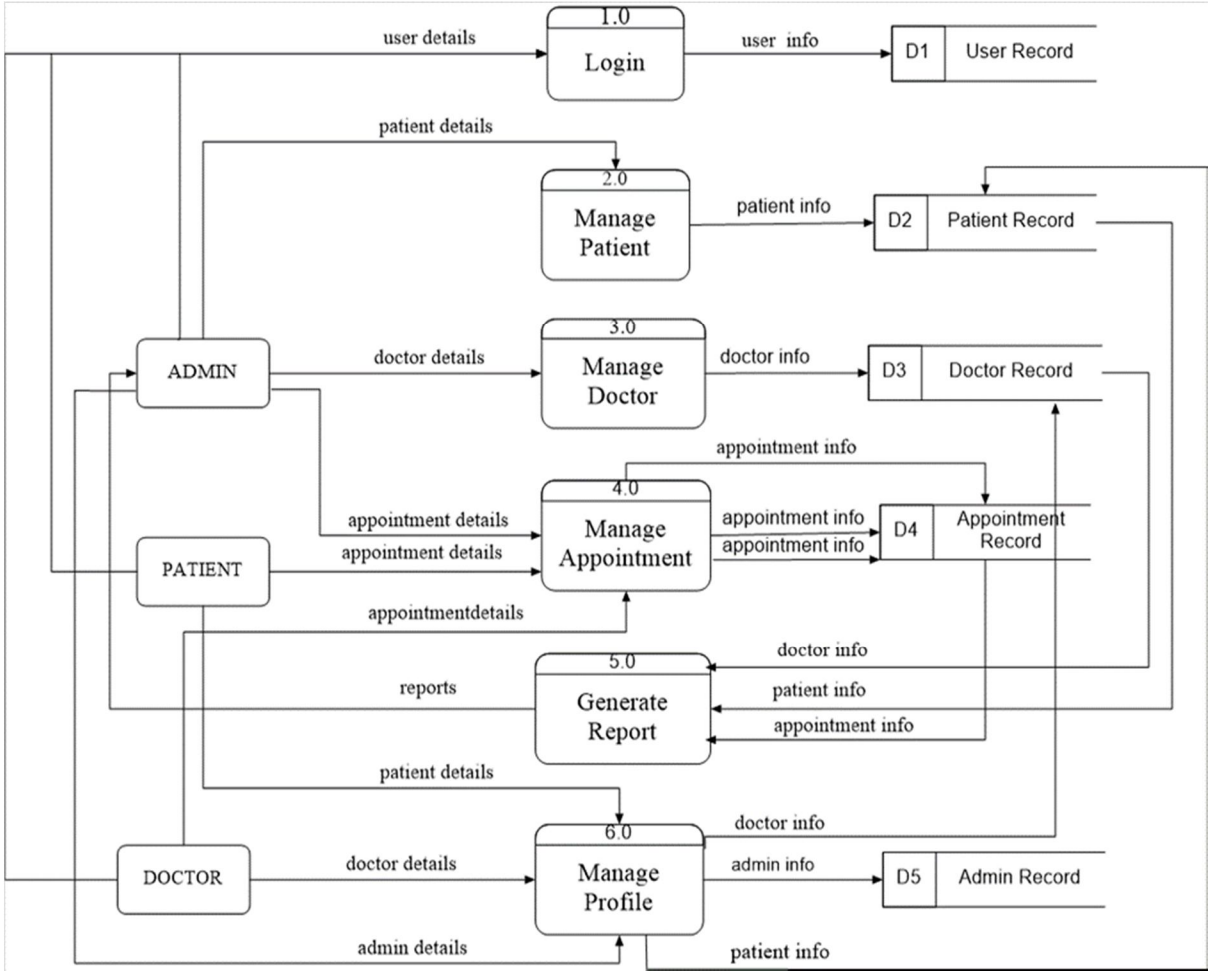


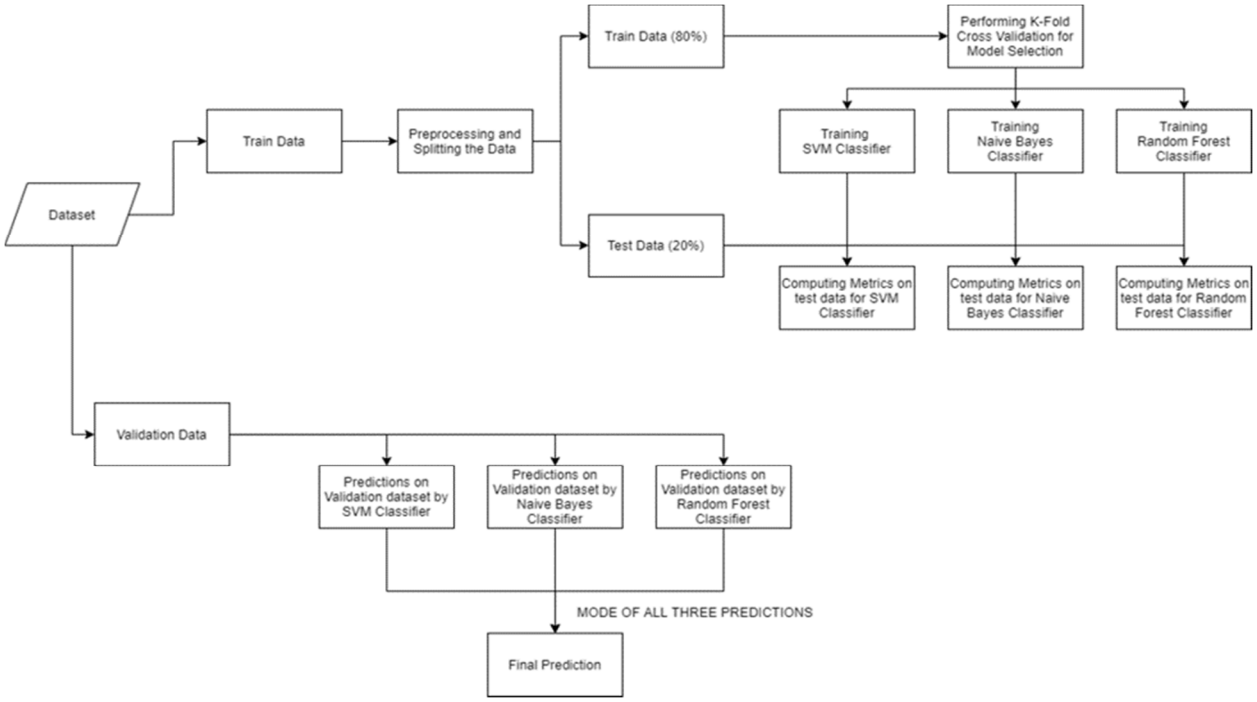
Fig 2.Data Flow diagram

The project uses powerful machine learning that can predict people's illnesses based on their symptoms. Let's see how to solve this machine learning problem:

1) *Data Collection:* Data preparation is the first step in any machine learning problem. We will use Kaggle's dataset to solve this problem. This file contains two CSV files, one for training and one for testing. There are 133 rows in the dataset, 132 of which represent symptoms, and the last row represents prognosis.
2) *Data Cleaning:* Cleaning is the most important step in machine learning. The quality of data determines the quality of machine learning models. Therefore, data always needs to be cleaned before being placed in the model for training. In our case, all columns are numbers, and the target column (i.e., prediction) is of string type and encoded using a tag encoder
3) *Design:* Once the data is collected and cleaned, it is ready and can be used to train the learning model. We will use this cleaned data to train support vector classifiers, naive Bayes classifiers, and random forest classifiers. We will use the confusion matrix to determine the quality of the model.
4) *Result:* After training our model, we will predict the disease for symptom feedback by combining our model's predictions. This ensures that all of our predictions are robust and accurate.

Finally, we will define the function to be used as a separate expression of the symptoms, predict the disease based on the symptoms using the training model, and return the prediction in JSON format.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



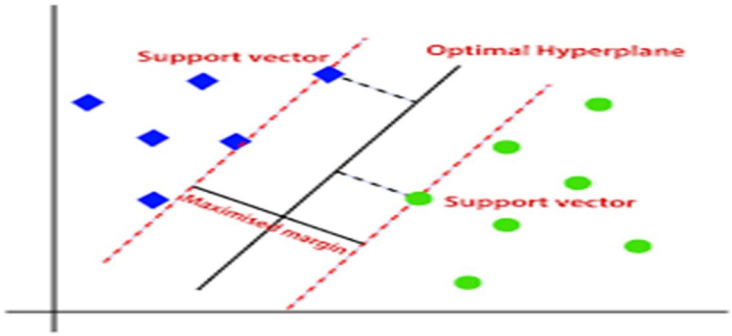
C. Algorithm

1) Random Forest

Random forests are a great way to train to test how they work early in the design process, and because of their simplicity, it's hard to create "bad" random forests. This rule is also a good choice if you need to create a model in a short time. Best of all, it's a pretty accurate representation of the weight that gives you the choice. Random forest is difficult to eliminate from a performance perspective. Most importantly, they will cover different types of variables, including binary, categorical, and numeric. Random Forest is a (mostly) fast, simple, and flexible tool, but it has its limitations. Random Forest is a combinatorial learning method for classification, regression, and other tasks that works (regressively) by building multiple decision trees during training and then displays class predictions based on the class (distribution) mode or mean. Random call forests are true for call trees that overfit the characteristics of the training sets.

2) SVM (Support Vector Machine):

Support vector machine (SVM) is a supervised learning algorithm used in machine learning to solve classification and regression tasks. SVM is particularly good at solving binary classification problems that require data points to be classified into two groups.The goal of the support vector machine algorithm is to find the best possible line or decision boundary that separates different data points in the data. When working in a high-dimensional feature space, this boundary is called a hyperplane. The idea is to obtain the best, that is, the distance between the hyperplane and the closest data for each class, thus making it easier to distinguish classes.



3) Naive Bayes Algorithm

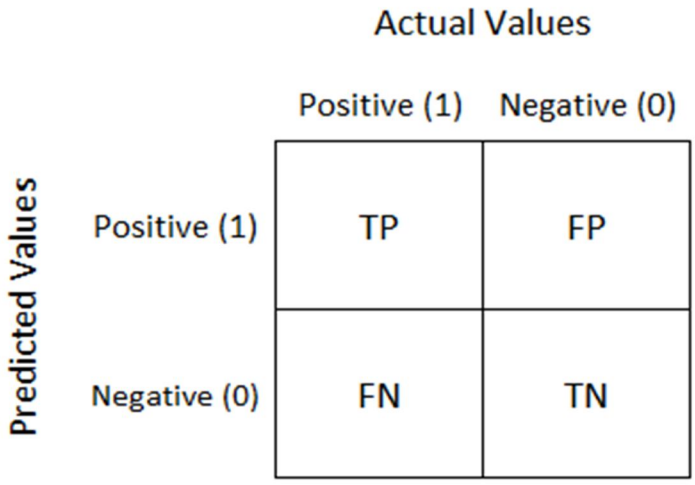
Naive Bayes is an algorithm that learns the probability that an object with certain properties belongs to a group or class. For example, when you look at fruits according to their color, shape, and smell, you will see a fruit that is orange, spherical, and has a strong orange-like smell. All these features lead to the conclusion that this fruit is orange, which is why it is called "naive." The "Bayes" section refers to author and scientist Thomas Bayes and his eponymous Bayes theorem, which is the basis of the negative Bayes algorithm. In basic terms, Bayes' theorem is expressed by the following equation:

P (A / B) = (P (B / A) * P (A)) / P (B)

D. Equation

Confusion matrix:

It is used to understand the functioning of the distribution model. It does this by using an N × N matrix, where N is the target nu mber.



Lets,understand the concept TP,TN,FP,FN.

1) True Positive (TP)

The predicted value is equal to the true value or the predicted class matches the true class.
The actual value is good and the model predicts a good value.

2) True Negative (TN)

The predicted value matches the actual value or the predicted class matches the actual class.
The actual value is negative and the model predicts the value to be negative.

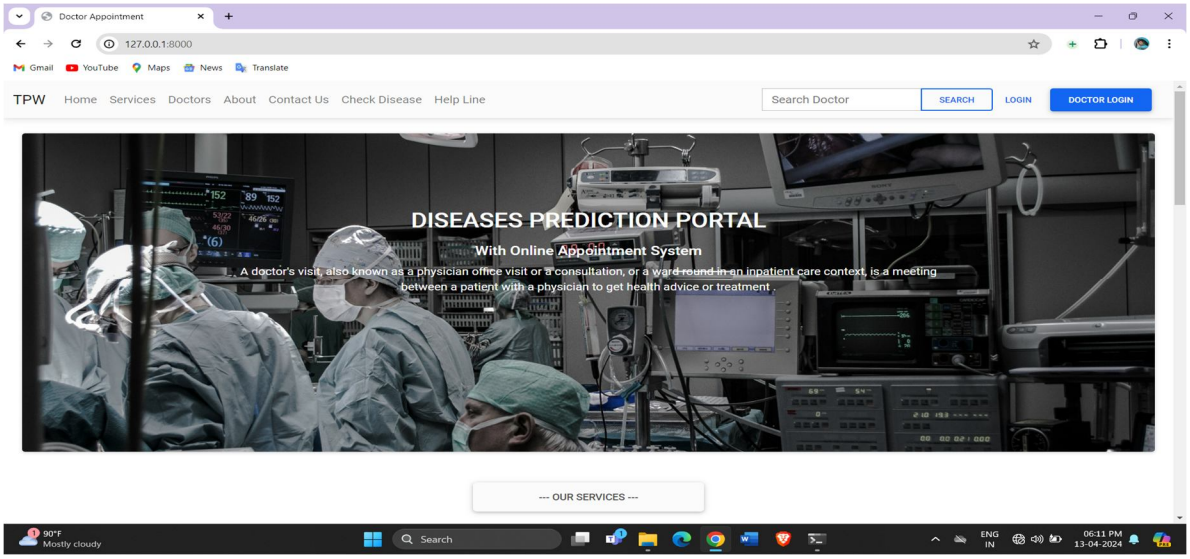
3) False Positive (FP) – Type I Error

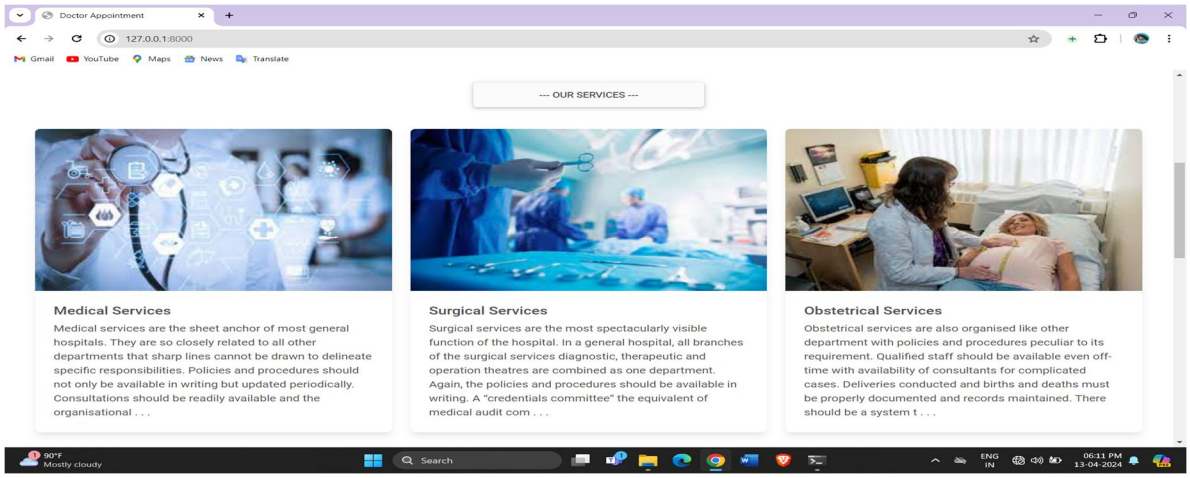
The estimated value is incorrect.
The actual value is negative but the model predicts the value well.
Also called Type I error.

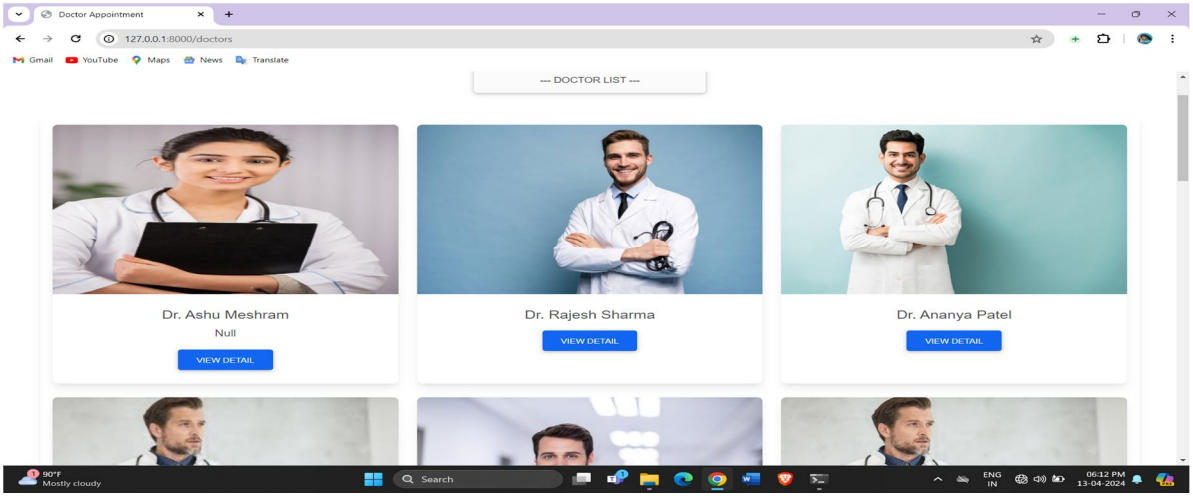
4) False Negative (FN) — Type II Error

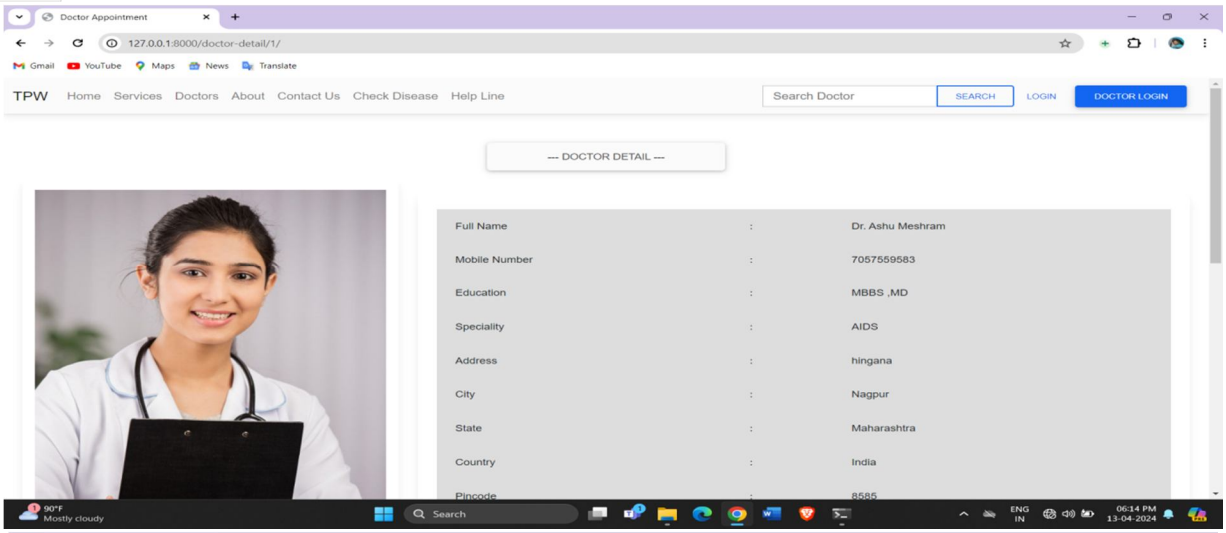
The estimated value is incorrect.
The actual value is good, but the model predicts the value is bad.
Also called type 2, error.

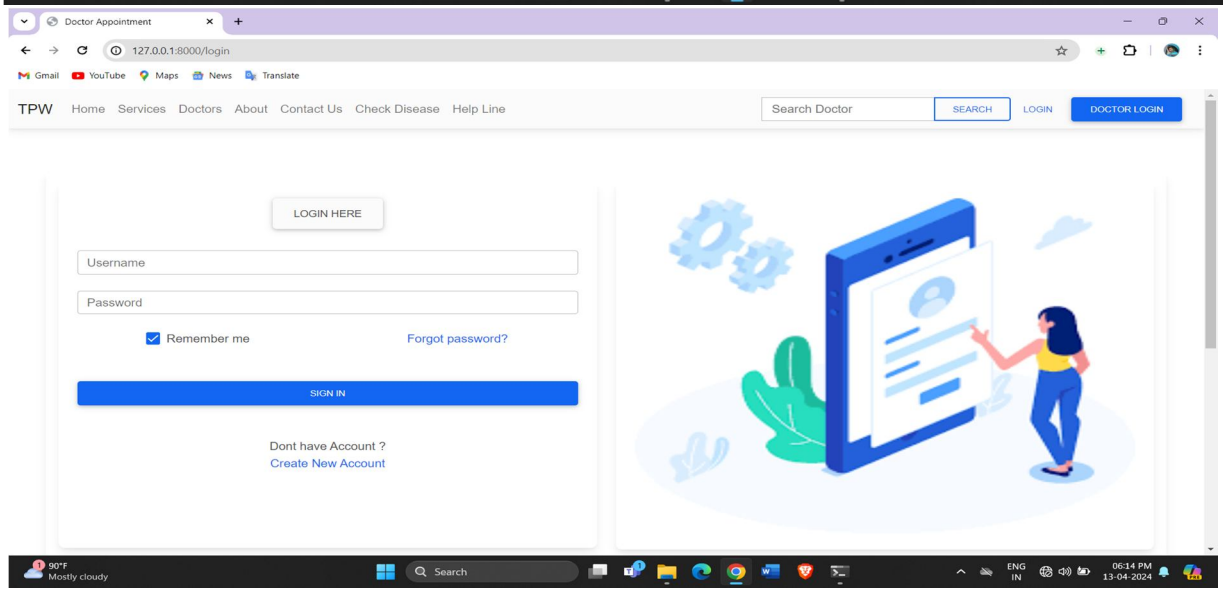
V. RESULT AND DISCUSSION

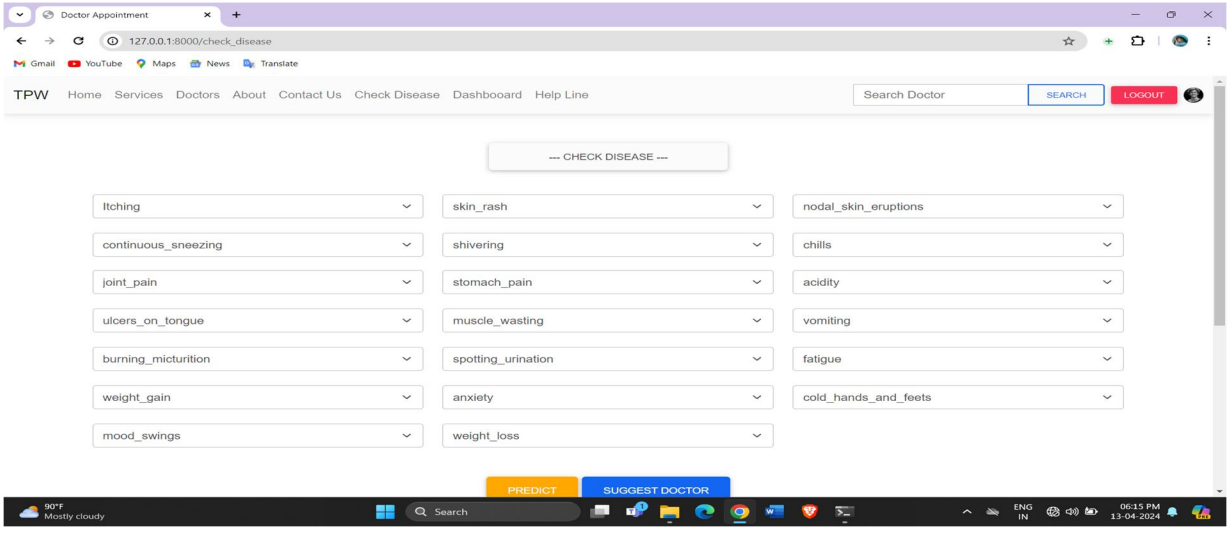




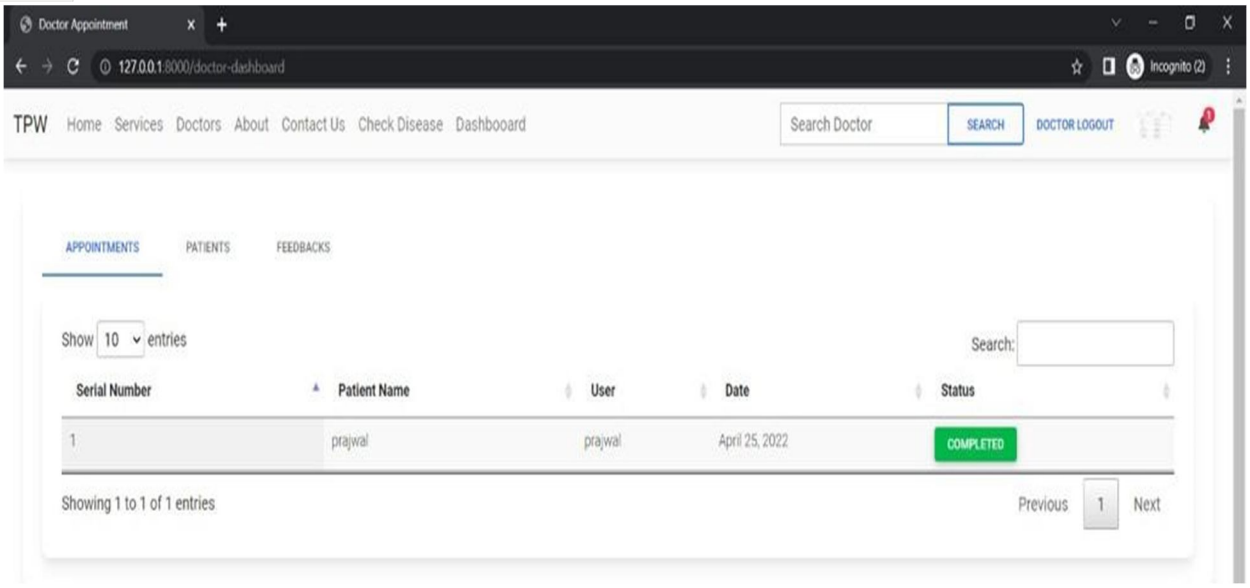


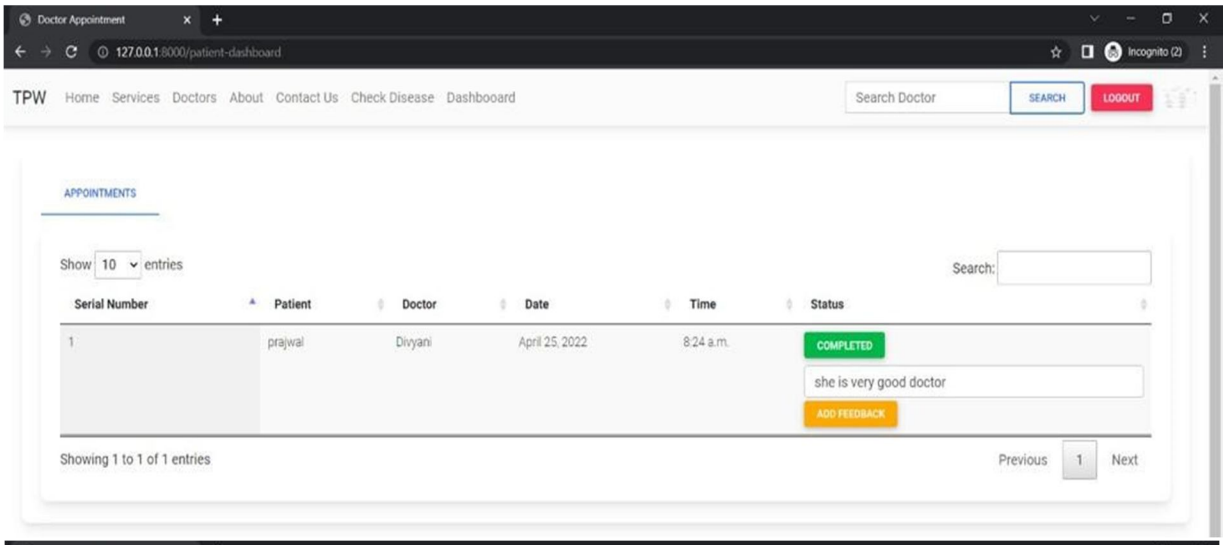


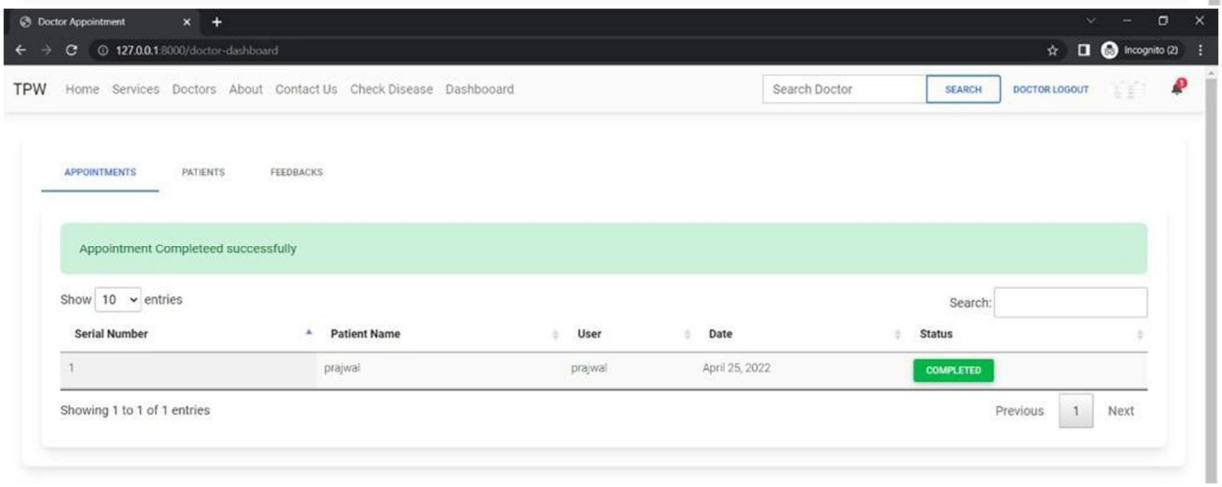




©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



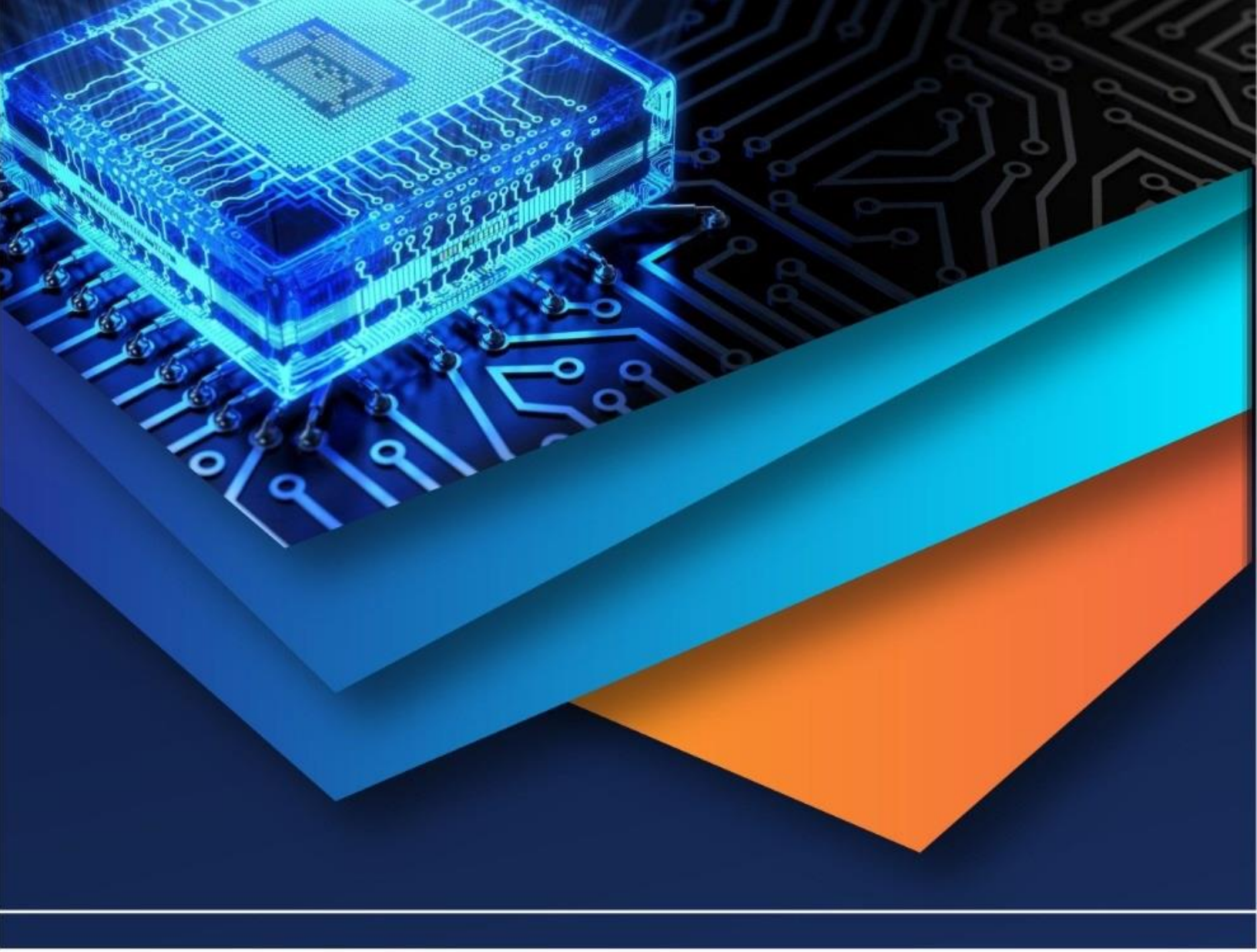




Comparing the accuracy between random forest, naive bayes and SVM algorithm. We conclude that naive bayes has the highest accuracy as compared to the other 2 algorithms. But for our project all 3 models are combined to give the best accuracy output.

REFERENCES

[1] Ali, A. 2001.Macroeconomic variables as common pervasive risk factors and the empirical content of the Arbitrage Pricing Theory. Journal of Empirical finance, 5(3): 221–240.
[2] Basu, S. 1997. The Investment Performance of Common Stocks in Relation to their Price to Earnings Ratio: A Test of the Efficient Markets Hypothesis. Journal of Finance, 33(3): 663-682
[3] Bhatti, U. and Hanif. M. 2010. Validity of Capital Assets Pricing Model.Evidence from KSE-Pakistan.European Journal of Economics, Finance and Administrative Science, 3 (20).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)