Django Website for Disease Prediction using Machine Learning

Sayali Kulkarni[1], Isha Sawant[2], Megha Shinde[3], Vishal Sonar[4] and Vaishali Latke[5]

[1,2,3,4]B.E (Computer Engineering), Pimpri Chinchwad College Of Engineering And Research, Ravet, Pune, Maharashtra, India

[5]Asst. Professor, Department of Computer Engineering, Pimpri Chinchwad College Of Engineering And Research, Ravet, Pune, Maharashtra, India,

[1]Email: sayali.kulkarni17@gmail.com [2]Email: ishasawant2012@gmail.com
[3]Email: meghas9640@gmail.com [4]Email: vibrus77@gmail.com
[5]Email: vaishali.latke@pccoer.in

Abstract: *Decision making has become easier due to advancement in machine learning and data mining. Machine learning algorithms have helped to solve a lot of real-world problems in many fields. Taking about medical field, data mining, deep learning has made it possible to process huge data and provide more efficient health care to the patients. The types of diseases are increasing day by day and it has been necessary for people to know about the disease. The early-stage prediction of a disease based on the symptoms becomes difficult for the patient alone. The information available online may always not be correct and may lead to tension and unnecessary panic. To avoid this people should look for their health-related queries on the right place. Thus, to make it easier for people to predict the right disease, development of a machine learning based system has become important. The system collects the symptoms from the user and predicts the correct disease. This will help people to recognize the disease at an earlier stage and take the further decision based on it.*

Keywords: **Python, Machine Learning, Disease prediction by symptoms, Django, PostgreSQL**

1. Introduction

Advanced technologies are been used in almost all fields now-a-days to make life easier. Technological advancements have made a large impact in medical field as well. The healthcare domain provides improved treatments and are coming up with more efficient ways to identify and further diagnose the diseases. The increase in types of diseases and disorders, the rapid increase in patients has resulted in huge demand of advanced technologies like AI, ML, Big data. The use of such technologies would save a lot of money in healthcare domain and could enhance the patient care.

Machine learning algorithms can help a lot in detection of various diseases, the algorithms would be used to predict the best diagnosis of a crucial disease, the automation of different processes in hospitals, etc. Among these, the detection of a disease at an earlier stage by a patient can be done with the help of various machine learning algorithms. It can be made very easy and cheap to predict a disease with the initial symptoms. This earlier prediction can help common people to decide the further route of treatment and which doctor to consult. So, in this project we have implemented a system which accurately predicts a disease with the help to symptoms entered by the patient. Machine learning algorithms like Decision tree, KNN, Random Forest and Naïve Bayes have been used. The comparison of these algorithms in terms of accuracy is also

done. The trained model based on these models is used in the prediction system built using Django.

2. Literature Review

In the paper [1], comparison of various supervised machine learning algorithms has been made. The most frequently used algorithms have been studied in a detailed manner, their results and performances are learnt so that the correct algorithms can be used by other researchers for their disease prediction related work. Logistic regression, Decision tree, SVM, Naïve Bayes, Random Forest, K-Nearest Neighbour, ANN, etc are studied in deep. ROC curves are used to see the results. Tables comparing the accuracies, precision, limitations and advantages are shown.

The main aim in research paper [2] is the discussion about various decision parameters, features, attributes, etc that are held into consideration while choosing an algorithm and their importance. A comparative study of heart disease as well as cancer is done to study the pattern. The accuracies and predictions are displayed in table which makes easy to see the comparisons. The data mining techniques and machine leaning algorithms are used in predicting the disease. It was stated that neural networks predict with a greater efficiency followed by decision tree and naïve bayes. Sometimes the new medical staff may face some difficulty to know a disease in other field of their study. Taking this into consideration a system which could go hand in hand with the knowledge of doctor will help more in correct treatment.

The form of availability of data plays a major role in the most effective methods for chronic disease prediction. In the paper [3], machine learning algorithms like NB, DT, KNN are used for the prediction of disease on structured data. While on structured and text data, CNN based models are used. The use of CNN is made here on both structured as well as unstructured data. So accordingly, we understand that use of NB, KNN, DT can be made for structured data-based predictions.

The need of early prediction of an illness can help it treat early is been highlighted in the paper [7]. The comparisons of various methods for the early disease prediction have been made. The accuracies, strengths, weakness and applications are also stated which helps is proper choice of the algorithms for further.

3. Methodology

This system mainly focuses on simplifying process of disease prediction which generally takes a lot more time and needs knowledge of talented doctor. We have provided a system where we can predict the disease of a person from the symptoms that he's facing. This will be helpful to those who seek help online and at any time he can consult the doctor. The rush time at hospitals can be avoided by this. There are multiple ML algorithms which could generate result if used alone. But to ensure and get accurate result, we have used 4 different algorithms to predict the disease.

The algorithms that we have used are Decision Tree, Random Forest, Naïve Bayes, k-Nearest Neighbours. We predict the disease using above mentioned classifiers. The use of various algorithms helps in giving maximum accuracy. The accuracies are compared and according the model is trained. This trained machine learning model is used in the system. Thus, the higher accuracy makes the system efficient for patients and then they can consult a doctor accordingly. The dataset used has about 5000 records with 132 symptoms and 41 diseases. Data cleaning and data reduction is performed on dataset to avoid overfitting, the algorithms are applied and at the end the model predicts the most likely disease.

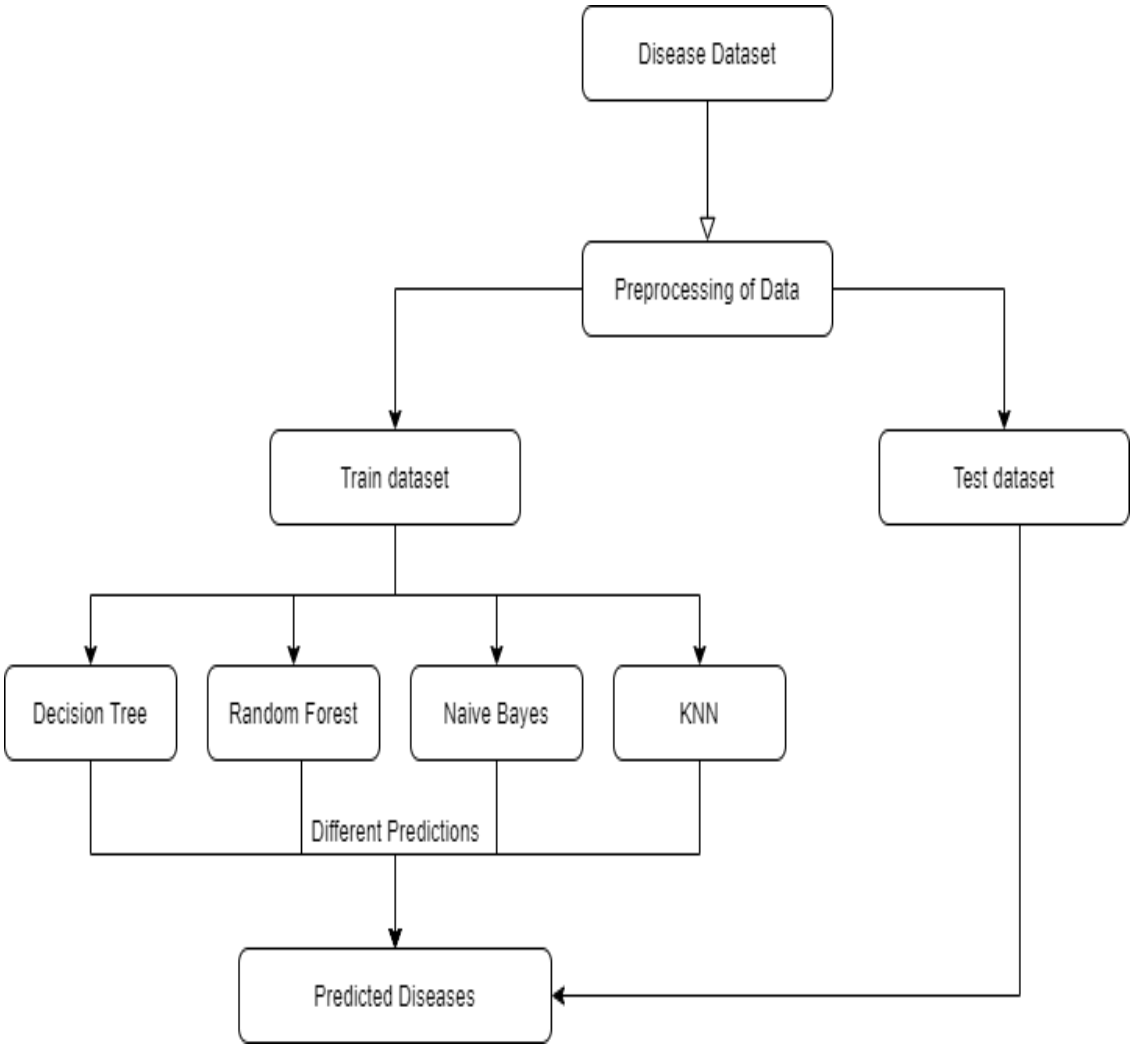

Figure 2. System Architecture

Implementation details -

The overall system is user friendly and mainly consists of two users –

- Doctors
- patients

The modules and functionalities present in patient dashboard –

- Check disease
- Consultation
- Feedback
- View and update profile

The modules and functionalities present in doctor dashboard –

- Consultation
- Feedback
- Update Profile

The user needs to register himself and after signing in he can view the functionalities provided. If the user is a patient, the functionalities provided are to check the disease by proving symptoms, consultation details and feedback option. The patient can view and update his details provided at the time of registration. The system predicts the disease using the machine learning model. Also, an option is provided to know more about that disease. A list of doctors is also provided based on the disease predicted. This will help the patient to get the necessary consultation and help at an earlier stage. This can be done by conversing with the doctor. The

patients can view the doctor profile and decide whether he wants to consult the doctor or not. The facility to view the consultation details and history has been provided. Also, ratings can be given and the system can be given a feedback as well. If the user registered is a doctor, he can fill his details which will be seen by the patients. The doctor can see the patient's disease related queries and he can answer them. The system having simplistic yet elegant user interface will work as a connecting bridge of doctor and patient. The UI is designed using HTML, CSS, JavaScript, jQuery. Django has been used for backend. PostgreSQL database and PgMyadmin is also used.

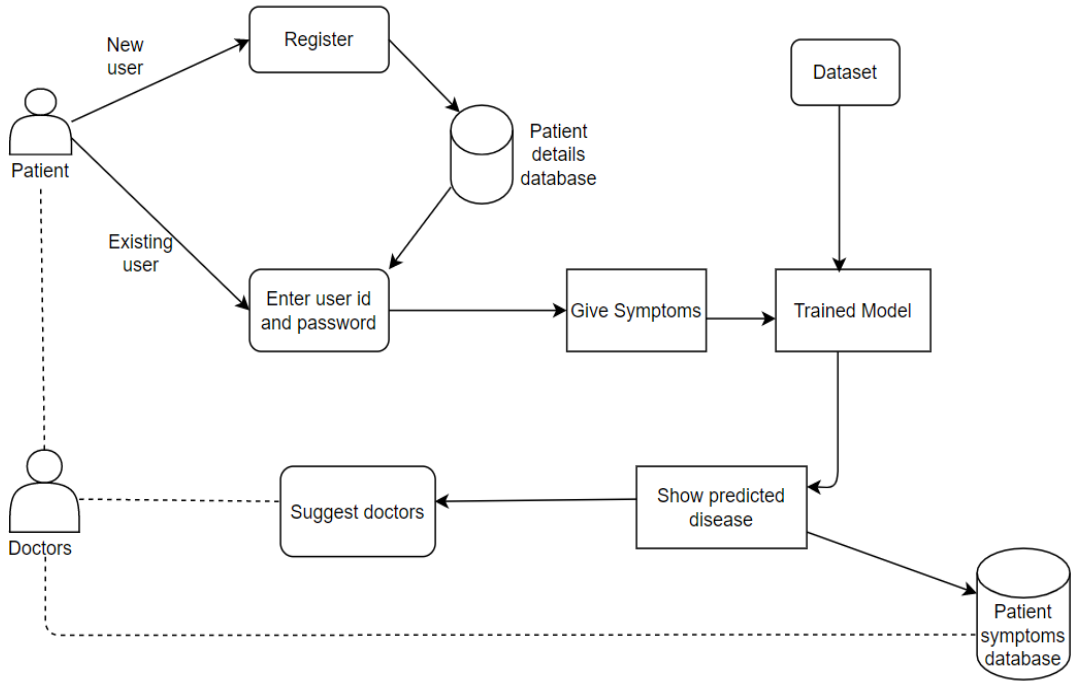

Fig. 2. ARCHITECTURE DIAGRAM

Models Used –

1. Decision Tree algorithm –

Decision tree is classified as a very effective and versatile classification technique. It is used in pattern recognition and classification for image. It is used for classification in very complex problems due to its high adaptability. It is also capable of engaging problems of higher dimensionality. It mainly consists of three parts root, nodes and leaf. Roots consists of attribute which has most effect on the outcome, leaf tests for value of certain attribute and leaf gives out the output of tree. Here gain ratio decision tree is used. It uses entropy approach due to which information gain is also maximized. Input is given here in the form of symptoms (attributes) to the nodes and the decision tree helps splitting the large data set into smaller segments. The interior nodes help predicting the output (disease) and the leaf node shows the output. Information gain is calculated using the formula –

IG(C,A) = E(C)-E(C,A)

where E(C)= entropy of frequency table using one attribute.

E(C,A) = entropy of frequency table using two attributes

C=current state

A=attribute considered

Decision tree is the first prediction method we have used in our project. It gives us an accuracy of ~95%.

Advantages-

a. Resultant classification tree is easier to understand and interpret.

b. Data preparation is easier.

c. Multiple data types such as numeric, nominal, categorical are supported.

d. Can generate robust classifiers and can be validated using statistical tests.

2.Random Forest algorithm –

Random Forest algorithm is one of the machines learning algorithm, belonging to the supervised learning technique. It solves both Classification and Regression problems in machine learning. This technique combines multiple classifiers which solves a complex problem resulting in improved performance of the model also called as ensemble learning concept. Random Forest is a technique that contains number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on only one decision tree, the random forest algorithm takes the prediction from each tree and based on the majority votes of predictions, it predicts the final output. The greater number of trees in the forest, higher is the accuracy thus, preventing the problem of overfitting. The procedure of random forest algorithm in disease prediction is –

1.The algorithm selects k symptoms randomly from total symptoms m where it builds a decision tree using these k symptoms.
2.In next step it repeats step1 multiple times to get n decision trees.
3.Pass a random variable to n decisions to predict the disease.
4.Frequent predicted disease is calculated and the most frequent one is decided as final.

Advantages-

a. Lower chance of variance and overfitting of training data compared to DT, since RF takes the average value from the outcomes of its constituent decision trees.

b. Empirically, this ensemble-based classifier performs better than its individual base classifiers, i.e., DTs.

c. Scales well for large datasets. It can provide estimates of what variables or attributes are important in the classification.

3. Naïve Bayes algorithm –

The Naive Bayesian classifier is based on a theorem called Bayes' theorem which assumes the independence between predictors. Naive Bayes model is easy to build, where there is no complicated iterative parameter estimation making it useful for very large datasets. The Naive Bayesian classifier is a simple algorithm, yet it often does well and is widely used because it often outperforms more sophisticated classification methods. It uses the Bayes' theorem with strong independence assumptions between the features to procure results.

Bayes Theorem - Bayes theorem works on conditional probability. Conditional probability is where it depends on the already happened event and assumes that something will happen based on earlier occurrence. The conditional probability gives the probability of an event using its prior knowledge.

Conditional probability: -

$P(A|B) = P(B|A).P(A)/P(B)$

Where,

P(A): The probability of a hypothesis being true also known as prior probability.

P(B): The probability of the evidence.

P(A|B): The probability of the evidence given that the hypothesis is true.

P(B|A): The probability of the hypothesis given that the evidence is true.

Advantages –

a. Simple and very useful for large datasets.

b. Can be used for both binary and multi-class classification problems.

c. It requires less amount of training data.

d. It can make probabilistic predictions and can handle both continuous and discrete data.

4. K-Nearest Neighbours algorithm –

K-Nearest Neighbour is also based on Supervised Learning technique. K-NN algorithm guess the similarity between the new case/data and cases that are already available. After that it put the new case into the category that is most similar to the available categories. All the available data is stored and classified to a new data point based on the similarity. This states that the newly appearing data can be easily classified into a suitable category by using K-NN algorithm. K-NN algorithm is mainly used for the Classification problems. K-NN is a non-parametric algorithm and does not make an assumption on underlying data. KNN does not learn from the training set immediately, so it also called a lazy learner algorithm. Instead, it stores the dataset and it performs an action on the dataset at the time of classification. KNN algorithm stores the dataset at the training phase and when it gets new data, it classifies that data into a category which is much similar to the new data.

Advantages -

a. Simple algorithm and can classify instances quickly.

b. Can handle noisy instances or instances with missing attribute values.

c. Can be used for classification and regression.

4. Results and Discussion

The use of k fold cross validation is used to compare the performance of all the algorithms. It shows that Naïve bayes and Random Forest are a bit more accurate than the other two. The results of accuracy after testing the model of 41 diseases is as given below-

Decision Tree – 0.95
Naïve Bayes - 0.95
KNN – 0.9268
Random Forest -0.95

Out of 41 diseases 38 diseases were correctly classified by K nearest neighbour and 39 were correctly predicted by Naïve bayes, Decision tree and Random Forest was seen by confusion matrix.

The percentage accuracy is as high as 96%. Hence, trained model of naïve bayes is used in the system as it is most accurate. The final results and its confidence score are displayed.

5. Conclusion

The performance is analysed using accuracy, confusion matrix, etc. Comparison of the accuracies of each algorithm like Decision tree, Random Forest, KNN, Naive Bayes has been made and it is around 96%. Thus, we have reached to a conclusion that our system provides better accuracy of disease prediction. Comparisons show that Naive Bayes given a bit more accuracy. Hence, the model is trained accordingly and use of it in this system will be helpful to those patients who are always worrying about their health and need to know what happens with their body.

The main aim to develop this system is to help such people with their health. Also, this

system can be used by small scale doctors or dispensaries to predict the disease and decrease the rush at OPDs of hospitals and reduce the workload on medical staff. The makes available the doctor's list of that particular predicted disease to get instant appointments which helps the patients and the hospitals. This ensures that the system is not affecting doctor's profession and ensures safety of patients along with increase gain of patients in the prediction system.

Acknowledgements

We thank our internal guide **Prof. Vaishali Latke** for helping us and guiding us whenever needed. The suggestions were very useful. We are also grateful to **Prof. Archana Chaugule**, Head of Computer Engineering Department, Pimpri Chinchwad College of Engineering and Research for her constant support. Special thanks to all the **teachers and friends** for providing their support.

REFERENCES

[1] *Shahadat Uddin , Arif Khan, Md Ekramul Hossain and Mohammad Ali Moni. Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making.*

[2] *Ahelam Tikotikar and Mallikarjun Kodabagi. A SURVEY ON TECHNIQUE FOR PREDICTION OF DISEASE IN MEDICAL DATA. School of Computing & IT REVA UNIVERSITY*

[3] *M. Chen, Y. Hao, K. Hwang, L. Wang, and L.Wang,"Disease prediction by machine learning over big data from healthcare communities", ," IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.*

[4] *Mr Chintan Shah,Dr. Anjali Jivani, "Comparison Of Data Mining Classification Algorithms for Breast Cancer Prediction", IEEE-31661*

[5] *Balasubramanian, Satyabhama, and Balaji Subramanian. "Symptom based disease prediction in medical system by using K-means algorithm." International Journal of Advances in Computer Science and Technology 3.*

[6] *Pingale, Kedar, et al. "Disease Prediction using Machine Learning." (2019).Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics, 2018.*

[7] *Chinmayi Chitnis and Roger Lee. Improving Health-Care Systems by Disease Prediction. 2018 International Conference on Computational Science and Computational Intelligence (CSCI)*

[8] *A.Davis, D., V.Chawla, N., Blumm, N., Christakis, N., & Barbasi, A. L. (2008). Predicting Individual Disease Risk Based On Medical History. Adam, S., & Parveen, A. (2012). Prediction System For Heart Disease Using Naive Bayes.*

[9] *P. Groves, B. Kayyali, D. Knott, and S. V. Kuiken, "The 'big data' revolution in healthcare: Accelerating value and innovation," 2016.*

[10] *Gopi Battineni, Getu Gamo Sagaro, Nalini Chinatalapudi and Francesco Amenta. Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis. J. Pers. Med. 2020, 10, 21; doi:10.3390/jpm10020021*