Smart Health Prediction (SHP)

Author: Prof. Sangeeta Kurundkar
Electronics & Telecommunication Engineering
Vishwakarma Institute of Technology, Pune, India
Mail Id: sangeeta.kurundkar@vit.edu

Author: Mayank Jha
Electronics & Telecommunication Engineering
Vishwakarma Institute of Technology, Pune, India
Mail Id: mayank.jha20@vit.edu

Author: Nirmay Meshram
Electronics & Telecommunication Engineering
Vishwakarma Institute of Technology, Pune, India
Mail Id: nirmay.meshram20@vit.edu

Author: Shweta Munjewar
Electronics & Telecommunication Engineering
Vishwakarma Institute of Technology, Pune, India
Mail Id: shweta.munjewar20@vit.edu

***Abstract*— The "Smart Health Prediction" system uses predictive modelling to identify a patient's or user's ailment based on the symptoms they supply as input to the system. The system evaluates the user's or patient's reported symptoms as input and generates a chance of disease based on algorithmic prediction. The implementation of the Naive Bayes algorithm allows for smart health prediction. All of the features that were trained during the training phase are taken into account by the Naive Bayes algorithm when calculating the disease's % likelihood. A precise interpretation of the disease data aids in early disease prediction for the patient or user and gives the user a clear picture of the disease. Following a prediction, the user/patient can seek the advice of a specialist doctor by going to or calling the specific information.**

Keywords—* *Disease Prediction, Naïve Bayes, Predictive Model, Precaution of Disease, Food Precaution.

I. Introduction

Machine Learning Disease Prediction is a system that predicts disease according to the data the user has provided. It also predicts the patient's or user's based on the data or symptoms entered into the system to determine an illness and provides the result based on that information. It is a system which provides extra information to users about disease through which the user can maintain his/her health. Nowadays industry plays an important role in curing the disease of the patients. So, this will be a kind of help to the health industry to tell the user and also it will be useful for the user. If he/she doesn't want to go to the hospital or any clinic, then just by giving the symptoms to this system the user can get to know the disease from which they are suffering from. Sometimes we require immediate assistance from doctors, but for various reasons, they are unavailable. Reverend Thomas Bayes introduced Naive Bayes as one of the most widely used classification techniques.

In this paper we have compared three different algorithms which includes Naïve Bayes, Decision Tree and Random Forest. And using comparable results, we discovered that Naive Bayes is more accurate than the other two.

The paper is structured as follows- section II outlines the main goal of this paper. Section III represents the study of various reviews. Section IV presents a thorough explanation of the techniques employed. Followed with section V and VI with the dataset and proposed system. The various algorithms compared for the better accuracy of the system are stated in section VII. The performance is quantified in section VIII and the accuracy of different methods is compared in section IX. The result of the overall system is discussed in section X and section XI concludes the paper.

II. Aim

The goal of this research is to develop an application of machine learning techniques effectively for health prediction, which will eventually shape a suitable health prediction system for patients. This project hopes to implement a system which not only predicts the disease but also provides the precaution of disease and food precautions with a doctor's list of specific diseases.

III. Literature Review

S. Naveenkumar et al. [1] proposed that the Naive Bayes Classifier be utilised to produce smart health predictions. It extracts new patterns from historical data utilizing machine learning techniques and database management methods. H. N. Ravuvar et.al [2] mentioned Cluster size 12 has an overall performance of 87.85% compared to cluster size 10, and forecast the dataset in 10ms of computing time. The significant contribution is developing a new android-based application for a smart health-prediction system based on the K-Means algorithm. According to Ali et al. [3], illness prediction is performed by data mining and machine learning. Some of the categorization models used include decision trees, artificial neural networks, SVM, and k-nearest neighbor. Compared to other categorisation methods, artificial neural networks have a higher accuracy rate of 97%. In the study, Vivek Joshi et al. [4] employed the Naive Bayes Algorithm, Support Vector Machine, Decision Tree, KNN, and Logistic Regression.Research shows that the best method for predicting heart disease is logistic regression, which has an accuracy of 85.07%, and the best method for predicting liver disease is k nearest neighbors, which has an accuracy of 79.55%. Dr. Bindu Garg et.al [5] aims to assess data mining methods and medical industry techniques to create an accurate disease-predicting application. It is a procedure that involves techniques from database systems, artificial intelligence, machine learning, and statistics to uncover patterns in massive data sets. Kumar et al. [6] evaluated a few conventional prediction calculations and found that the proposed calculation has an anticipated exactness of 94.8% and a union speed that is faster than the CNN-UDRP computation of the unimodal disease hazard expectation. Sathya, D et al [7] targeted stacking ensemble classification algorithms to outperform naive bayes, random forest, support vector machine, k-nearest neighbor, decision tree, and logistic regression in terms of accuracy and correct prediction for various sorts of 149 illnesses. The model's accuracy is 94.19% in significantly less processing time.

Rashbir Singh [8] focused on integrating technology. This technique employs several biosensors to provide medicine and earlier disease detection. It operates machine learning with an accuracy of up to 97.50%. Nisha Gupta et.al [9], illustrated that the Random Forest algorithm is effective for detecting heart problems. Decision Tree produces roughly 75% accuracy, whereas Random Forest obtains 71.50% accuracy. The work is carried out using the libraries sci-kit-learn, pandas, matplotlib and other required libraries. For diabetes mellitus prediction, Ashok Kumar Dwivedi [10] recommended using computer intelligence techniques such as classification trees, logistic regression, support vector machines, naive Bayes, and artificial neural networks. With F 1 values of 0.83 and 0.84, respectively, artificial neural networks and logistic regression achieved classification accuracy of 77 and 78%. The decision tree was reported by Olta Llaha et al [11] as a data mining classification strategy that properly recognised diabetes data 79% of the time. The structure was developed by research into the categorization of data mining methods such as Naive Bayes, Decision Tree, SVM, and Logistic Regression. According to P. K. Sahoo et al. [12], tThe intra- and inter-cluster correlation analysis of huge healthcare data sets is what the IaCE and IeCE algorithms are designed for. With 98% accuracy, an FHCP algorithm is intended to forecast patients' future health state. The cloud-based MapReduce paradigm is used as the processing basis for huge data analysis. The approach might be used to forecast cardiac disease. S. Palaniappan et al. [13] described the Intelligent Heart Disease Prediction System (IHDPS), which was created utilising data mining techniques like Decision Trees and Naive Bayes. According to Shuo Tian et al. [14], smart healthcare is producing a new generation of medical technology that intends to entirely replace the old medical system.Smart healthcare uses big data, cloud computing, artificial intelligence, and the internet of things (loT).

M. Chen et al. [15] introduceda new method for multimodal disease risk prediction using convolutional neural networks that has a prediction accuracy of 94.8% and converges more quickly than CNN-UDRP. R. Venkatesh et al. [16] presented the BPA method, which has a greater accuracy than the previous system, 97.12%, and divides the dataset into training (60%) and testing (40%). The suggested BPA approach made use of the Spark execution environment and its in-memory cluster computing capabilities. In addition, the Naive Bayes classification technique was applied. Neesha Jothi et. al[17] have discussed various techniques researched by many researchers and various data analysis methods for various implementations in the medical field. Kun-Hsing Yu et. al [18] examined AI's economic, legal, and societal consequences in healthcare. This Review Article highlights current advances in AI technology and their biomedical applications and also identifies barriers to further advancement in medical AI systems. According to Q. Cai et al. [19], smart healthcare systems need new criteria for managing data and making decisions, with uses including disease analysis, triage, diagnosis, and therapy. also gave a comprehensive examination of existing procedures, including cutting-edge ways. Y. Liu et al. [20] suggested a digital twin healthcare-based cloud healthcare system framework (CloudDTH). CloudDTH aspires for interacting and coming together of the physical and virtual domains of medicine. To solve the issue of real-time monitoring and the precision of aging crisis warning in healthcare services, a DTH model was suggested.

IV. Methodology

Our project is a smart health prediction application, which predicts the disease using the given input symptoms and also provides the extra features like precaution of diseases, food precaution and doctor’s list. At present, when a person suffers from a disease, then that person has to visit the doctor which is time consuming and costly too. This will be a prediction model so that one can have the name of that condition based on its symptoms. And also it can be used as a confirmation tool as anyone can see the disease whatever he/she is suffering from and then they can confirm that by visiting the doctors and checking. It also provides information on available cure options such as homoeopathy, allopathy, and ayurveda, as well as their side effects and success in healing that specific condition. A successful implementation of this user interface will above all provide faster guidance to the patients in case of emergencies and help them to choose the best form of precautions considering all the available options.

V. Dataset

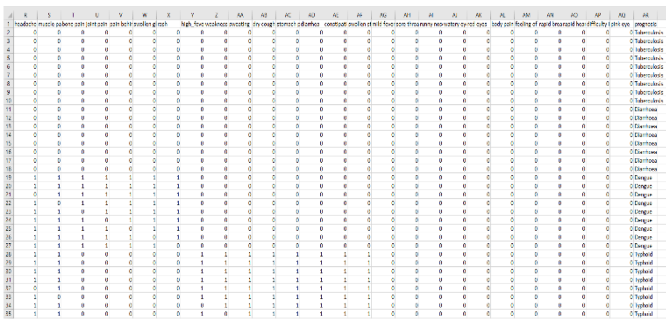

Fig 5.1. Custom Dataset (87R x 44C) for Training and (8R x 44C) for testing.

The dataset is made up of numerous symptoms and illnesses. This dataset has 44 columns in total, 43 of which are symptoms experienced by patients, and the last column gives the diagnosis for those symptoms. It consists of 7 different diseases. This is the custom dataset but we used pre pre-made dataset for having the good result and prediction which consist of 133 columns which is symptoms and 49210 rows in training dataset whereas in testing dataset, there are 42 rows which contains the disease and 133 columns are there for symptoms.

VI. Proposed System

The suggested approach is intended to address the shortcomings of the present system and offer patients with instance guidance. A lack of doctors has emerged from the pandemic catastrophe. There are also several more instances such as late night emergencies, curfews, and so forth. Our application's main goal is to give timely counsel to patients in such situations and assist them in obtaining the greatest possible precaution and doctor's availability. In our approach, illnesses are predicted automatically by utilizing the dataset to train our model. After receiving the projected disease, the system will provide the possible treatment options for the ailment. It also suggests the doctors that treat such ailment. The method allows the patient to provide symptoms, and the machine will forecast a condition based on those symptoms. Then it provides the therapy options for that ailment, along with their efficacy, side effects, and so on. It also offers doctors who treat such ailment and gives the opportunity to call the doctor. The patient can then consult a doctor whenever he wants.

VII. Algorithm

1. Naïve Bayes

The term "Naive Bayes classifiers" refers to a set of classification methods based on Bayes' Theorem. It is a family of algorithms that all share a similar premise, namely that every pair of characteristics being categorised is independent of each other.

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

Working with continuous data typically involves making the assumption that the continuous values related to each class are distributed normally (or Gaussian).

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

Continuous valued information is accepted by Gaussian Naive Bayes, which models them as Gaussian (normal) distributions.

By assuming that the data has a Gaussian distribution and that there is no covariance (independent dimensions) between the dimensions, a simple model may be developed. The only thing needed to create such a distribution is to compute the mean and standard deviation of the points within each label, which is all that is necessary to fit this model.

2. Decision tree

A decision tree is a tree structure that resembles a flowchart, with each core node standing for an attribute test, each branch for the test's result, and leaf nodes for classes or class distributions. Decision Tree algorithms like ID3, C4.5, and CART are widely used. Using a decision tree technique, the ID3 method is simple. The splitting criterion are information gain. An improvement to ID3 is C4.5. As a splitting factor, gain ratio is used. When choosing a test attribute for a given collection, the CART algorithm uses the Gini coefficient as the selection criterion and always chooses the attribute with the lowest Gini coefficient. Decision trees are advantageous because they are simple to understand and interpret when used for data classification. However, disadvantages of decision tree are:

1) The majority of algorithms (such as ID3 and C4.5) demand that the target characteristic have only discrete values.

2) Decision trees typically function well because they use the "split and conquer" strategy when several highly significant traits are available, perhaps less so in the presence of numerous complicated relationships.

3. Random Forest

The supervised machine learning algorithm known as random forest is frequently used in classification and regression applications. It creates decision trees from various samples, using the average for regression and the majority vote for classification.
The Random Forest Algorithm's capacity to work with data sets containing both continuous and categorical variables, as in regression and classification, is one of its key features. When used for categorization jobs, it gives superior results.
Before we can grasp how the random forest works, we have to first see the ensemble approach. Ensemble is just the combination of numerous models.

Ensemble employs two categories of techniques:

1. Bagging- It creates a unique training subset via replacement from the sample training data, and the result is decided by a majority vote. Consider Random Forest.

2. Boosting- It turns feeble learners become powerful learners by constructing consecutive models with the maximum accuracy. For example, ADA BOOST, XGBOOST.

Steps involved in random forest techniques:

Step 1: In a random forest model, n records at random are chosen from a set of k records.

Step 2: Each sample has its own decision tree constructed for it.

Step 3:Every decision tree will yield outcomes.

Step 4: For Classification and Regression, the final result is evaluated using Majority Voting or Averaging.

VIII. Quantifying the Performance

To quantify the accuracy of classification, we use sensitivity, specificity, and precision as the performance indicators. All of these performance indicators are based on the confusion matrix, which represents the two states of actual and the predicted.

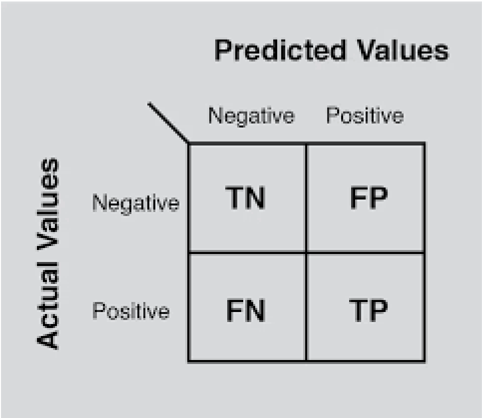

Fig 8.1. Confusion Matrix

$$Sensitivity = \frac{TP}{TP+FN}$$

$$Specificity = \frac{TN}{TN+FP}$$

$$Precision = \frac{TP}{TP+FP}$$

Here, TP - true positives (true instances predicted correctly), FP - false positives (false instances predicted as true), TN - true negatives (false instances predicted correctly), |N| -the total of true instances and |P| - total of false instances in the testing sample. Furthermore, we define F-measure, which is also called the harmonic mean between the precision and the sensitivity.

$$F\ measure = 2\left(\frac{Precision \times Sensitivity}{Precision + Sensitivity}\right)$$

This F-measure represents the weighted average of the two quantities of precision and sensitivity. The maximum value of F-measure is 1, while the minimum value is 0. Another two measures one can use to evaluate classifications method are Classification Accuracy and Error Rate. They are defined as follows.

$$Classification\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Error\ Rate = \frac{FP + FN}{TP + TN + FP + FN}$$

The use of various performance indicators will bring more insight to interpret the accuracy of the data prediction. For instance, precision is the proportion of related information out of all the retrieved information. This is a valuable indicator in almost all applications. Sensitivity measures the true-positive recognition rate. This becomes very useful where there is a high importance of classifying positives such as in security checking. In contrast, specificity measures the rate of actual negatives and it is useful in areas such as diagnosing health conditions prior to treatments.

IX. Accuracy

S. No.	Algorithm	Accuracy Rate
1.	Logistic Regression	85.07%
2.	KNN	84.51%
3.	Random Forest	84.50%
4.	Decision Tree	75.00%
5.	FHCP	98.00%
6.	CNN	94.80%
7.	BPA	97.12%
8.	Naïve Bayes(Our Approach)	80.00%

Table 9.1. These are the approaches taken by the other Authors which we have mentioned above in literature review and our approach is also here with the other ones

X. RESULT

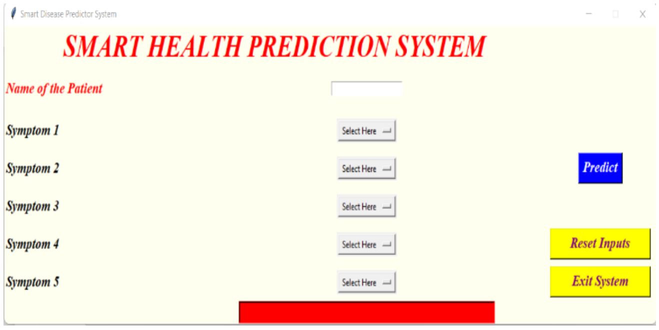

Fig 11.1. Home Screen of Smart Health Prediction System.

This is the layout of our project in which you can see the different sections where you have to give the Name and Five Symptoms. From that our system will have the Symptoms as input and there are three buttons which are Predict, Reset and Exit. From the predict button the prediction is done and from the Reset Inputs button the inputs get reset and you have to re-enter the name and symptoms. And from Exit the system gets close.

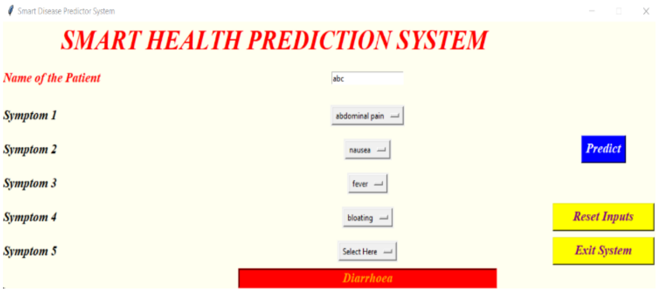

Fig 11.2. After filling the symptoms the Disease is getting predicted.

So, here you can see that we have taken the name and symptoms and from that we have the most suitable disease from which a user will be suffering.

XI. CONCLUSION

The project addresses the topic of disease prediction using symptoms and providing the disease and food precaution with the doctor's list. The project is set up so that the user's symptoms serve as input to the system and predicts disease as an output. And this will be very helpful to people who will be using this as from this one will know the disease from which he/she is suffering and they can have the treatment of that specific disease. In our project, there are extra features also from which the user will do some immediate actions like precaution on food and how they will be better if they are suffering from that disease. Doctor's list is also there in which the details of doctors are given and from that they can call doctors and can go to the doctors whoever address is given there. So one will be having all things at one place.

XII. REFERENCES

[1] *Naveenkumar, S., R. Kirubhakaran, G. Jeeva, M. Shobana, and K. Sangeetha. "Smart health prediction using machine learning." Int. Res. J. Adv. Sci. Hub 3, no. 3 (2021): 124-128.*

[2] *H. N. Ravuvar, H. Goda, S. R. and P. Chinnasamy, "Smart Health Predicting SystemUsing K-Means Algorithm," 2020 International Conference on Computer Communication and Informatics (ICCCI), 2020, pp. 1-4, doi: 10.1109/ICCCI48352.2020.9104206.*

[3] *Ali, N. Shabaz, and G. Divya. "Prediction of Disease in Smart Health Care System using Machine Learning." International Journal of Recent Technology and Engineering 8, no. 5 (2020): 2534-2537.*

[4] *Joshi, Vivek, Shipra Goswami, and Shalini Goel. "Smart Health Prediction System." Heart Disease 83 (2017): 79.*

[5] *Garg, Bindu, Aamir Hafiez, and Mayank Kumar Shalini. "E-Smart Health Prediction System using Datamining Tools." (2020).*

[6] *KUMAR, N. VIJAY, and M. UDAYA PRAKASH. "Smart Health Prediction using Data Mining with Effective Machine Learning." (2019).*

[7] *Sathya, D., T. Primya, S. Vinothini, J. Priya, and D. Jagadeesan. "SMART HEALTH SYSTEM USING STACKING ENSEMBLE CLASSIFICATION ALGORITHM."*

[8] *Singh, Rashbir. "IoT Based Smart Health Care." In IREHI. 2018.*

[9] *Nisha Gupta, Gulbakshee Dharmale, Darshana Parmar, "Heart Disease Prediction using machine learning", 2021 JETIR March 2021, Volume 8, Issue 3.*

[10] *Dwivedi, Ashok Kumar. "Analysis of computational intelligence techniques for diabetes mellitus prediction." Neural Computing and Applications 30, no. 12 (2018): 3837-3845.*

[11] *Llaha, Olta, and Amarildo Rista. "Prediction and Detection of Diabetes using Machine Learning." In RTA-CSIT, pp. 94-102. 2021.*

[12] *P. K. Sahoo, S. K. Mohapatra and S. -L. Wu, "Analyzing Healthcare Big Data With Prediction for Future Health Condition," in IEEE Access, vol. 4, pp. 9786-9799, 2016, doi: 10.1109/ACCESS.2016.2647619.*

[13] *S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," 2008 IEEE/ACS International Conference on Computer Systems and Applications, 2008, pp. 108-115, doi: 10.1109/AICCSA.2008.4493524.*

[14] *Tian, Shuo, Wenbo Yang, Jehane Michael Le Grange, Peng Wang, Wei Huang, and Zhewei Ye. "Smart healthcare: making medical care more intelligent." Global Health Journal 3, no. 3 (2019): 62-65.*

[15] *M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," in IEEE Access, vol. 5, pp. 8869-8879, 2017, doi: 10.1109/ACCESS.2017.2694446.*

[16] *Venkatesh, R., Balasubramanian, C. & Kaliappan, M. "Development of Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique". J Med Syst 43, 272 (2019). https://doi.org/10.1007/s10916-019-1398*

[17] *Neesha Jothi, Nur Aini Abdul Rashid , Wahidah Husain. "Data Mining in Healthcare – A Review". (2015) The Third Information Systems International Conference.*

[18] *Yu, Kun-Hsing, Andrew L. Beam, and Isaac S. Kohane. "Artificial intelligence in healthcare." Nature biomedical engineering 2, no. 10 (2018): 719-731.*

[19] *Q. Cai, H. Wang, Z. Li and X. Liu, "A Survey on Multimodal Data-Driven Smart Healthcare Systems: Approaches and Applications," in IEEE Access, vol. 7, pp. 133583-133599, 2019, doi: 10.1109/ACCESS.2019.2941419.*

[20] *Y. Liu et al., "A Novel Cloud-Based Framework for the Elderly Healthcare Services Using Digital Twin," in IEEE Access, vol. 7, pp. 49088-49101, 2019, doi: 10.1109/ACCESS.2019.2909828.*

[21] *Marouane Fethi Ferjani , "Disease Prediction Using Machine Learning", Computing Department Bournemouth University, p. 1*

[22] *G.Pooja reddy, M.Trinath basu, K.Vasanthi, K.Bala Sita Ramireddy, Ravi Kumar Tenali, "Smart E-Health Prediction System Using Data Mining", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-8 Issue-6, April 2019, pp. 788-789*

[23] *Wilson Wibamanto, *Debashish Das and Sivananthan, "Smart Health Prediction System with Data Mining", A/L Chelliah School of Computing & Technology, Asia Pacific University of Technology & Innovation, Kuala Lumpur, Malaysia, pp. 2-3*

[24] *Ahmad Ashari, Iman Paryudi, A Min Tjoa, "Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No. 11, 2013*

[25] *Analytics Vidhya, "Random Forest: Introduction to Random Forest Algorithm", June 24, 2021*

[26] *Yashaswi G Sagar1, Sahana Gajanana Acharya2, Vishal S Chincholi3, Riyal Vivek A4, Swetha P M5, "Medi-Insight: A Smart Health Prediction System", International Research Journal of Engineering and Technology (IRJET), Volume: 08 Issue: 06 | June 2021*