DISEASE PREDICTION SYSTEM USING SYMPTOMS

Er. Harjeet Singh[1], Mr. Ankit Mehta[2],

*[*1,2]Faculty, Department of Computer Science Lloyd Institute of Engineering & Technology Gr. Noida*

---***---

Abstract - Technology has revolutionized the health domain largely. This project aims to design a diagnostic model for various diseases relying on their symptoms. This System has used data mining techniques such as classification in order to achieve such a model. Datasets consisting of voluminous data about patient diseases are gathered, refined, classified, and used for training the intelligent agent. Here, the Naive Bayes Algorithm is used for classification purposes. Naïve Bayes Classifier calculates the probabilities of the disease. Based on the result, the patient can contact the doctor accordingly for further treatment. It is an exemplar where technology and health knowledge are sewn into a thread perfectly with a desire to achieve “prediction is better than cure”

***Key Words*: Machine learning, Naïve Bays, Database, Dataset.**

1. INTRODUCTION

The use of the internet has been stimulating curiosity among people and, if it of any kind is, they are trying to find a solution to their problems through the internet only. It is a matter of fact that people have much easier access to the Internet than hospitals and doctors. So, with the help of this system, a user can consult a doctor by sitting at home itself. There will not be any fuss about visiting a clinic or hospital and making your health condition worse. This Disease Prediction system is a web-based application that predicts the most probable disease of the user in accordance with the given symptoms with the help of the data sets collected from different health-related sites. It often happens that someone nearer or dearer to you may need a doctor’s help immediately for some serious reasons but the doctor is not available for consultation for some prior commitments or other obvious reasons. That is when the role of this automated program comes into play. This Disease Prediction system can be used for urgent guidance on their illness according to the details and symptoms they will feed to the web-based application. Here, we use some intelligent data processing techniques to get the most accurate disease that would be related to the patient’s details. And then based on the results, the patient can contact the respective disease specialist for any further treatments. This system can be used for a free consultation regarding any illness.

1.1 NEED OF STUDY

The need of this document is to present a detailed description of a Web application Based Disease Prediction System an interactive web application for users. It will explain the purpose and features of the system, the interfaces of the system, what the system will do, theconstraints under which it must operate, and how the system interface will react to external stimuli:

1.1 Product Scope

A web application-based Disease Prediction System, which predicts diseases that the user may be suffering from, based on the symptoms that the user provides. Basically, the portal is designed for three entities. They are three Users, Doctor, Admin; The User can fill in the Patient details along with the symptoms. The symptoms provided are used by a probability-based algorithm to Display a list of probable Diseases according to their relevance to Symptoms.

It also displays a list of doctors from the user’s provided city who can cure at least one of the predicted diseases. More efficient and robust Data Mining and Machine Learning algorithms that provide well structure and comparatively larger Datasets that are well known for their accurate prediction can replace the current algorithm. We can also increase records in our disease and symptom database by asking doctors and admin to share with us valuable information. We can also add features such as registering for doctor’s appointments from the portal itself.

- The disease prediction system has three users as doctor,patient, and admin.
- Each user of the system are authenticated by the system.
- There is role-based access to the system..
- The system allows the patient to give symptoms and accordingto those symptoms, the system will predict a disease.
- The system allows the patient to give symptoms and accordingto those symptoms, the system will predict a disease.
- The system allows online consultation for patients.

1.2 Related Work

Here is a list of symptoms are required medication, for promoting any program, Potential disease involves either accurate or false. Thisreport Explain the nature of some of the diagnoses and related symptoms such as a disease but it may not give complete information about the site Symptoms/diagnosis are not related to the patient or family Record or other factor. The Iliad is an expert disease diagnosis system used to describe a relationship to finding Disease. This system uses the Naïve classification to calculate Diagnosis DX plain is a medical decision Support System It generates a ranking list of features most likely; the disease is in the lowest place storage. To use Information, prevalence, and importance of each disease, the system identifies the difference between a common disease and a rare disease. This system also serves as a physician reference with a searchable database of Diseases and clinical manifestations. Clinical decision support systems are used to identify diagonals the patent was recorded. It has three broad categories.

- Improve patient safety.
- Improve the quality of care.
- Improve the efficiency of health care delivery.

Patient safety in terms of minimizing and correcting errors Drug. The second category describes clinical improvement Documentation and patient satisfaction. Describes the third category Reduce the price and duplicate list, minimize the negativity of the event

Use Novell to separate features of all datasets here Classifiers based on bias discrimination function. The hybrid algorithm was used to extract unique features from the throat Biological dataset. Machine learning algorithms are used in the Training set. The main goal is to find a relationship among the features that can be used in decision-making. This is a method of preventing many problems with medical data such as missing Prices, Rare Information, and Temporary Data. Machine learning the algorithm is well suited for this type of data. There are two types of use:

1. To find the relationship between the features.
2. Test prediction for future disorders.

2. REVIEW OF LITERATURE

Al-AIdaroos KM [1] Research Reviews and literature for the best and simple medical clinical mining technology. Was done by. Forever. The authors compared five other taxonomies for Nivea. The basis is Logistics Regression LR), K Star (K *), Decision Tree (DT), Neural Network (NN) and Simple Rule-Based Algorithm (Zero R). He uses15 real-world medical issues from the UCI Machine-Learning repository [2]. Those selected for evaluation. 8th bale not found 8 out of 15 data sets are considered to be a better estimation technique than others to overcome and as shown by other algorithms in figs. 1. The result shows that the tree works well and is sometimes as accurate as the decisive tree in the Bayesian taxonomy, others Estimation methods such as KNN, neural network, clustering based classification do not work properly [3]. In addition, he suggested it Decision trees accuracy and Bayesian classification further improved after applying genetic algorithms to reduce actual data Quantity to obtain an appropriate subset of symptoms to assess cardiovascular disease. You, which are compared to trained values, and then diagnoses heart disease [4]. The result of masathe H.D [5] etc. showed 99% evaluation accuracy, which has been used in data mining techniques Health sector to evaluate samples from dataset. There are many documents that diagnose heart disease. Based on a predetermined tree Systems have been used to diagnose heart disease. Showman Maui [6] and. Al. Alternative decision evaluation tree performance, J4.8 decision tree pattern and bagging algorithm in the diagnosis [7]. of heart disease patients. They are tenderness, uniqueness and Found to be better than accuracy and other algorithms.

Medical Problems	NB	LR	K*	DT	NN	Zero R
Breast cancer wise	97.3	92.98	95.72	94.57	95.57	65.52
Breast cancer	72.7	67.77	73.73	74.28	66.95	70.3
Dermatology	97.43	96.89	94.51	94.1	96.45	30.6
Echocardiogram	95.77	94.59	89.38	96.41	93.64	67.86
Liver Disorders	54.89	68.72	66.82	65.84	68.73	57.98
Pima Diabetes	75.75	77.47	70.19	74.49	74.75	65.11
Haeberman	75.36	74.41	73.73	72.16	70.32	73.53
Heart-c	83.34	83.7	75.18	77.13	80.99	54.45
Heart-statlog	84.85	84.4	73.89	75.59	81.78	55.56
Heart-b	83.95	84.23	77.83	80.22	80.07	63.95
Hepatitis	83.81	83.89	80.17	79.22	80.78	79.38
Lung cancer	53.25	47.25	41.67	40.83	44.08	40
Lymphpgraphy	94.97	78.45	83.18	78.21	81.81	54.76
Postooerative patient	68.11	61.11	61.67	69.78	58.54	71.11
Primary tumor	49.71	14.62	38.02	41.39	40.38	24.78
wins	8\15.	5\15	0\15	2\15	1\15	1\15

The same researchers found 83.3% accuracy in adding k-mean clustering Diagnosis vembandasamy K. et. Al. The innocent bias algorithm was used to classify and identify medical data 86.4-198% accuracy with minimum time [8]

Methodology

The Dataset : for getting some dataset and training our model, so fore that we have made some surveys in medical field explored some data on internet and made arrow dataset by combining all of that so now we have a dataset. This CSV file contains 5000 rows and 133 columns, 132 columns properties. And last column for the disease class (40 unique disease classes). Some rows of disease with their corresponding symptoms in the dataset

	Disease	Symptoms
1	Malaria	{chills, vomiting, high_fever, sweating, heada...
2	Allergy	{continuous_sneezing, shivering, chills, water...
3	Fungal infection	{skin_rash, nodal_skin_eruptions, dishromic_...
4	Gastroenteritis	{vomiting,sunken_eyes, dehydration, diarrhoea
5	arthritis	{muscle_weekness, stiff_neck, swelling_joints,...
6	Typhoid	{chills,vomiting, fatirgue, high_fever, headac...
7	Hypertension	{muscle_weakness, stiff_neck, swelling_joint,....

Data Pre-processing -After collecting that data, as that data is raw data we have to make it suitable for training our machine learning model. By using some python libraries like NumPy, and pandas, we have made that data suitable for machine learning models.

Now, our data is ready to use with machine learning algorithms to predict some output. As our problem come under unsupervised machine learning technique, we have used Naive Bayes algorithm.

Model building-After applying these algorithms, we have to select which is most fitted with our dataset and which gives us more accuracy. So, we have used a confusion matrix for that and mapped out the accuracy of each model. And, we have found that all are giving the same 100% accuracy, so we have selected Random Forest Classifier for building our model

Naïve Bayes - Machine Learning Field Monitoring and its various model decisions there are a wide range of algorithms to predict various diseases with predictive characteristics in tree, neo base and random forests. There Three different models of naïve Bayes, namely Gaussian, multinational and Bernoulli naïve Bayes. Each model has its own accuracy Application and data fitting to assess disease outcome are almost identical in all three models. Since the Gaussian naïve Bayes Relatively easy to understand and much simpler than the other two, this project work was done using it.

Code: from sklearn.naive_bayes import GaussianNB

```
gnb = GaussianNB()
```

Naïve Bayes classifier depends on Bayes Theorem.

Bayes theorem:

$$P(Y/X_1,X_2,\ldots\ldots,X_n)= P(Y)\,\frac{P(X_1,X_2,\ldots\ldots,X_n)}{P(X_1,X_n)}$$

Where, Y is the class Variable

$X_1,X_2,\ldots,X_n$ are the dependent features.

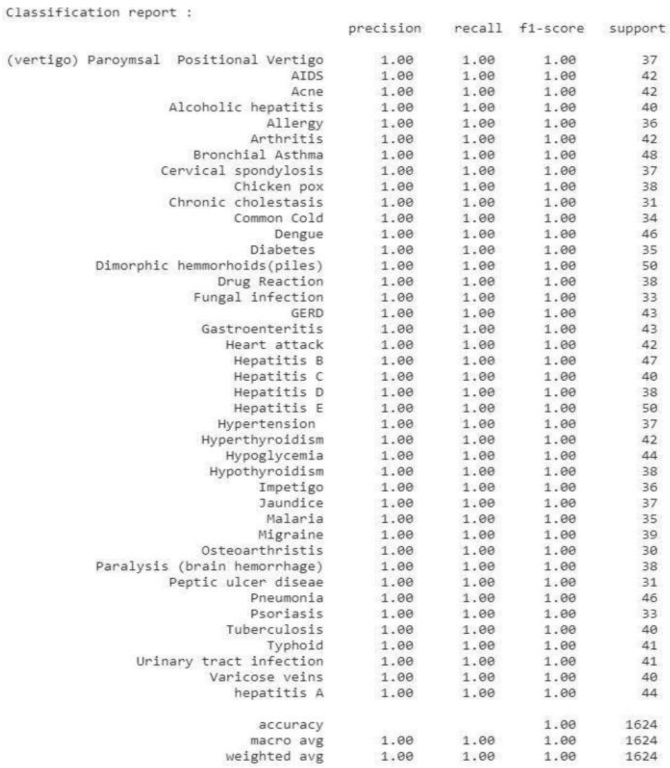
Confusion matrix-

A confusion matrix is basically a table that is used to describe the performance of a classification model. First test the data that has thecorrect values. From the confusion matrix table, it is clear that the Naïve Bayes algorithm is predicting everything the diseases correctly in the test set.

This confusion matrix showing the "actual class" versus 'predicted class" ratios:

```
[[18,    0,    0,   .........,    0,    0,   0],
 [ 0,   30,    0,   .........,    0,    0,   0],
 [ 0,    0,   24, .........,      0,    0,   0],
 ......,
 [ 0,    0,    0, ...........,    0,    0,   0],
 [ 0,    0,    0, ...........,    0,   22,   0],
 [ 0,    0,    0, ..........,     0,    0,  34]]
```

Classification report- Classification report visualizes the precision recall and F1 score of a model.



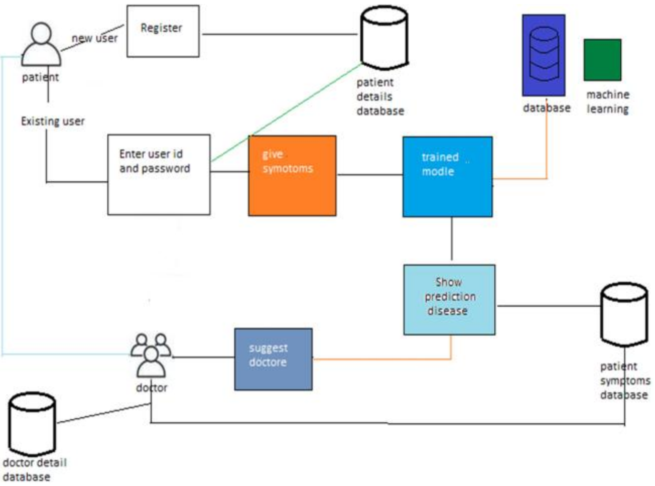
```
Classification report :
                                          precision    recall  f1-score   support

(vertigo) Paroymsal  Positional Vertigo       1.00      1.00      1.00        37
                                   AIDS       1.00      1.00      1.00        42
                                   Acne       1.00      1.00      1.00        42
                    Alcoholic hepatitis       1.00      1.00      1.00        40
                                Allergy       1.00      1.00      1.00        36
                              Arthritis       1.00      1.00      1.00        42
                       Bronchial Asthma       1.00      1.00      1.00        48
                   Cervical spondylosis       1.00      1.00      1.00        37
                            Chicken pox       1.00      1.00      1.00        38
                    Chronic cholestasis       1.00      1.00      1.00        31
                            Common Cold       1.00      1.00      1.00        34
                                 Dengue       1.00      1.00      1.00        46
                              Diabetes        1.00      1.00      1.00        35
        Dimorphic hemmorhoids(piles)          1.00      1.00      1.00        50
                          Drug Reaction       1.00      1.00      1.00        38
                       Fungal infection       1.00      1.00      1.00        33
                                   GERD       1.00      1.00      1.00        43
                        Gastroenteritis       1.00      1.00      1.00        43
                           Heart attack       1.00      1.00      1.00        42
                            Hepatitis B       1.00      1.00      1.00        47
                            Hepatitis C       1.00      1.00      1.00        40
                            Hepatitis D       1.00      1.00      1.00        38
                            Hepatitis E       1.00      1.00      1.00        50
                          Hypertension        1.00      1.00      1.00        37
                        Hyperthyroidism       1.00      1.00      1.00        42
                           Hypoglycemia       1.00      1.00      1.00        44
                         Hypothyroidism       1.00      1.00      1.00        38
                               Impetigo       1.00      1.00      1.00        36
                               Jaundice       1.00      1.00      1.00        37
                                Malaria       1.00      1.00      1.00        35
                               Migraine       1.00      1.00      1.00        39
                        Osteoarthristis       1.00      1.00      1.00        30
        Paralysis (brain hemorrhage)          1.00      1.00      1.00        38
                   Peptic ulcer diseae        1.00      1.00      1.00        31
                              Pneumonia       1.00      1.00      1.00        46
                              Psoriasis       1.00      1.00      1.00        33
                           Tuberculosis       1.00      1.00      1.00        40
                                Typhoid       1.00      1.00      1.00        41
                Urinary tract infection       1.00      1.00      1.00        41
                         Varicose veins       1.00      1.00      1.00        40
                            hepatitis A       1.00      1.00      1.00        44

                               accuracy                           1.00      1624
                              macro avg       1.00      1.00      1.00      1624
                           weighted avg       1.00      1.00      1.00      1624
```

System Architecture

Diseases are automatically assessed by the system using our model trained in medical dataset In the proposed system. This system also reflects the confidence score of the prediction. Once the expected disease has arrived, the system will refer the doctor for that disease and therefore the patient can consult to the doctor online. The proposed system acts as a decision support system and will prove to be an aid for the physicians with the diagnosis



The diagnostic system has 3 users such as doctor, patient and admin. Every user of the system is verified by the system. The system allows the patient to give symptoms and the system predicts the disease with an accurate score according to those symptoms. It then refers doctors to online consultation. The patient can then consult a physician at any time at his convenience.

3. CONCLUSIONS

Our proposed method, the Disease prediction system aim to provide better out- put results. We found almost 100% accuracy on our dataset which is a higher than any existing systems. This system will provide batter support regarding your current health status on the basis of a small survey of personal details. This project aims to predict the disease on the basis of the symptoms. The project is designed in such a way that the system takes symptoms from the user as input and predicts disease as an output. Because of our system, Patients won't have to wait for Doctors appointments, due to our system patients save their money and time. After getting the anticipated disease, the system will suggest doctors associated with that disease and therefore the patient can consult the doctor online. The proposed system acts as a decision support system and will prove to be an aid for the physicians with the diagnosis.

REFERENCES

1. Al-Aidaroos, K.M., Bakar, A.A. and Othman, Z.: Medical data classification with Naïve Bayes approach. Information Technology Journal. 11(9), 1166 (2012).

2. Asuncion, A. and Newman, D.: UCI machine learning repository downloaded from https://ergodicity.net/2013/07/.

3. J Soni, J., Ansari, U., Sharma, D. and Soni, S.: Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applica- tions, 17(8), 43-48 (2011).

4. Pattekari, S.A. and Parveen, A.: Prediction system for heart disease using Naïve Bayes. International Journal of Advanced Computer and Mathematical Sciences. 3(3), 290-294 (2012).

5. Masethe, H.D. and Masethe, M.A.: Prediction of heart disease using classification algorithms. In Proceedings of the world Congress on Engineering and computer Science. 2, 22-24 International Association of Engineers, Francisco (2014).

6 Shouman, M., Turner, T. and Stocker, R.: Using decision tree for diagnosing heart disease patients. In: Proceedings of the Ninth Australasian Data Mining Conference, Volume. 121, 23-30. Association of Computing Machinery, Victoria (2011).

7. Shouman, M., Turner, T. and Stocker, R., 2012. Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients. In: Proceedings of the International Conference on Data Science (ICDATA), pp. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (World Comp), Monte Carlo (2012).

8. Vembandasamy, K., Sasipriya, R. and Deepa, E., 2015. Heart diseases detection using Naive Bayes algorithm. International Journal of Innovative Science, Engineering & Technology, 2(9), pp.441-444.

9. https://www.kaggle.com/neelima98/disease-prediction-usingmachine- learning

BIOGRAPHIES

Er. Harjeet Singh, Assistant Professor and Researcher at LIET Gr. Noida

Mr. Ankit Mehta, Assistant Professor and Researcher at LIET Gr. Noida