SMART HEALTH SYSTEM USING STACKING ENSEMBLE CLASSIFICATION ALGORITHM

Sathya D.(1), Primya T.(2), Vinothini S.(1), Priya J.(3), Jagadeesan D.(4)*

[1] Dept. of Computer Science and Engineering, Dayananda Sagar College of Engineering, Bangalore; [2] Dept. of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore; [3] Dept. of Information Technology, Bannari Amman Institute of Technology, Tamilnadu; [4] Dept. of Management Studies, Hindusthan College of Art's and Science, Coimbatore
India

* Corresponding Author, e-mail: sathy.spj@gmail.com

Abstract: Data Mining is a powerful technology and is used to identify useful and understandable patterns by analyzing large sets of data. It gives a detailed view of various disease predictions. It will be more useful especially in pandemic times. During these days, doctors are in the front line and battling with the COVID-19 virus. It will be hard for people to immediately get medical guidance or appointments. Our proposed system, the smart health application will come in handy at these times. The system allows people to get medical guidance for their health issues. Also, the system is fed with symptoms and the disease-related with it which will give high accuracy for disease prediction. Our model aims to use Stacking Ensemble Classification Algorithms to give high accuracy and correct prediction than Naive Bayes, Random Forest, Support Vector Machine, K – Nearest Neighbor, Decision Tree, Logistic Regression for different types of 149 diseases. The GUI is designed which can be used easily to predict the different types of disease accurately.

Key words: Classification, Diseases, Ensemble Learning, Machine Learning, Symptoms.

1. INTRODUCTION

Diseases in humans are one of the major causes of death in the world [1]. The number of people affected by many diseases increases irrespective of age in both men and women. Symptoms like heart attack, asthma, cancer and soon contribute too many diseases. Prediction and analysis of major diseases by considering the parameters like diseases and symptoms will be more useful these days since it is hard for people to get immediate medical guidance. In health concerns, Data mining plays an important role in predicting diseases. Data mining helps to predict the pattern of similar symptoms accurately to analyse the diseases. Classification is one of the important methods in machine learning to predict diseases more accurately. Machine Learning algorithms is a probabilistic model and considered as one of the major optimization methods to learn from past experience and detect useful patterns from the complex dataset [2].The medical field has enormous amount of data that can be processed by the data mining techniques. It could be useful if some patients need doctor's help but they are not nearer to patient's location. Many existing approaches had predicted the medical data using traditional classification algorithms which does not provides accurate results. The accurate predictions save many lives [19]. So, the proposed work has concentrated on giving accurate decision and also developed the GUI system which provides fast and accurate results, and easy to use. The paper is structured as follows: Section 2 describes the related work, Section 3 about materials and methods, Section 4 and 5 explains the implementation, section 6 discusses the results and Section 7 explains the conclusion.

2. RELATED WORK

In [3], mainly uses data mining techniques such as clustering, classification, regression, etc. The system uses various algorithms like Naive Bayes, Linear Regression, Discriminant Analysis, etc. It concludes that the Neural Networks reported higher accuracy compared to other models and it should necessarily be trained in a better way to get higher accuracy. In [4], proposed a system on health diagnosis, and the system is used for patient data analysis and disease diagnosis at various levels for some patients who require continuous check-up. A Naive Bayes algorithm is used in this system to get analysis to predict the diseases.

In [5], focused on developing a web application for predicting diseases by mining data sets and provides remedial solutions for effective treatment based on the symptoms. The system uses Apriori and FB growth algorithms, which helps to reduce the human effort and cost in terms of human resources. The system is used for Heart Disease or Diabetes prediction. In [6], focused on developing a model used for predictive analysis on the data mining and machine learning algorithms to predict the diseases on patients. Naive Bayes and Logistic Regression algorithms are used for implementing the system. The system helps to decrease the gap between medical patients and the doctor's availability time. In [7], the system uses Logistic regression algorithm. The system used the dataset consists of various features of breast cancer which allows patients to share their symptoms related to breast cancer and get the results for cancer that could be associated with it. The system automatically shows the result with a specific accurate percentage for breast cancer. In [8], implemented a system that predicts the disease by the symptoms. It also suggests a doctor available in the nearest possible area. The system has two types of modules such as Treatment Specialist Doctor and Patient, which uses the

Naive Bayes algorithm for the implementation to find out the possible diseases. In [9], focused on providing the solution for detecting heart diseases given various symptoms. Data Pre-Processing, Classification, and Anamoly detection were done for processing and used Random Forest classification algorithm. In [10], system uses IoT cloud based system and BigData techniques for storing the real time health data and gives alert on emergency situations. The cloud based health information system is simple to use and also helpful for future references [14]. In [11], proposed a solution for predicting diseases using Naive Bayes and Decision Tree Algorithms. The system predicts the diseases when the symptoms are given by the patient and implemented using the tomcat server. In [12], focuses on Heart Diseases and Diabetes. The system uses the Naive Bayes Algorithm and R shiny an open-source package in R to display the result. All the existing disease prediction system gives accuracy with basic classification algorithms like Naive Bayes, Random Forest, Support Vector Machine, etc. All the existing solutions covered specific diseases by considering the related symptoms. To overcome all these problems, we proposed a solution using a database from the case study of disease-symptom associations [13]. Our model uses Stacking Ensemble classification algorithms along with Naive Bayes, Random Forest, Support Vector Machine, K – Nearest Neighbor, Decision Tree, Logistic Regression for the prediction of major 149 diseases as a replacement for existing systems and displaying the predicted result using GUI.

3. MATERIALS AND METHODS

The main purpose of the project is to create a system that will help people to get medical guidance because it is hard to get doctor appointments immediately these days. The smart health prediction system helps people in a timely manner. The system helps people for the identification of major 149 diseases accurately. The system fed with diseases and symptoms. Based on the symptoms, it predicts diseases. A GUI is used for the purpose where people can enter the symptoms and get the results back. Smart Health application uses the Stacking Ensemble classification algorithm for the analysis of diseases and the associated symptoms which proves that this algorithm gives higher accuracy and helps to make predictions accurately. Users have to enter the name and can use the drop-down box feature to enter the symptoms. The system predicts the diseases with high accuracy. People can use this system anywhere and get guidance immediately. Figure 1 describes the overall system implementation. A System Architecture defines the project structure and the different modules of the system. Data Collection is the first module where the dataset is collected. Then Data Pre-Processing step gives a cleaned dataset. The cleaned dataset is trained and made the prediction using different models and displaying the predicted result using the GUI system followed by evaluation of different algorithms prediction.

3.1. System Initialization Steps

The system initialization is the first module consists of three steps that explain the system implementation.

- **Data Collection**
  The dataset is collected from the study of the University of Columbia at New York-Presbyterian Hospital where patients were admitted during 2004 [13]. The dataset mainly consists of three attributes namely the different types of disease, the number of occurrences of the patient, and the different types of symptoms. Each type of disease has its associated symptoms. The dataset contains 149 different types of diseases. Each disease has a minimum of 15 to a maximum of 50 symptoms. Diseases and symptoms are of string type and the patient count is of numeric type.
- **Data Preprocessing**
  The majority of real-world data consists of incomplete, inconsistent, and containing many errors. All the data gets transformed or encoded to give the data in a state in which the machine can easily understand. Data pre-processing is a necessary step before building a model with the main features. We considered 4 steps for data pre-processing. The data preprocessing steps are given as follows,
- **Eliminating Unwanted Numbers, Codes, and Strings**
  We eliminated all the unwanted codes, characters, and numbers from the original dataset. The symptoms and diseases in the dataset contain different names for similar types of diseases and symptoms. We stored all the same types of symptoms with different disease names for better prediction and classification and also for getting higher accuracy.
- **Converting String into Binary Data**
  We replaced all the symptoms in the column with binary values [15]. We replaced all the categorical data such as disease and symptoms in the dataset with binary values. Dealing with binary values can be easily predicted because it lies in the case like true or false. After replacing all the symptoms, each disease in the rows having a single symptom in column value depicting the disease - symptoms association.
- **Converting Diseases as Indices**
  Having numerical values is easier for making predictions than handling the combination of string and numeric values. We replaced diseases with numerical values starting from 0 to 148. After replacing, the dataset contains similar numeric values for some diseases because all the diseases having a minimum of 15 to a maximum of 50 symptoms.
- **Grouping all common Diseases**
  Handling similar kinds of data results in over-fitting. We grouped all the similar kinds of repeated data for higher-level prediction. Finally, the fully

cleaned dataset looks like having Diseases in the rows followed by Symptoms in the column.

- **Training the Dataset**
  Proper training of the dataset leads to higher accuracy. The prediction of the dataset is made by training the whole dataset because each disease is new to everyone. Different types of diseases and related symptoms are trained properly to get accurate prediction results using different types of classification algorithms.

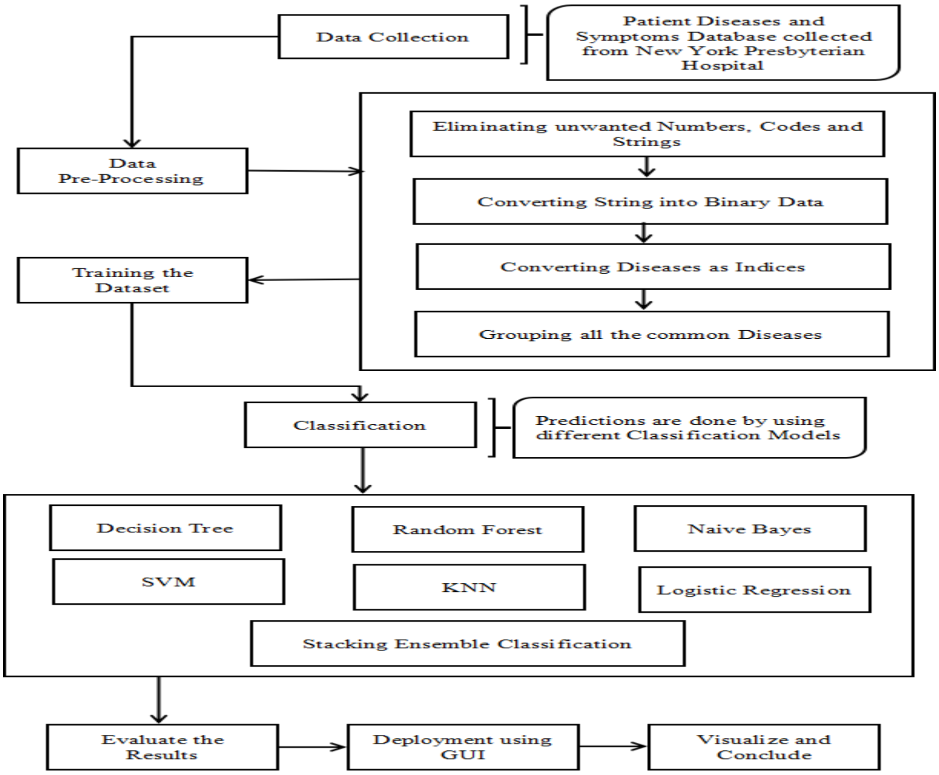

Data Collection
Patient Diseases and Symptoms Database collected from New York Presbyterian Hospital
Data Pre-Processing
Eliminating unwanted Numbers, Codes and Strings
Converting String into Binary Data
Converting Diseases as Indices
Grouping all the common Diseases
Training the Dataset
Classification
Predictions are done by using different Classification Models
Decision Tree
Random Forest
Naive Bayes
SVM
KNN
Logistic Regression
Stacking Ensemble Classification
Evaluate the Results
Deployment using GUI
Visualize and Conclude

Fig. 1. System Architecture

4. IMPLEMENTATION AND EVALUATION OF CLASSIFICATION ALGORITHMS

The dataset majorly consists of various types of diseases and associated symptoms. Different models are built using different types of supervised machine learning techniques or classification approaches and the performances are analysed to get a better insight. Important features are extracted from the training dataset and are fed to the model along with the class labels for making predictions. We trained our dataset and made predictions using different types of models.

The system uses 6 different types of individual characteristic classification algorithms and a combination of algorithms such as Ensemble Stacking classification. The list of different classification algorithms used for comparison are:

i. Decision Tree
ii. Random Forest
iii. Naive Bayes
iv. Support Vector Machine
v. K – Nearest Neighbor
vi. Logistic Regression
vii. Ensemble Stacking Classification

4.1. Ensemble Stacking

Stacking is considered as an Ensemble Machine Learning Algorithm [25]. It uses a meta-learning algorithm to find out the way to combine the predictions from two or more base machine learning algorithms and gives the best results. It has the advantage to harness the capabilities of different models on classification or regression tasks and gives higher performance than any single characteristics model. Ensemble Stacking involves two or more base models called level-0 and a meta-model called level-1. Every time with the use of the Stacking Ensemble classification algorithm, the dataset is divided into k splits for higher-level prediction with the class labels. Each classifier made some kind of prediction. With the prediction, all the outputs are compared and give the final prediction based on the higher similar prediction made by individual characteristics algorithms. Fig.2. shows the diagrammatic representation of the Ensemble Stacking classification algorithm that helps to combines the results from different algorithms.

x = Dataset (Diseases and Symptoms) and, Classifiers are referred to as different types of algorithms like Random Forest, Decision Tree, etc. and, Different Classes are referred to as the output of the particular algorithms.

Stacking is considered one of the best Ensemble Machine Learning algorithms which learn how to best combine the predictions from different types of well-performing individual characteristics machine learning models.

Level-0 Model (Base-Model): It fits the training data and compiles the predictions.

Level-1 Model (Meta-Model): Ensemble model learns the way to best combine the predictions of the base-level models.

K-fold cross-validation is used to prepare the training dataset for the meta-model [25]. The dataset is used to train the model set to be k-1. It is a resampling method used for the evaluation of machine learning models on a limited data sample. 'k' means the number of groups that a given data sample is to be split into. Linear models are frequently used as the meta-model such as linear regression for regression tasks and logistic regression for classification tasks. By using Ensemble

Stacking classification, we got 94.19% accuracy. Ensemble Stacking classification algorithms Accuracy, Precision, F1, Recall values are shown in Table 2.

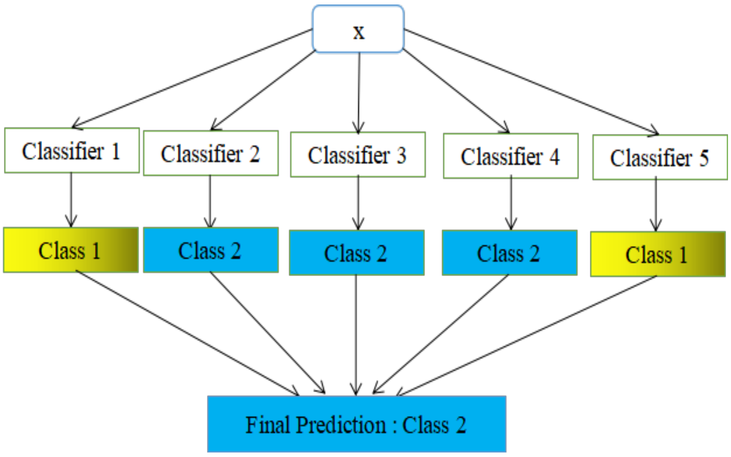


Fig. 2. Diagrammatic Representation of Ensemble Stacking Classification

Based on the prediction made by different classification algorithms, we got 88% (Random Forest) as high accuracy for the disease prediction dataset in the case of individual characteristics algorithms. The system aims to provide higher accuracy than individual characteristics classification algorithms. We analyzed the disease dataset using the Ensemble Stacking classification algorithm which uses a combination of many algorithms as base models and predicts using a meta-model algorithm. Successfully, we got higher accuracy than individual characteristic algorithms like Naive Bayes, Random Forest, Decision Tree, KNN, Logistic Regression, SVM. After the successful prediction using Ensemble Stacking Classification, we got 94.19% accuracy which is considered as the best algorithm for disease prediction.

5. GUI IMPLEMENTATION FOR SMART HEALTH PREDICTION SYSTEM

The system implemented using GUI element for displaying the type of disease, for entering different types of symptoms, patient name, and also for the disease prediction buttons. Fig. 3 shows the GUI template for Smart Health Application that helps to display the predicted diseases. We build our GUI and make use of the widgets where it needs to get pre-compile everything with window.mainloop() or root.mainloop(). Using Tkinter's, it is faster to develop an interface and basically, it comes bundled with Python [16]. Widgets are considered as the elements of Tkinter (like buttons, option menu, labels, frames, entries, etc.).

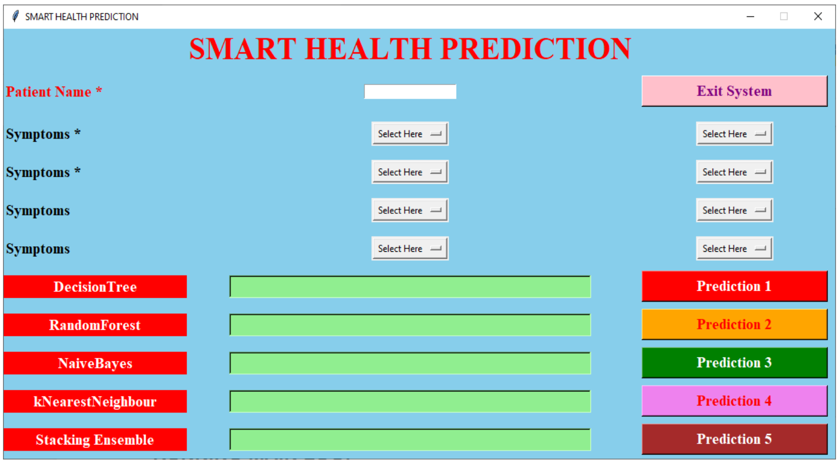


Fig. 3. GUI Template for Smart Health Prediction

We analysed 50 random diseases in the dataset and from the analysis, Ensemble Stacking predicted 49 diseases accurately. Table 1 shows different algorithms disease prediction scores and also reporting that Ensemble Stacking Classification correctly predicted 49 diseases out of 50 random diseases than other classification algorithms.

Table 1. Different Algorithms Disease Prediction Scores (out of 50 random diseases)

ALGORITHMS	***CORRECT PREDICTION RESULT***
Stacking Ensemble	*49*
Naive Bayes	*48*
Random Forest	*47*
K-Nearest Neighbor	*35*
Decision Tree	*32*

6. RESULTS AND DISCUSSION

After successful implementation of different types of classification algorithms, we are listing each of the different classification algorithm's performance like accuracy, precision, f1, and recall scores in table 2. Accuracy is the measurement that gives values closes to a specific value [18]. Both precision and recall give results by supporting relevance. F1 score is calculated from the weighted average of Precision and Recall. Fig. 4 shows the different models performance comparison based on accuracy, precision, f1, and recall scores using bar plot. Table 2 & Fig. 4 show that ensemble stacking gives very good performance than other algorithms. This is because ensemble stacking finds best results based on the combination of different algorithms. Our analysis proves that the performance of ensemble stacking classification is better than other models performance. Fig. 5 shows some sample diseases prediction results using the GUI system and Ensemble Stacking predicted the diseases correctly than other classification algorithms.

Table 2. Different Algorithms Performance Results

ALGORITHMS	***ACCURACY***	***PRECISION***	***F1***	***RECALL***
Decision Tree	*84 %*	*0.751572*	*0.764151*	*0.792453*
Naive Bayes	*84 %*	*0.751572*	*0.764151*	*0.792453*
Random Forest	*88 %*	*0.820261*	*0.833333*	*0.862745*
SVM	*86 %*	*0.785256*	*0.798077*	*0.826923*
KNN	*84 %*	*0.751572*	*0.764151*	*0.792453*
Logistic Regression	*82 %*	*0.719136*	*0.731481*	*0.759259*
Ensemble Stacking	*94.19 %*	*0.853468*	*0.868009*	*0.899329*

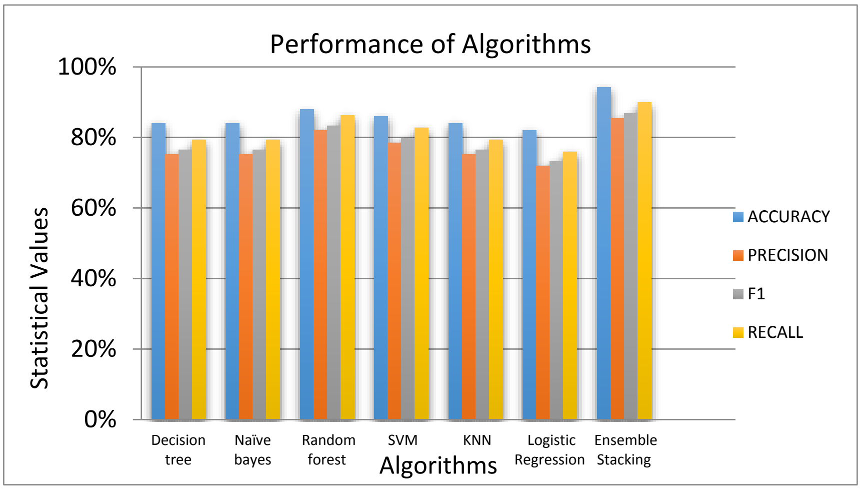
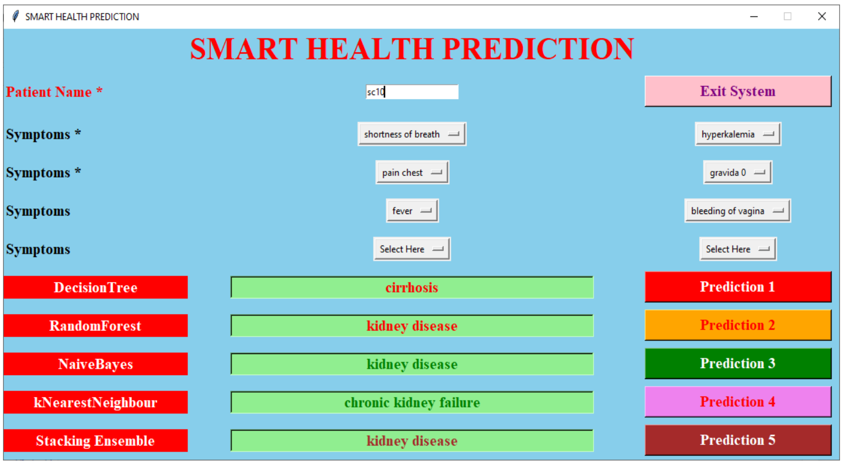


Fig. 4. Different Algorithms Performance Score using bar plot



SMART HEALTH PREDICTION

Patient Name * sc10

Symptoms * shortness of breath hyperkalemia

Symptoms * pain chest gravida 0

Symptoms fever bleeding of vagina

Symptoms Select Here Select Here

DecisionTree cirrhosis

RandomForest kidney disease

NaiveBayes kidney disease

kNearestNeighbour chronic kidney failure

Stacking Ensemble kidney disease

Fig. 5. Sample Disease Prediction

7. CONCLUSION

The proposed system can be used in hospitals, clinical and all health-related areas where human safety and prevention are of primary concerns. The benefits of Stacking Ensemble classification algorithms are quite remarkable. Ensemble Stacking classification combines five different algorithms in the system and predicts the disease based on the result of the algorithms to yield high accuracy. It helps to identify human health-related issues with high accuracy. We implemented a Smart Health system using GUI. The accuracy obtained for the Stacking Ensemble model is 94.19% respectively in a very less computational time and considered as the best algorithm. In the future, Web App development of the project can be developed and the algorithm needs to be implemented for specific type of disease.

REFERENCES

[1] https://www.health.harvard.edu/blog/important-health-problems-matter-2016091510267.

[2] https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-1004-8.

[3] N. Shabaz Ali, G. Divya. Prediction of Diseases in Smart Health Care System using Machine Learning. *International Journal of Recent Technology and Engineering*. Vol 8, No. 5, 2020, pp. 2277-3878.

[4] Prof. Krishna Kumar Tripathi, Shubham Jawadwar, Siddhesh Murudkar, Prince Mishra Professor. A Smart Health Prediction Using Data Mining. *International Research Journal of Engineering and Technology*, Vol.5, No.4, 2018, pp. 2395-0056.

[5] Pinky Saikia Dutta, Shrabani Medhi, Sunayana Dutta, Tridisha Das, Sweety Buragohain. Smart Health Prediction System using Data Mining. *International Journal of Current Engineering and Scientific Research*, Vol.4, No.8, 2017, pp. 2393-8374.

[6] Vidya Zope, Pooja Ghatge, Aaron Cherian, Piyush Mantri, Kartik Jadhav. Smart Health Prediction using Machine Learning. *International Journal for Scientific Research & Development*, Vol. 4, No.12, 2017, pp. 2321-0613.

[7] Manisha M S Pillai, Rahul Gopal, Roshitha Mariam Sunny3, Revathy Chandran, Akhila Balachandran. Smart Health Prediction System Using Python. *International Journal of Computer Sciences and Engineering*, Vol. 7, No.5, 2019, pp. 2347-2693.

[8] Nikita Kamble, Manjiri Harmalkar, Manali Bhoir, Supriya Chaudh. Smart Health Prediction System Using Data Mining. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*,Vol. 2, No.2, 2017, pp. 2456-3307.

[9] E.Kodhai, K.Vagulamaliga, S.Mirudhula, S.R.Pavithra, K.Rama. Smart Disease Prediction using Effective Vector Machine Algorithm. International Journal of Pure and Applied Mathematics. Vol. 116, No.5, 2017, pp. 55-59.

[10] Awotunde Joseph Bamidele, Rasheed Gbenga Jimoh, Roseline Oluwaseun Ogundokun, Sanjay Misra, Oluwakemi Christiana Abikoye, Big Data Analytics of IoT-Based Cloud System Framework: Smart Healthcare Monitoring Systems. *Artificial Intelligence for Cloud and Edge Computing,* 2022, pp. 181-208, https://doi.org/10.1007/978-3-030-80821-1_9.

[11] 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*, Vol.8, No.6, 2019, pp. 2278-3075.

[12] Harshitha M, B M Sagar. Smart Health Care Implementation Using Naïve Bayes Algorithm. *International Journal of Innovative Research in Computer Science & Technology,* Vol.7, No.3, 2019, pp. 2347-5552.

[13] https://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html

[14] Ajayi, Priscilla, Nicholas A. I. Omoregbe, Sanjay Misra, Davies Adeloye. Evaluation of a cloud based health information system. *Innovation and Interdisciplinary Solutions for Underserved Areas. Springer, Cham*, 2018, pp. 165-176.

[15] https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9feb.

[16] https://www.geeksforgeeks.org/what-are-widgets-in-tkinter/.

[17] Sanjay Misra. A Step by Step Guide for Choosing Project Topics and Writing Research Papers in ICT Related Disciplines, *Communications in Computer and Information Science*, *Springer*, Vol. 1350, 2021.

[18] N. Shabaz Ali, G. Divya, Prediction of Diseases in Smart Health Care System using Machine Learning, *International Journal of Recent Technology and Engineering,* Vol. 8 , No.5, 2020, pp. 2277-3878.

[19] https://www.javatpoint.com

Information about the authors:

Dr. D. Sathya is presently working as Associate Professor in the department of Computer Science and Engineering at Dayananda Sagar College of Engineering, research area is data security in sensor networks.

Ms.T.Primya is presently working as Assistant Professor in the department of Computer Science and Engineering at KPR Institute of Engineering and Technology, research interest in cloud computing and IoT.

Dr. S. Vinodhini is presently working as Associate Professor in the department of Computer Science and Engineering at Dayananda Sagar College of Engineering, research area is cloud computing.

Ms. J. Priya, Research scholar at Bannari Amman Institute of Technology, research area is machine learning techniques.

Mr. D. Jagadeesan is presently working as Assistant Professor in the department of Management studies at Hindusthan College of Arts and Science, research area is marketing.

Manuscript received on 01 June 2022