e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 8, August 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

© 2024 IJMRSET | Volume 7, Issue 8, August 2024| DOI: 10.15680/IJMRSET.2024.0708045

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Machine Learning Model for Predicting Chronic Diseases

Dr. Rajesh K S[1], Mamatha K R[2], Abishyanth S[3], Shrunga G V[4], Srujana V H[5], Varsha M K[6]

Professor, Department of Artificial Intelligence and Machine Learning, Rajarajeswari College of Engineering, Bengaluru, Karnataka, India[1]

Assistant Professor, Department of Artificial Intelligence and Machine Learning, Rajarajeswari College of Engineering, Bengaluru, Karnataka, India[2]

UG Students, Department of Artificial Intelligence and Machine Learning, Rajarajeswari College of Engineering, Bengaluru, Karnataka, India[3,4,5,6]

ABSTRACT: The healthcare field is one of the most essential areas of research in the modern period, thanks to rapid changes in technology and data. It's difficult to keep track of a large amount of patient records. The use of Big Data Analytics allows us to manage this content. Various ailments can be treated in a variety of ways all throughout the world. Machine Learning is one of the approaches for disease prediction and diagnosis. This study reveals how symptoms can be utilized as input parameters in machine learning to produce disease predictions. On the dataset, the machine learning technique Random Forest is used to forecast the disease. The predicted disease and the symptoms are stored on the database. It is implemented using the Python programming language, and a graphical user interface (GUI) has been created to display the findings. The Prediction of disease can be done by classifying the given dataset.

KEYWORDS: Artificial Intelligence, Machine Learning, Graphic User Interface, Python Programming, Data Science, Random Forest, Support Vector Machine, Naive Bayes.

I. INTRODUCTION

Chronic diseases pose a significant challenge to healthcare systems globally. However, machine learning (ML) is emerging as a powerful tool, fundamentally transforming how we predict, diagnose, and manage these conditions.
At the heart of ML's impact lies its ability to analyze vast datasets encompassing diverse information. This includes detailed patient medical records, intricate genetic data, and even environmental factors. Traditional methods often struggle to uncover the subtle correlations and complex relationships hidden within this data. Machine learning algorithms, however, excel at identifying these nuanced connections, empowering healthcare professionals with valuable insights.

As technology advances and healthcare data sets become even more complex, the ongoing refinement of ML models holds immense potential. With continued development, we can expect even earlier interventions and significantly improved patient outcomes in the fight against chronic diseases. The integration of ML into healthcare practices marks a paradigm shift, paving the way for a future of precision medicine and more effective chronic disease management.

II. RELATED WORK

[1] Nowadays, humans face various diseases due to the current environmental conditions and their living habits. 'e identification and prediction of such diseases at their earlier stages are very important, so as to prevent the extremity of it. It is difficult for doctors to manually identify the diseases accurately most of the time.[2] Utilizing data mining and machine learning approaches to diagnose chronic diseases due to their cost-effectiveness and ability to analyze large amounts of patient data to identify patterns that may not be detectable by human experts. This enables early detection of diseases, which can improve patient outcomes.[4] need an accurate, feasible, reliable, and robust system to diagnose

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

diseases in time so that these can be properly treated. With the growth of medical data, many researchers are using these medical data and some machine learning algorithms to help the healthcare communities in the diagnosis of many diseases. In this paper a survey of various models based on such algorithms, techniques is presented and their performance is analyzed. Researches have been conducted on various models of supervised learning algorithms and some of them are Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), Naïve Bayes and Random Forest (RF). [5] Model is used to identify the relation between web search queries submitted by population and clinical data available in official health. For measuring chronic diseases, the clinical data has been collected from centers Disease Control and Prevention (CDC). These clinical data were obtained from a non-communicable disease surveillance system in the USA. In addition, these data sets reported a weekly percentage of patients who are suffering from chronic diseases. The Log transformation method proposed can normalize the data. An evaluation phase used the same metrics namely MSE, RMSE, and MAE for testing the proposed model. [7]Data mining is a process of knowledge discovery from unknown or useless datasets. There are various techniques of data mining that are used to process the data and convert them as useful information. Data mining can be used in various fields such as business analysis, healthcare, stock management etc. Medical field has a wide amount of data that can be processed by the help of data mining techniques. It might have happened before that you or someone near you wanted immediate help from a doctor but could not find anyone. Creating a model that can predict the diseases based on user symptoms is quite helpful in getting fast and appropriate medical facilities for patients. The timely analysis of data and gaining accurate prediction of diseases from symptoms can save many lives. Early detection of diseases helps doctors to give accurate medication. In the field of medicine different algorithms of machine learning are used for predicting different diseases and helps the physicians to diagnose fast. Based on the input of data the accuracy of results may vary.

III. METHODOLOGY

A Machine Learning based project to understand and analyze the symptoms to predict what kind of chronic disease the patient is suffering from is done. This Model is designed to read the symptoms of its user, analyze to what disease the symptom matches the most, and try to predict the actual disease as a result. User's or Patient's symptoms are the key considerations when determining the particular chronic disease the user is suffering from. Datasets from various medical sources regarding the symptoms that patients usually face during various chronic disease suffering are collected and used to train our Machine Learning model. Data processing, analysis and Data Visualization is done by means of data science techniques in order to prepare data for our model. Various ML Algorithms are also used to train our model for prediction analysis. The block diagram of the developed ML Model is shown in Figure 3.1.

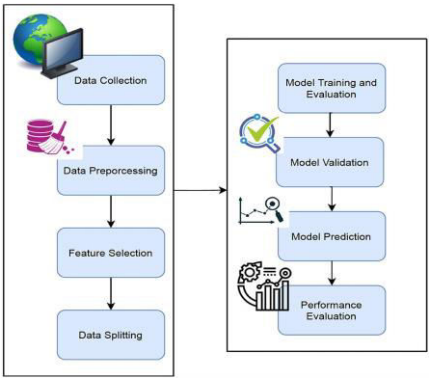


Figure 3.1 Block diagram of developed ML Model

The details of various algorithms and data processing techniques used in the developed Prediction Model are discussed as follows:

Scikit Learn Library

scikit-learn is a free software machine learning library for the Python programming language.[3] It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient

© 2024 IJMRSET | Volume 7, Issue 8, August 2024| DOI: 10.15680/IJMRSET.2024.0708045

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

boosting, *k*-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. Utilizing this, we have incorporated various algorithms of this library to build our model.

Decision Tree Classifier

Decision tree classifier or learning is a supervised learning approach used in statistics, data mining and machine learning. In this formalism, a classification or regression decision tree is used as a predictive model to draw

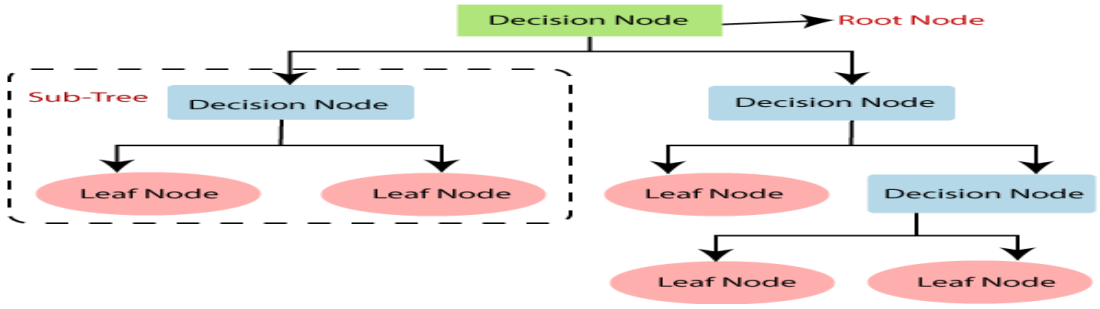


Figure 3.2 Decision Tree Classifier

Naive Bayesian Model

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

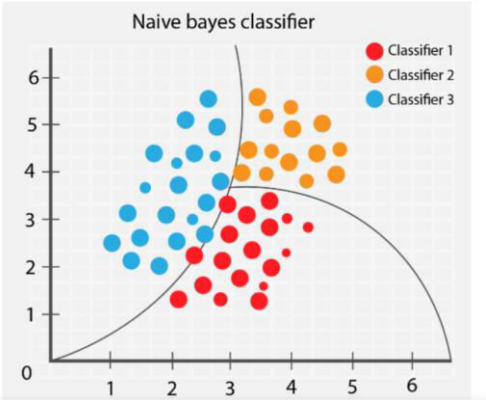


Figure 3.3 Naive Bayesian Classifier

Random Forest Classifier

Random forests or **random decision forests** is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

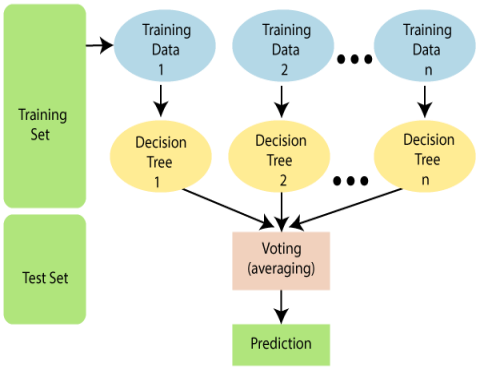


Figure 3.4 Random Forest Classifier

Support Vector Machine

In machine learning, support vector machines (SVMs, also support vector networks) are supervised max-margin models with associated learning algorithms that analyze data for classification and regression analysis.

Symptoms Data Set

Here, we define, collect, and report chronic disease data that are important to public health practice and available for states, territories and large metropolitan areas.

Figure 3.5 Chronic Disease data information.

Streamlit

Streamlit is an open-source Python library that makes it easy to build beautiful custom web-apps for machine learning and data science. It is a simple and powerful app model that lets you build rich UIs incredibly quickly.

The Machine Learning models require large amounts of data to train effectively. In the case of chronic disease prediction, relevant data may include medical records, lab results, patient demographics, lifestyle factors, genetic information, and possibly even environmental data.

Features are the variables used by the model to make predictions. Feature selection involves identifying the most relevant features from the collected data that are likely to have an impact on the prediction of chronic diseases. This step often involves domain expertise and statistical analysis.

Before feeding the data into the ML model, it needs to be preprocessed. This includes handling missing values, normalizing or standardizing numerical features, encoding categorical variables, and possibly performing feature engineering to create new features that may better capture relationships in the data.

There are various ML algorithms that can be used for predictive modeling, including logistic regression, decision trees, random forests, support vector machines, and neural networks. The choice of algorithm depends on factors such as the

© 2024 IJMRSET | Volume 7, Issue 8, August 2024| DOI: 10.15680/IJMRSET.2024.0708045

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

nature of the data, the size of the dataset, interpretability requirements, and computational resources available. Once the algorithm is selected, the model is trained using the preprocessed data. During training, the model learns the underlying patterns and relationships in the data to make predictions about whether a person is likely to develop a chronic disease based on their features.

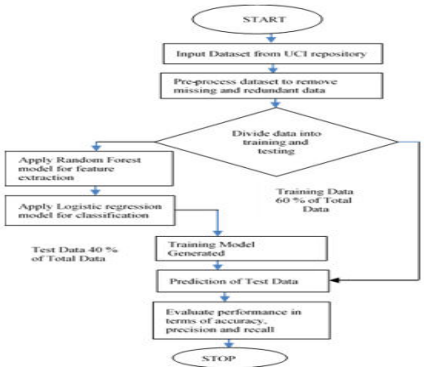
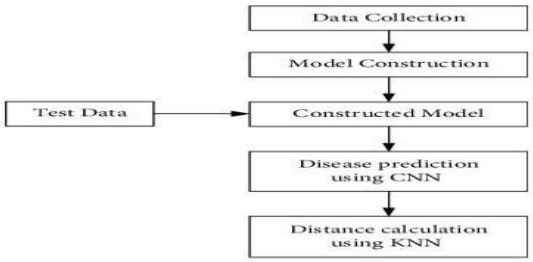


Figure 3.6 Work flow of the model

The model is then integrated with User Interface for the user to interact and use the model to get their results instantaneously, by simply selecting their symptoms and submitting them.

Methodology of Machine Learning model.

This Chapter consists of the Dataset used for training the model, training, and classification of the trained model, and feature selection. To train the model and evaluate its effectiveness, a machine learning model was used. The model based on the Multinomial Naïve Bayes algorithm was applied.



Collecting the Raw Data

Data collection is the process of compiling and examining data from various sources. Data collecting allows one to maintain track of prior events so that data analysis can be used to find repeating patterns. The Kaggle website is where the dataset for "Crop Recommendation" is gathered. The dataset includes 2 characteristics and class labels for 22 different crops. In the below dataset, we are taking (i) Temperature in degrees Celsius and (ii) Relative Humidity percentage.

Data Preprocessing

Data preprocessing is the process of transforming raw data into a form learning algorithms can use to uncover insights or predict outcomes. Finding missing values is the data processing technique used in this study. It is difficult to obtain every data point for each record in the dataset. A lack of data may be indicated by empty cells, values like null, or a particular character like a question mark. There were no missing values in the dataset that was used for the research.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Train and Test Split
Using the train test split() method of the sci-kit learn module, the dataset is divided into a training dataset and a testing dataset. The dataset was split into two parts: a training dataset that makes up 80% of the total, and a testing dataset that makes up 20%.

Fitting the model
Fitting is the process of altering the model's parameters to improve accuracy. On data for which the target variable is known, an algorithm is run to build a machine-learning model. By contrasting the model's outputs with the target variable's actual, observed values, the accuracy of the model is assessed. A machine learning model's capacity to generalize data similar to that with which it was trained is known as model fitting. A model that accurately approximates the output while the inputs are uncertain is referred to as having a good model fit.

Checking the score over a training dataset
Using a trained machine learning model, scoring, also known as prediction, is the process of generating values from fresh incoming data. The score of each model over a training dataset is calculated using the model.score() method, which demonstrates how effectively the model has learned.

Predicting the model
"Prediction" refers to the outcome of an algorithm after it has been trained on a prior dataset
and applied to new data when predicting the likelihood of a particular result. Utilizing the test feature dataset and the predict() method to predict the model The output was provided as an array of forecasted numbers.

Gaussian Naive Bayes algorithm.
The naive Bayes classifier calculates the probability of an event in the following steps:
Step 1: Determine the class labels' prior probability given the current temperature and relative humidity.
Step 2: Find the Likelihood probability with each attribute for each class.
Step 3: Put these values in Bayes Formula and calculate posterior probability.
Step 4: See which class has a higher probability, given the input belongs to the higher probability class.
How well our Sklearn Multinomial Naive Bayes model predicted crop using the test data is indicated by the accuracy model score is **~100%**

Confusion Matrix and Classification Report
Confusion Matrix and Classification Report are the methods imported from the metrics module in the scikit learn library that are calculated using the actual labels of test datasets and predicted values.

Confusion Matrix gives the matrix of frequency of true negatives, false negatives, true positives and false positives.

Classification Report is a metric used for evaluating the performance of a classification algorithm's predictions. It gives three things: Precision, Recall and f1-score of the model.

Precision refers to a classifier's ability to identify the number of positive predictions which are relatively correct. It is calculated as the ratio of true positives to the sum of true and false positives for each class.

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

Where Precision-Positive Prediction Accuracy; TP-True Positive; FP-False Positive.

Recall is the capability of a classifier to discover all positive cases from the confusion matrix. It is calculated as the ratio of true positives to the sum of true positives and false negatives for each class.

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

© 2024 IJMRSET | Volume 7, Issue 8, August 2024| DOI: 10.15680/IJMRSET.2024.0708045

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Where Recall- The percentage of positives that were correctly identified; FN-False Negative.

F1 score

A weighted harmonic mean of precision and recall, with 0.0 being the worst and 1.0 being the best. Since precision and recall are used in the computation, F1 scores are often lower than accuracy measurements.

$$F1\ score = \frac{2 * PR}{(P + R)}$$

Where P-Precision; R-Recall

Accuracy

The number of correct predictions divided by the total number of predictions is known as model accuracy. The accuracy of the model is calculated using the accuracy score() method of the scikit learn metrics module. The accuracy model score is 88.67%

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

IV. EXPERIMENTAL RESULTS

This section is separated into outputs from User Interface and outcomes from machine learning models. The output will be predicted by the machine learning model if the output of the model is equivalent to the input.

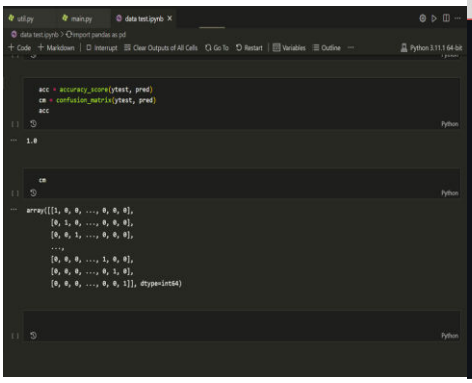


Figure 4.1 Evaluation of the model

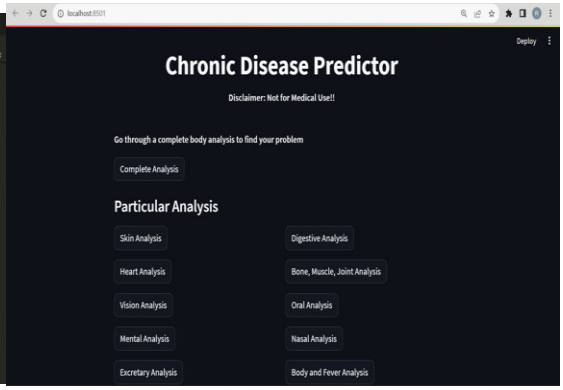


Figure 4.2 Appearance of UI

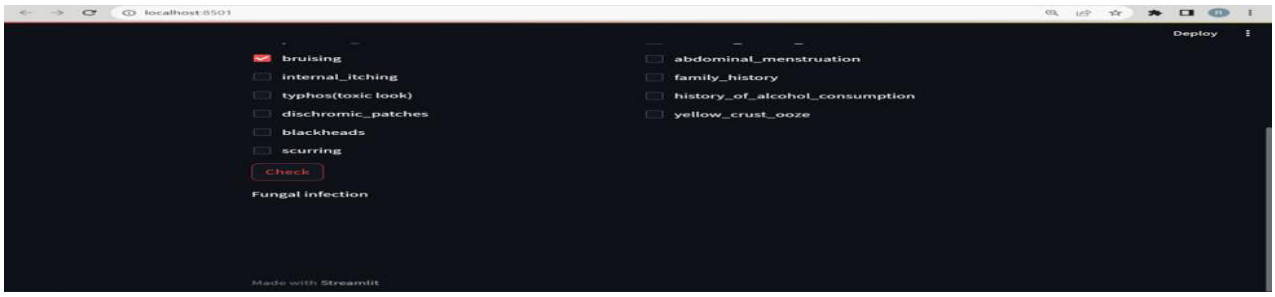


Figure 4.3 Actual operation at frontend / user end

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. CONCLUSION

In this paper, we set out to create a system which can predict disease on the basis of symptoms given to it. Such a system can decrease the rush at OPDs of hospitals and reduce the workload on medical staff. We were successful in creating such a system and used 4 different algorithms to do so. On an average we achieved accuracy of ~97%. Such a system can be largely reliable to do the job. Creating this system we also added a way to store the data entered by the user in the database which can be used in future to help in creating a better version of such a system. Our system also has an easy to use interface. It also has various visual representations of data collected and results achieved. This model can be used in future wherein it provides an accurate result and predictions. This particular model can be upgraded and enabled with medication recommendation to form a complete Health Care System for Chronic Diseases.

REFERENCES

[1]Vaibhav Kumar, Bhagwinder Singh, Dr Shilpi Sharma, Dr Dolly Sharma and Vipul Narayan. A Machine Learning approach for predicting onset and progression "Towards Early detection of chronic diseases". Journal of Pharmaceutical Negative Results Volume 13, 2022

[2]Rayan Alanazi, Hindawi Journal of Healthcare Engineering Identification and Prediction of Chronic Diseases Using Machine Learning Approach, volume 2022

[3]Shahadat Uddin , Arif Khan, Md Ekramul Hossain and Mohammad Ali Moni Comparing different supervised machine learning algorithms for disease prediction, 2019. BMC Medical informatics and decision making

[4] Md. Ehtisham Farooqui, Dr. Jameel Ahmad, A Detailed Review on Disease Prediction Models that uses Machine Learning, International Journal of Innovative Research in Computer Science & Technology (IJIRCST), Volume 8, July - 2020

[5] Theyazn H.H Aldhyani, Ali Alshebami, Yahea Alzahrani, Soft Computing Model to predict chronic diseases, Article in Journal of Information Science and Engineering · January 2020

[6] Divya Gopu, Saveetha University, Prediction of Diseases in Smart Health Care System using Machine Learning, Article in International Journal of Recent Technology and Engineering (IJRTE) · May 2021

[7] Ibrahim Mahmood Ibrahim, Adnan Mohsin Abdulazeez, The Role of Machine Learning Algorithms for Diagnosing Diseases, Journal of Applied Science and Technology Trends, vol. 02, No. 01, pp. 10-19(2021)

[8] Abraham Jacob Frandsen, Brigham Young University, Machine Learning for Disease Prediction, Brigham Young University BYU Scholars Archive, 2016-06-01

[9] Wiley, Hindawi, Journal of Food Quality, Analyzing the Performance of Machine Learning Techniques in Disease Prediction, Volume 2023, Article ID 9852606, Published 20 December 2023

[10]Vijay Kumar Korke, Vaibhav Chowdhary, Sagar M Keri, Miss Pallavi Patil, Dr. Suvarna Nandyal, Chronic Disease Prediction Using Machine Learning, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 10 Issue VIII August 2022- Available at www.ijraset.com

[11] Aditya Sharad Ahirrao, Research Publications, Multi Disease Detection and Predictions Based on Machine Learning, Article in SSRN Electronic Journal · February 2020

[12] Sunil Vilas Awari, M.Sc Scholar, Department of Computer Science, Tilak Maharashtra Vidyapeeth, Nagpur, Maharashtra, India. Diseases Prediction Model using Machine Learning Technique. International Journal of Scientific Research in Science and Technology, Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com), Volume 8, Issue 2, Page Number : 461-467, March-April-2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com