GMR Institute Of Technology, Rajam

Department of IT

DISEASE PREDICTION AND MEDICATION ADVICE USING MACHINE LEARNING ALGORITHMS

Project Supervisor
Ms. Pooja Panapana
Assistant Professor
Dept. of Information Technology

K. Rakesh Reddy (20341A1246)
J. Deepika (20341A1238)
G.Rushivardhan (20341A1226)
A.Drakshayani (20341A1203)

ABSTRACT

People today deal with a variety of diseases as a due to their lifestyle choices and the surroundings. Prediction of disease is an integral part of treatment. So, it becomes crucial to make disease predictions early on. The hardest task is making an accurate diagnosis of a disease. So, Machine learning is crucial in predicting the disease in order to solve this issue. By examining the patient's symptoms, the disease is accurately predicted in this project. The patient can enter the symptoms and the system will identify the disease they are likely to have. We use Classification Algorithms for disease prediction, such as Naive Bayes (NB), Random Forest, Logistic Regression, and KNN, with a variety of accuracy levels. Once the disease has been predicted by the system, it asks the doctor to supply the appropriate medications. The doctor will next suggest the best kind of medication to treat the disease. The patient may also seek an appointment with the doctor. Based on the doctor's availability, the patient is given a slot time, day, and date as well as the doctor's phone number. To make it easier for users to engage with the symptoms, an interactive interface is designed as the front-end. Django web application is used to implement the model.

Keywords: Naive Bayes (NB), Random Forest, Logistic Regression, K-Nearest Neighbour (KNN), Disease Prediction, Django.

INTRODUCTION

Health and pharmaceutical are going gradually important in today's rapidly changing society, as arising technology is being employed to battle virtually each known conditions. Due to the present state of the surroundings and their life choices, individuals are today exposed to a broad range of situations. According to estimations, more than 70% of Indians are exposed to common disorders including the flu, cold wave, cough, and viral infections for every two 53exmonths. Since numerous people are ignorant that general health problems could be lateral goods of reality more dangerous. The absence of early discovery of numerous conditions, including cancer, diabetes, and others, is the primary reason for adding the cause of deaths on a global scale.

Therefore, it is important to identify the illness as early as possible to maintain a strategic distance from any unfavourable losses. In order to identify the diseases early, Machine learning plays a vital role. Machine Learning is the study of a computer system in which the Machine Learning model learns from data and experience. Training and Testing are the two iterations of the machine learning algorithm. ML model helps us to develop models to collect rapidly cleaned, processed data, and deliver results quickly. The suggested system detects the illness from the symptoms data set and delivers a basic medical advice suggestion depending on the patient's symptoms by utilising Machine learning approaches like Gaussian Naive bayes Algorithm and Random Forest Classifier.

2.LITERATURE SURVEY

To create this project, we looked at five publications from various external sources. We examined the relationship between various algorithms' performance in various disease prediction scenarios.

2.1. "Disease Prediction and Doctor Recommendation System using Machine Learning Approaches". International Journal for Research in Applied Science and Engineering Technology. 9. 10.22214/ijraset.2021.36234.

This paper looks at the use of Machine Learning to develop a prediction of diseases and doctor recommendation system. Several classification approaches like Logistic Regression, Random Forest Classifier, KNN and Naïve Bayes are applied. It can give more accuracy for random forest has an accuracy of 90.2%. These methods can anticipate a person's condition based on their symptoms and then advises which sort of doctor to contact. An engaging interface is constructed as front-end and is linked to the Server. Here the model is implemented by using Django. This system might have a significant impact on how doctors treat patients in the future. But, due to the complexity and variety of diseases, there may be possible accuracy issues as well as bias in the data used to train the algorithm**.** And the major defect of this paper is it cannot predict any medication advices. It can only give disease prediction and doctor recommendation.

2.2. Gomathy, C K. (2021). "The Prediction of Disease using Machine Learning".

In this study, the use of machine learning is discussed to predict diseases from patient symptoms. It determines the probability of what disease could be present by using the supervised machine learning algorithm like Naive Bayes classifier, Decision tree, random forest, and SVM. In this system shows highest accuracy for random forest with an accuracy of 98.95%. Accurate analysis can help in early identification and improve patient care along with the development of biomedical and healthcare data. System mostly interacts with user. It also shows how the linear regression and decision tree algorithms can be used to predict specific diseases like Diabetes, Malaria, Jaundice Dengue or Tuberculosis. The benefits of applying Machine Learning model is the ability to correctly examine medical data, leading to early detection and better patient care. Additionally, it can be used to predict specific diseases with a

high degree of accuracy. However, this approach needs a lot of data for the algorithms to function well, and bias might result if insufficient varied datasets are used.[2]

2.3. Kunal Takke, Rameez Bhaijee, Avanish Singh;"Medical Disease Using Machine Learning Algorithms";2022 International Journal for Research in Applied Science & Engineering Technology (IJRASET), May 2022

The standard method of diagnosis involves a patient seeing a doctor, going through several tests, and then coming to a decision. It takes a long time to do this task. It is organized in such a way such that when the user is introduced to the chatbot system, they are offered the option of obtaining an estimate or forecast of their disease depending on the data they have supplied to the chatbot. Data is gathered from Columbia University to help with disease prediction, and Kaggle provides a source for the diseases' related symptoms. In this system user have given the maximum of five symptoms only. This system proposes an automated disease prediction to save time required for initial process of disease prediction relies on user input. There are numerous machine learning techniques used, like, K-Nearest Neighbours, Naive Bayes, Support Vector Machine and Random Forest classifier. However, when fewer symptoms are submitted, the accuracy will be reduced. [3]

2.4 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. 1211-1215, doi: 10.1109/ICCMC.2019.8819782.

This research study suggests that by utilising data mining techniques like K-Nearest Neighbour (KNN) and Convolutional Neural Network (CNN) machine learning to analyse enormous volumes of medical data to uncover trends and patterns in order to effectively forecast illnesses. CNN has an accuracy of 84.5% for general disease prediction, which is greater than KNN, as well as consuming less time and memory use than KNN. After general disease prediction, the system can indicate how likely someone is to suffer from a certain disease. The disadvantage is that it requires a large amount of medical data to be able to find patterns in order to make accurate predictions. The main limitation of using data mining methods like KNN and

CNN for general disease prediction is that it requires a large amount of medical data to be able to find patterns in order to make accurate predictions.

2.5. P. Hamsagayathri and S. Vigneshwaran, "Symptoms Based Disease Prediction Using Machine Learning Techniques," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 747-752, doi: 10.1109/ICICV50876.2021.9388603.

The study examines the use of machine learning (ML) to computer-aided diagnosis (CAD), a field of medical research. In this system using algorithms like Naïve bayes, random forest, Decision tree and SVM. System can provide highest accuracy for SVM with an accuracy of 94.60%. It talks over how Machine Learning can be used to assure impartiality in decision-making processes and to increase the accuracy and reliability of disease identification. The study also looks at several machine learning (ML) algorithms and methods for identifying conditions like diabetes and heart disease and drawing conclusions based on. This approach may also be used to analyse huge datasets from several medical sources and forecast illnesses. However, the downside of using ML in CAD is the possibility of errors based on by inaccurate data or algorithmic assumptions. The amount of data that can be used and evaluated may also be constrained, and the dataset may contain biases.

3.METHODOLOGY

3.1 Problem definition:

The Main purpose of this paper is to develop a system to predict the disease by evaluating the various patient's symptoms followed by recommending a appropriate medication advice for the disease. This system applies machine learning techniques like Naive bayes and Random Forest to study and predict the disease and prescribe medication. The flowchart Fig.3.1 shows the workflow the proposed methodology.

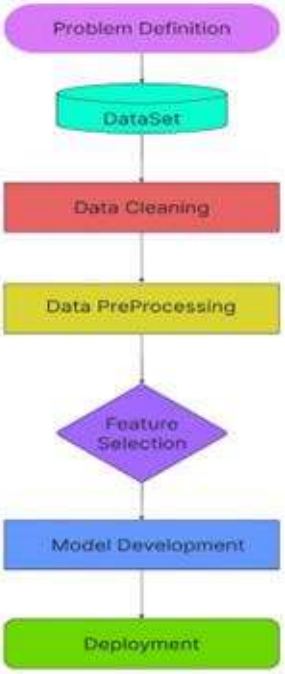


Fig 3.1: System Architecture

3.2. Data gathering:

Data preparation is an important task in any Machine Learning model. For this system, we have taken the dataset from Kaggle. In the first step, we will divide the dataset into two parts One is for Testing part and another is for Training part. This dataset consists of 133 columns out of which 132 are of symptoms and last column is for Prognosis. To eliminate the null columns from the dataset we import pandas library from the folders.

	itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	...	blackheads	sc
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	

Fig3.2.1: Dataset

lackheads	scurring	skin_peeling	silver_like_dusting	small_dents_in_nails	inflammatory_nails	blister	red_sore_around_nose	yellow_crust_ooze	prognosis
0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	Fungal infection

Fig3.2.2: Dataset

3.3. Data Preparation:

Deleting inaccurate, improperly formatted, duplicate, or incomplete data from a dataset is an important phase in Machine Learning. The accuracy of the model depends on the quality of the data [5]. So, cleaning of data should be done carefully. The given dataset is free of null values, and each feature includes 0s and 1s. Every time we perform a classification operation, we must determine whether our target column is balanced or not.

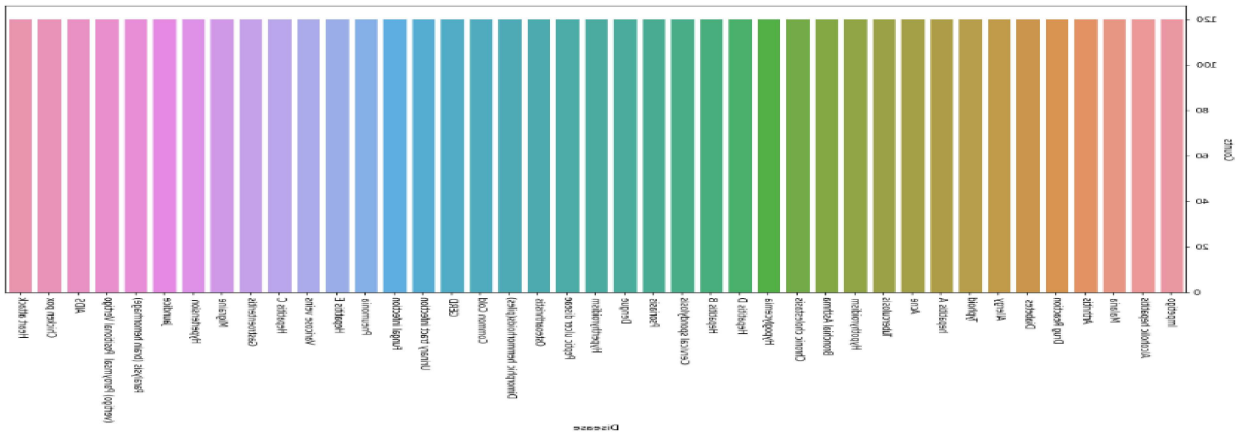


Fig3.3.1: Balanced Dataset

The figure above illustrates that the dataset is balanced, i.e., there are about 120 samples for each illness, and no more balancing is required. We see that our target column is of object datatype and such format is improper for training a machine learning model. A label encoder is used to transform the string-type prognosis target column to numerical form.

3.4. Training and testing the model:

Now we are going to split the data into training and testing data after cleaning the data by eliminating Null values and converting all these labels into numerical format. We will divide the data into an 80:20 ratio, indicating that 80% of the information will be used to train the model and 20% of the data is going to be used to evaluate the model's performance.

3.5. Model Building:

Finally, the data is ready to train a machine learning model after it has been collected and cleaned. This cleaned data will be used to train the Naive Bayes and Random Forest classifiers.

3.5.1. Random Forest:

A Random Forest Method is a supervised machine learning algorithm that is often used in Machine Learning to solve classification and regression problems. The more tress in the algorithm indicates the it predicts the result more accurately and it has good problem-solving ability [7]. Random forest is a classification algorithm which takes the average of multiple decision tress on diverse subsets of a given dataset to increase its prediction accuracy. This working is based on the concept of Ensemble learning which is a process of combining multiple classifiers to solve a complex problem and to enhance the model's performance. Ensemble

indicates the combining of different models. Ensemble generally uses two types of techniques Bagging(parallel), Boosting(sequential). Random forest approaches have three main parameters that must be changed before training begins. These parameters include number of trees, node size and number of characteristics sampled.

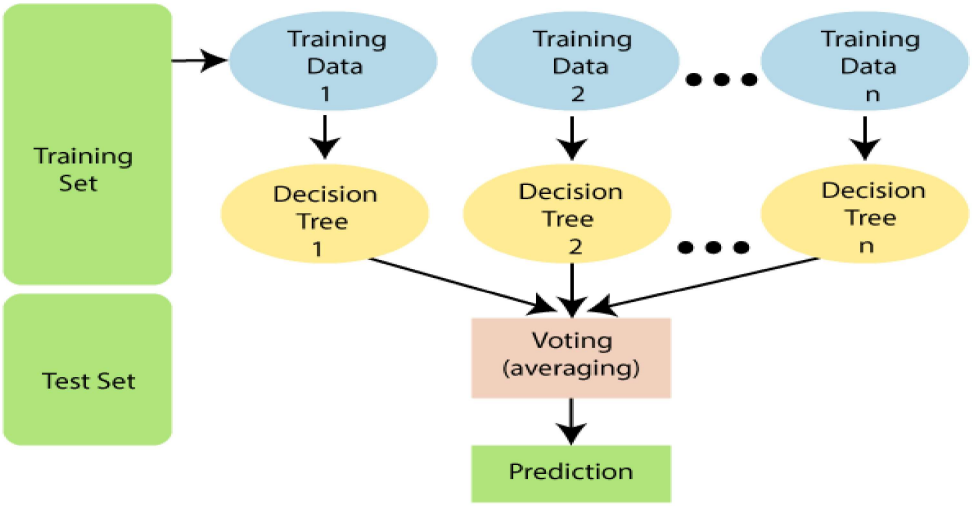


Fig3.5.1.1: Working of Random Forest

Functioning of Algorithm:

Step 1: Select some random samples from the data set or training set.

Step 2: This stage will construct a decision tree for each and every training data.

Step 3: Selection would be done by averaging the decision tree.

Step 4: Finally, choose the prediction result with the greatest number of votes as the final prediction result.

Here, from the disease dataset it will create a Bootstrap dataset i,e it will randomly picks up some symptom values from dataset and the values can be repeated. The dataset is separated into subgroups and supplied to each decision tree. Each decision tree generates a prediction result. This algorithm takes the results from multiple tress rather than depending on a single decision tree. Then it evaluates the results from each decision tree and finally it predicts the output based on the majority vote of predictions.

3.5.2. Naive Bayes:

The Naive Bayes method is a probabilistic machine learning technique which is utilized for broad variety of classification problems. It is based on Bayes Theorem. The word "Naive" represents that the algorithm incorporates features in its model that are independent from each other. Any changes in the value of one feature have no influence on the value of the algorithm's other elements i,e The value of one characteristic is considered to be independent of the value of any other feature. In practise, naive bayes classifiers are often found to be quite efficient, especially when the number of dimensions of the feature set is large.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Fig 3.5.2.1: Bayes Theorem

In order to calculate the posterior probability P(B|A), We need to create a Frequency table for each feature against the target variable. Then these frequently tables are converted into Likelihood tables. Finally, we calculate the probability for each class by using the Nave Bayesian equation. The class with the greatest posterior probability is the result of the prediction. [1]

Gaussian Naive Bayes (GNB):

Gaussian Naive Bayes (GNB) is a probabilistic classification method that applies the concept of Bayes theorem with strong independence assumptions [1]. It is a Naive Bayes variation that accepts continuous data and resembles the Gaussian normal distribution. Each parameter (also known as a feature or a predictor) is assumed to have an independent ability to predict the output variable by Gaussian Naive Bayes. It is a fast and flexible model that produces highly reliable results on large data sets. There is no need to spend a lot of time towards training. It also improves grading performance by removing unimportant specifications.

The final prediction is the combination of the predictions for all parameters that returns the probability of the dependent variable would be placed in each group. The final classification is given to the group with the highest probability.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Fig 3.5.2.2: Formula

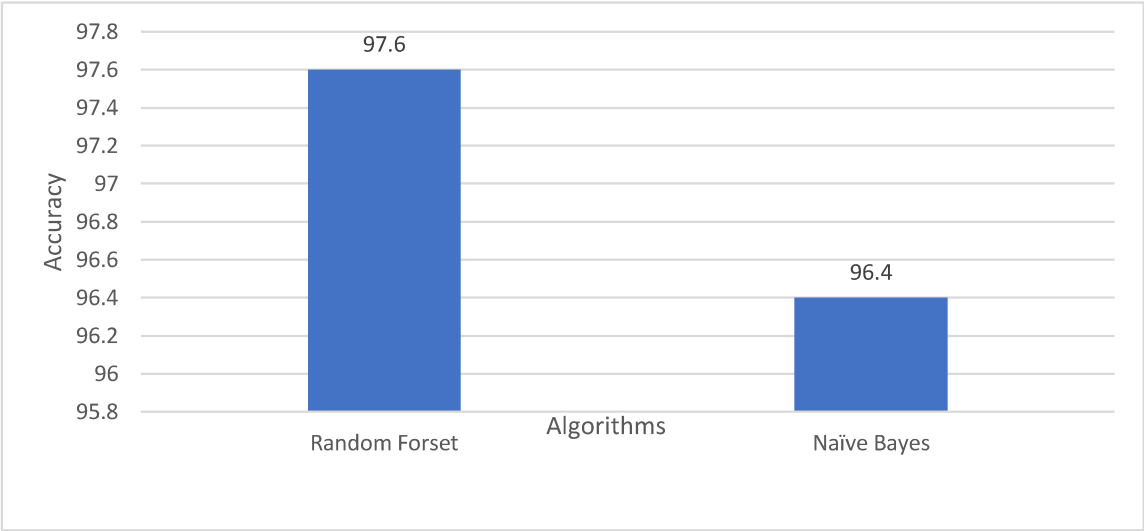
The Gaussian probability density function could be used to produce predictions by changing the parameters with the new variable input value, and the Gaussian function will offer an estimate for the probability of the new input value.

4.RESULTS AND DISCUSSION

The Disease prediction and Medication advice based on the Machine learning will predict the Disease of a person based on their Symptoms and Provide the medication for that disease. This system would use this data to train a machine learning algorithm such as Random Forest and Naive Bayes.

Table 4.1: Algorithms and accuracies for Disease Prediction

S.NO	Algorithm	Accuracy
1	Random Forest	97.6
2	Naïve Bayes	96.4



Graph 4.1: Analysis of Algorithms for Disease Prediction

1.Homepage: This is the home page of Disease Prediction and Drug Recommendation System. Here we have brief information about Different Disease Prediction and Drug Recommendation.

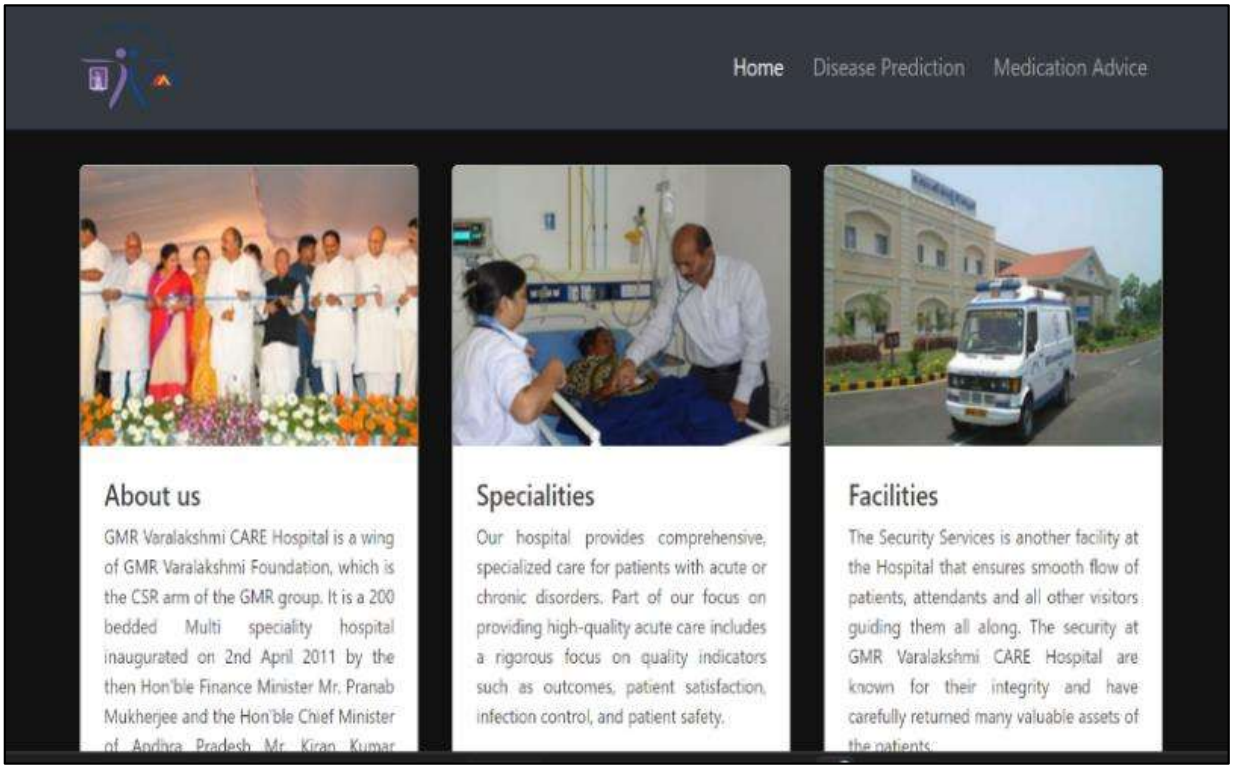


Fig 4.1: Home Page

2) Disease Prediction Form: Prediction form is a general Prediction that is based on Symptoms. User must input any symptoms and this Disease prediction and Drug

recommendation system will be able to predict the Probable underlying Disease. Prediction is based on two most famous algorithms that are 1. Random Forest Algorithms and 2. Naive Bayes.

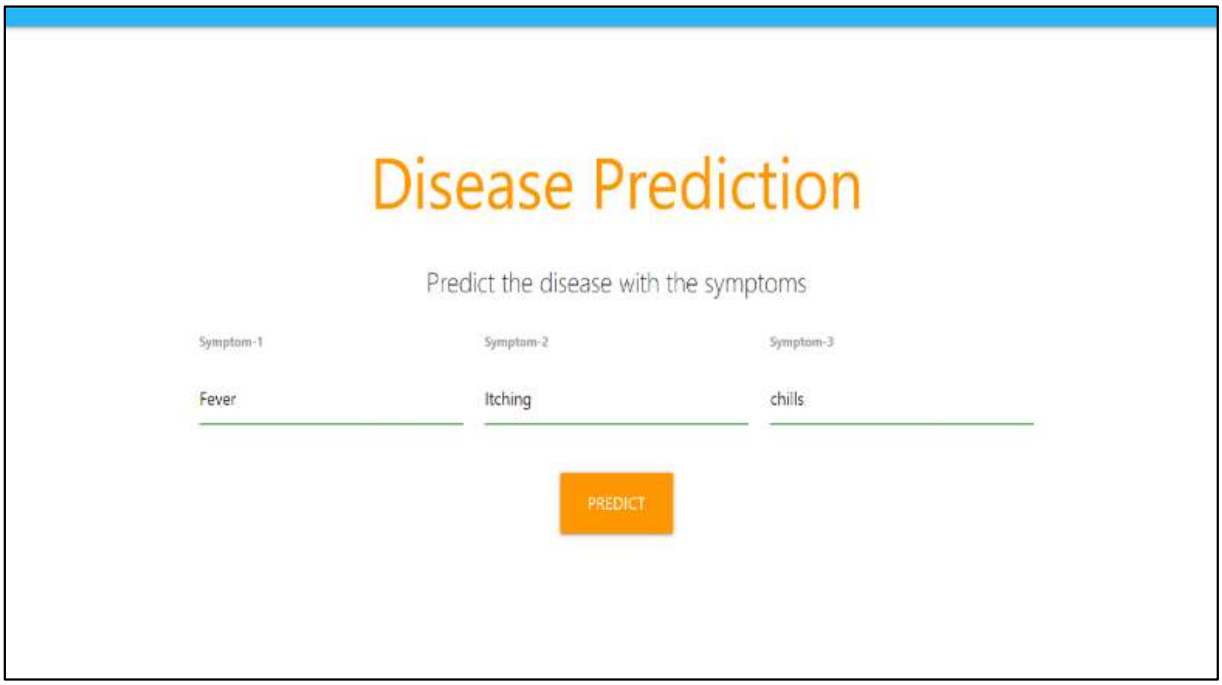


Fig 4.2: Disease prediction Page

3) Disease Prediction Output: This gives the output of the disease which the user is suffering from based on the symptoms provided by the user. Once the user submit the symptoms the Disease is anticipated.

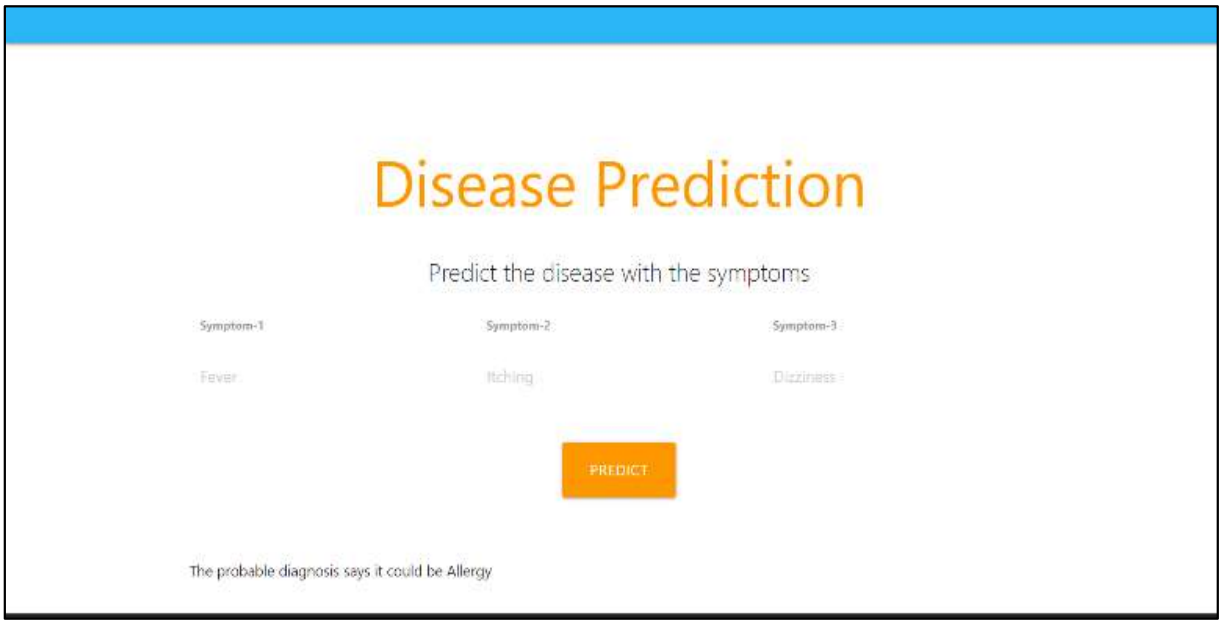


Fig 4.3: Disease prediction output

4) Drug Recommendation: This is user input form where the user Enters the Symptoms, Gender, Age and submits it. The system displays the drug for the specific disease.

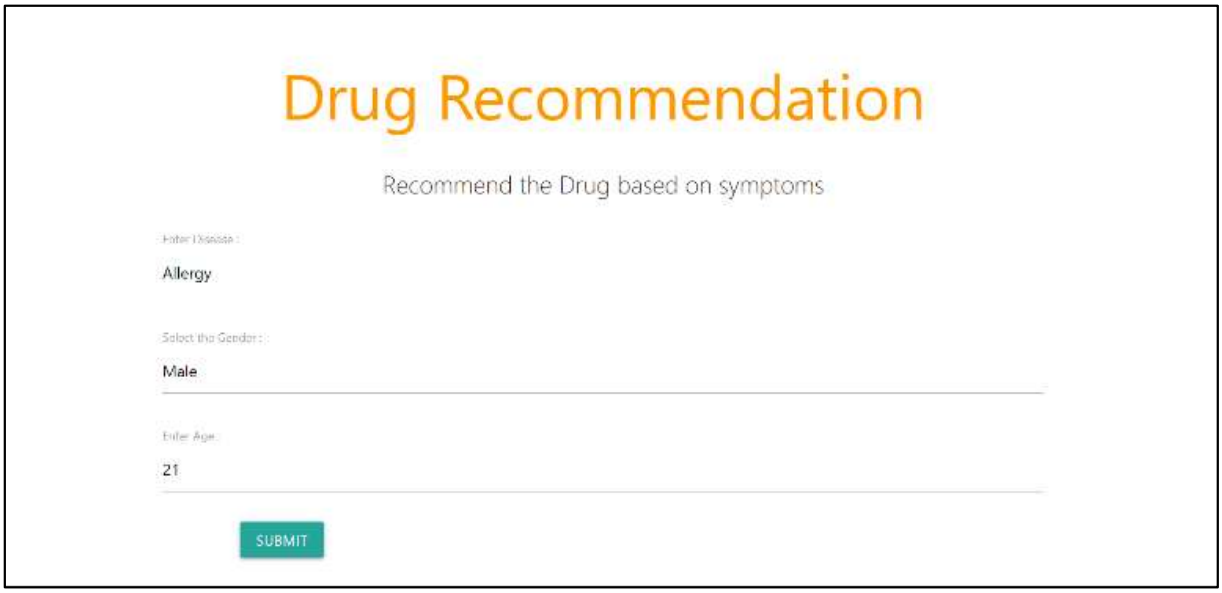


Fig 4.4: Drug Recommendation page

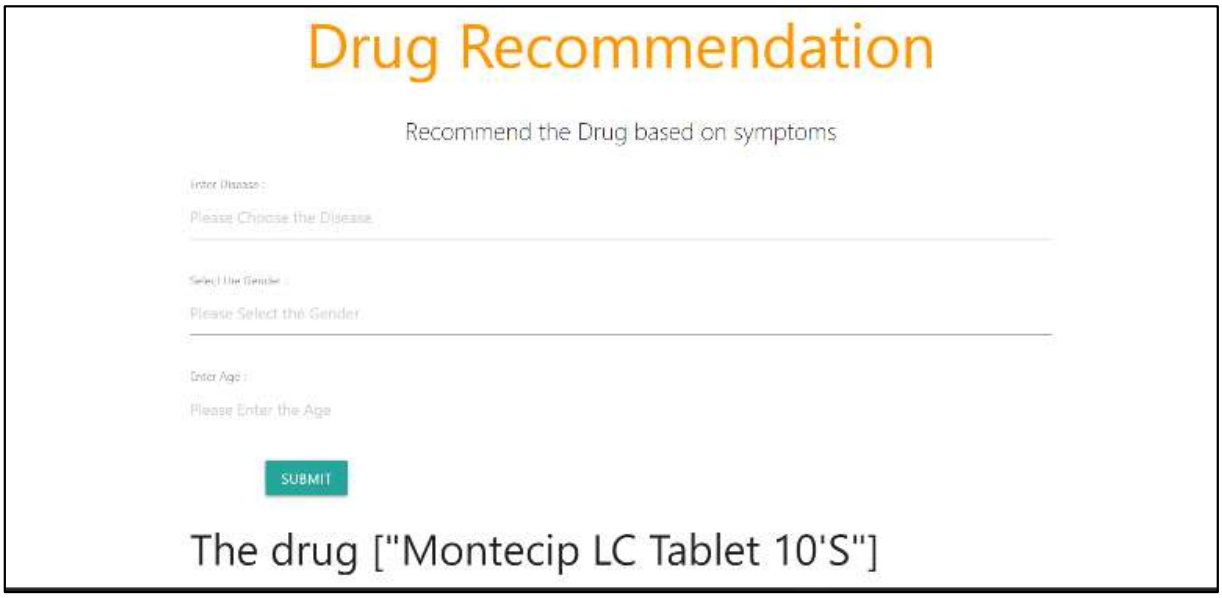


Fig 4.5: Drug Recommendation output

5. CONCLUSION AND FUTURE SCOPE:

This disease prediction system's main function is to generate disease prediction based on symptoms. This system accepts the user's symptoms as input and generates a result in the form of a disease prediction and it gives medication advice. This study presented a methodology for predicting the existence of a disease in a person by using algorithms such as Naive Bayes and Random Forest. We discovered that the Random Forest algorithm is the most used, algorithm followed by the Nave Bayes algorithm. Naive Bayes produces the best results because it is faster and offers highest accuracy of 97.6. It is widely considered that the recommended strategy can minimise illness risk by detecting them early and reducing the cost of diagnosis, treatment, and medical consultation. However, the selection of symptoms has a substantial influence on illness prediction accuracy. In the future, we can further enhance the model by including Deep Learning Algorithms and using vast datasets directly obtained from hospitals. To make the project more user-friendly, we can implement it entirely within the Android application.

6.REFERENCES

[1] "Disease Prediction and Doctor Recommendation System using Machine Learning Approaches". International Journal for Research in Applied Science and Engineering Technology. 9. 10.22214/ijraset.2021.36234.

[2] Gomathy, C K. (2021). "The Prediction of Disease using Machine Learning".

[3] Kunal Takke, Rameez Bhaijee, Avanish Singh;"Medical Disease Using Machine Learning Algorithms";2022 International Journal for Research in Applied Science & Engineering Technology (IJRASET), May 2022

[4] 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. 1211-1215, doi: 10.1109/ICCMC.2019.8819782.

[5] P. Hamsagayathri and S. Vigneshwaran, "Symptoms Based Disease Prediction Using Machine Learning Techniques," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 747-752, doi: 10.1109/ICICV50876.2021.9388603.

[6] A. Tyagi, R. Mehra and A. Saxena, "Interactive Thyroid Disease Prediction System Using Machine Learning Technique," 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), 2018, pp. 689-693, doi: 10.1109/PDGC.2018.8745910.

[7] Dhanashri Gujar, Rashmi Biyani, Tejaswini Bramhane, P Vaidya "Disease Prediction and Doctor Recommendation System," March-2018 International Research Journal of Engineering and Technology(IRJET) .

[8] Juned Hakim Shaikh, Abhishek Ganesh Takale ,"Heart Disease Prediction and Doctor Recommendation System using Machine Learning," Nov-2022 International Research Journal of Modernization in Engineering Technology and Science(IRJMETS).

[9] Keniya, Rinkal and Khakharia, Aman and Shah, Vruddhi and Gada, Vrushabh and Manjalkar, Ruchi and Thaker, Tirth and Warang, Mahesh and Mehendale, Ninad and Mehendale, Ninad, Disease Prediction from Various Symptoms Using Machine Learning (July 27, 2020).