International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com
doi : https://doi.org/10.32628/IJSRCSEIT

Symptoms Based Disease Prediction Using Machine Learning

Revuru Ravi Teja[1], Mr. T. Saravanan[2]
MCA Student[1], Assistant Professor[2]
Department of Computer Applications, Madanapalle Institute of Technology and Science, Madanapalle, Andhra Pradesh, India

ARTICLEINFO

Article History:

Accepted: 01 Aug 2023
Published: 09 Aug 2023

Publication Issue
Volume 9, Issue 4
July-August-2023

Page Number
277-284

ABSTRACT

People's needs for health information are changing, and more and more of them are looking for information on diseases, diagnoses, and treatments online. Review mining could reduce the amount of time needed to implement a recommendation system for physicians and medications. However, because many users lack specialized knowledge in the topic, interpreting complicated medical vocabulary might be difficult. Furthermore, the wealth of medical information available on a variety of platforms might be daunting. Providing accurate, trustworthy, and plagiarism-free content is crucial for a successful system.

Keywords : Random Forest Algorithm, Naive Bayes, Support Vector Machine, Logistic regression

I. INTRODUCTION

Disease prediction using machine learning is a cutting-edge technique that gives accurate disease predictions based on data and symptoms reported by users. By examining the input data, the system may determine what kind of illness the person might be experiencing. Additionally, it includes suggestions for maintaining a healthy lifestyle as well as helpful health information. The healthcare industry as well as users gain from this arrangement.

It provides customers with the convenience of getting disease prognoses without having to go to a hospital or clinic. They can rapidly determine the potential ailment they may be dealing with by just entering their symptoms and other pertinent information. The system also offers advice on how to enhance their health and wellbeing.

This predictive method may be a useful tool in the healthcare sector for diagnosing patients. Medical Practitioners can obtain patient symptom information, enter it into the system, and instantly and roughly anticipate disease. This may help in delivering medical care more quickly and effectively.

It's critical to remember that privacy and data security should be carefully considered during the design and development of this system. It will be essential to ensure suitable encryption and access control mechanisms because it deals with sensitive health information in order to preserve user confidentiality and trust.

Machine learning is indeed a powerful tool that allows computers to optimize performance by learning from example data or past experiences. In the context of the medical field, machine learning has been widely applied to predict diseases using patient symptoms and historical data.

The process of applying machine learning in healthcare involves two main tracks: Training and Testing. During the training phase, the machine learning algorithm is exposed to a large dataset containing patients' symptoms, medical history, and disease outcomes. The algorithm learns patterns and relationships within the data to build a predictive model. In the testing phase, the model is evaluated on new and unseen data to assess its accuracy and performance.

Various machine learning algorithms can be used for disease prediction in healthcare, including but not limited to:

1. Linear Regression: A simple algorithm used for regression tasks to establish a linear relationship between the input features (symptoms, history) and the output (disease prediction).
2. K-Nearest Neighbors (KNN): An instance-based algorithm that predicts the disease based on the similarity of the patient's symptoms and history to those of other patients with known outcomes.
3. Decision Tree: A tree-like model that makes decisions based on a sequence of rules learned from the data.
4. Logistic Regression: A classification algorithm used when the outcome is binary (e.g., disease present or not) to estimate the probability of an event occurring.
5. AdaBoost: A boosting algorithm that combines the predictions of weak classifiers to create a strong classifier.

These algorithms can handle both structured data (e.g., numerical values) and unstructured or textual data (e.g., medical records, reports, notes). By leveraging machine learning technology and analyzing massive amounts of data, healthcare professionals can make more informed decisions for patient diagnoses and treatment options, leading to improved healthcare services.

It's important to ensure that the work done in applying machine learning to medical data is original and free from plagiarism. Ethical considerations and privacy concerns must also be taken into account when handling sensitive patient information.

Machine learning continues to advance and offers promising opportunities to revolutionize the medical field, improving patient outcomes and streamlining healthcare processes.

II. RELATED WORKS:

[1] **Monto et al:**

Based on a dataset of 3744 unvaccinated adults and adolescents who had fever and at least two other influenza symptoms, it appears that you have created a statistical model to determine whether a case has the illness or not. Out of these 3744 patients, 2470 had influenza that was determined in a laboratory.

The sensitivity (also known as recall or true positive rate) of your model is 79%. Sensitivity measures the proportion of actual positive cases (influenza cases) that the model correctly identified. In this case, it means that your model correctly identified 79% of the cases that had influenza based on the given symptoms and data.

If you have more details about the model or data, I could assist you further in evaluating the model's performance or suggest improvements if needed.

[2] **Colorful machine learning algorithms were streamlined for the effective prophecy of a habitual complaint outbreak by Chen et al**.:

It seems like you are describing the development of a new model called the "Convolutional Neural Network-grounded Multimodal Complaint Trouble Predictor" (CNN-MDRP) that was designed to overcome deficiencies in training data. The model utilizes an "idle factor model" to address the data limitation.

The use of the idle factor model suggests that the training data might have missing or incomplete information. This model likely fills in the gaps or synthesizes additional data to improve the training process. It's essential to handle data deficiencies carefully to ensure the model's effectiveness and avoid biases or erroneous predictions.

Furthermore, the CNN-MDRP model achieved a delicacy (sensitivity) of around 94.8%. Delicacy is a measure of a model's ability to correctly identify positive cases (true positives) from all the actual positive cases in the dataset. High delicacy indicates that the model can effectively detect instances of the target class.

[3] The DNN model performed more in terms of average performance and the LSTM model gave close prognostications when circumstances were large. Haq et al:

Three feature selection methods were used by the researchers: Relief, Least Absolute Loss and Selection Motorist, and Minimum Redundancy Maximum Relevance (mRMR). These techniques were probably employed to extract the dataset's most pertinent and instructive aspects.

In order to avoid overfitting and validate the selected features, the researchers made use of the K-fold cross-validation technique. The dataset is parted into K subsets (folds) utilizing a procedure called cross-approval, and the model is then prepared and tried K times, with an alternate subset filling in as the approval put down each point in time. As a result, the model's performance can be evaluated with greater confidence. Six different AI strategies were then used to group the picked highlights. Although the specific machine learning algorithms used were not specified, a variety of classifiers, such as logistic regression, decision trees, support vector machines, random forests, and neural networks, are frequently used to evaluate their efficacy in predicting the presence or absence of heart problems.

[4] Maniruzza- man et al.:

It seems like you want to classify diabetes complaints using Machine Learning algorithms and have used Logistic Regression (LR) to identify the factors associated with diabetes complaints. The overall sensitivity (also known as recall or true positive rate) of the ML-predicted system is reported to be 90.62%. Sensitivity measures the percentage of actual positive cases that the model correctly identifies.

To provide a more detailed analysis, it would be helpful to know additional information such as the dataset used, the number of features, the size of the dataset, and any preprocessing steps or feature engineering techniques applied before running the Logistic Regression model.

[5] Aiyesha Sadiya:

As a matter of fact, utilizing AI calculations to look at clinical information for sickness expectation, for example, the Choice Tree Classifier, has showed early illness distinguishing proof and better understanding consideration. These calculations may effectively handle a lot of natural and medical services information, bringing about exact gauges and redid therapy plans.

Because it is simple to understand and offers insights into the decision-making process, the Decision Tree Classifier is a well-liked algorithm for disease prediction. It builds a model that resembles a tree, with each node denoting a characteristic or attribute and each branch denoting a choice depending on that feature. The tree's leaves stand in for the anticipated results (in this case, the likelihood of contracting a specific disease).

By implementing the Decision Tree Classifier in this way, you can achieve a reliable model for disease prediction, which can be deployed to assist medical professionals in early diagnosis and personalized patient care.

III. Methodology

Proposed system:

This strategy utilizes side effects to foresee sickness. In this technique, the model is assessed utilizing a choice tree classifier. Clients are the ones who utilize this framework. The technology will be able to predict disease based on symptoms. Using AI strategies, this framework. To predict illnesses, researchers use the decision tree classifier method. We refer to this system as "AI Therapist." We integrated a few parts that recognize them and further develop their state of mind since this framework is intended for the people who stress significantly over their wellbeing. Consequently, the health awareness function known as "Disease Predictor" may be able to identify diseases based on their symptoms.

BLOCK DIAGRAM:

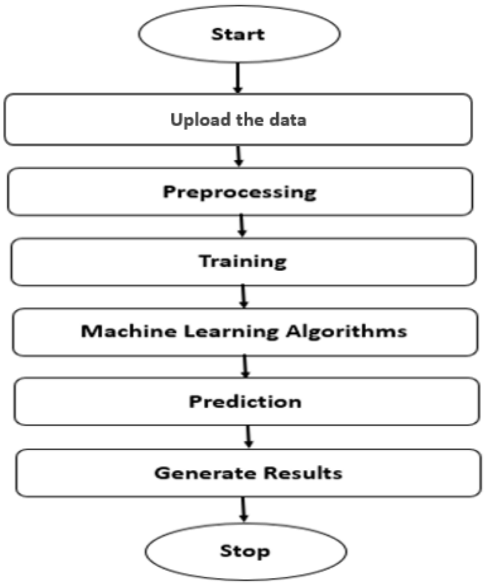


Figure 1: Block diagram

IV. Implementation

Random Forest:

A random forest is a machine learning technique used to solve classification and regression issues. It employs ensemble learning, a technique for addressing difficult problems by combining multiple classifiers.

The choice trees that can be utilized in an irregular woodland calculation are various. The irregular backwoods calculation produces a "woodland" and trains it through stowing or bootstrap conglomeration. AI frameworks' precision is improved by packing, a troupe meta-calculation.

The (random forest) algorithm chooses the outcome based on the predictions made by the decision trees. It makes expectations by averaging or averaging out the outcomes from various trees. The number of trees increases the result's accuracy.

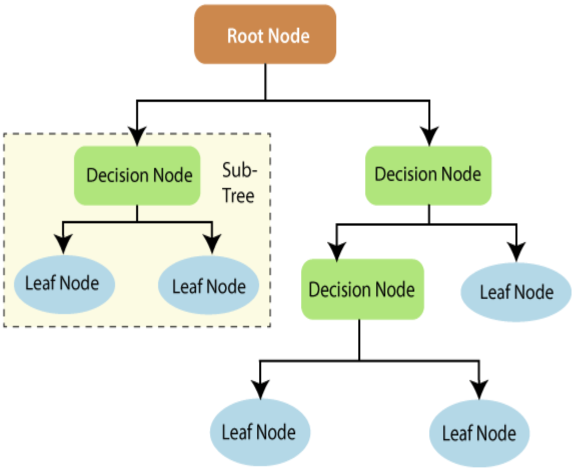
The random forest algorithm overcomes the drawbacks of the decision tree approach. Accuracy improves while dataset overfitting declines. It produces expectations without the requirement for various bundle changes, as opposed to Scikit-learn.

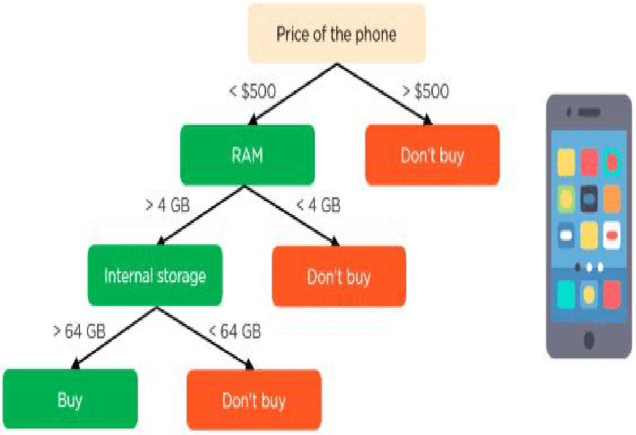
The Arbitrary Woodland Calculation enjoys the accompanying upper hands over the Choice Tree Calculation:

- It's more exact;
- It provides a useful method for dealing with missing data;
- It can make an accurate prediction without hyper-parameter adjustment;
- It settle the over fitting issue with choice trees.
- A subset of highlights is haphazardly chosen for every irregular backwoods tree at the hub's parting point.

Decision trees are the fundamental building blocks of a random forest algorithm. A choice help strategy with a tree-like design is known as a choice tree. We'll look at decision trees and how random forest methods work.

Choice hubs, leaf hubs, and a root hub are the three pieces of a choice tree. A preparation dataset is separated into branches by a choice tree approach, and those branches are then additionally partitioned. This procedure continues until it reaches a leaf node. The leaf node can no longer be split in any more ways. The characteristics that are utilized to estimate the result are shown by the choice tree's hubs. Decision nodes provide links to the leaves. The three various types of hubs that can be found in a choice tree are portrayed in the graph underneath.





The use of independent variables (features) to learn more about a target variable (class) is referred to as the "information gain idea." The entropy of the objective variable (Y) and the contingent entropy of Y (given X) are utilized to work out the data gain. The use of information gathering is necessary for the training of decision trees in this instance, as the conditional entropy reduces Y's entropy. It helps these trees feel less anxious. A high amount of vulnerability (data entropy) is taken out when there is a critical data gain. Entropy and data gain assume key parts in the parting of branches, a basic stage in the development of choice trees.

Look at this basic choice tree model. Let's assume we need to foresee regardless of whether a client would buy a cell phone. He chooses the phone based on its capabilities. A decision tree diagram can be used to illustrate this research.

The choice's root hub and choice hubs sub for the previously mentioned telephone qualities. The leaf node represents the outcome, regardless of whether a purchase is made. Cost, internal storage, and random access memory (RAM) are the primary deciding factors. The decision tree will look like the illustration below.

NAIVE BAYES:

A probabilistic machine learning model called a Naive Bayes classifier is utilized for classification tasks. The Bayes theorem serves as the foundation of the classifier.

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

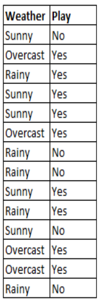
At the point when B has occurred, we might utilize the Bayes hypothesis to compute the probability that A will likewise happen. Here, An is the speculation and B is the supporting proof. Here, it is accepted that the indicators and highlights are autonomous. That is, the presence of one element doesn't change the way of behaving of another. The expression "innocent" is an outcome.

How about we utilize an outline to understand it. Below is a weather training data set and the objective variable "Play," which indicates the possibility of playing. Now, we need to classify whether or not people will play games based on the weather. How about we complete it by following the means beneath.

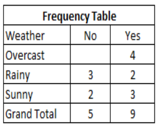
Step 1: Create a frequency table from the data set

Step 2: Find the probabilities and use them to create a Likelihood table, such as the Overcast probability of 0.29 and the Playing probability of 0.64.

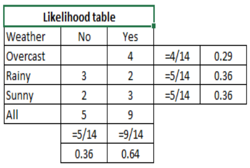
Step 3: Now, determine the posterior probability for each class by employing the Naive Bayesian equation. The class with the most noteworthy back likelihood is the result of expectation.



Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No



Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9



Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Problem: Players will play if weather is sunny. Is this statement is correct?

We can solve it using above discussed method of posterior probability.

P(Yes | Sunny) = P(Sunny | Yes) * P(Yes) / P (Sunny)

Here we have P (Sunny |Yes) = 3/9 = 0.33, P (Sunny) = 5/14 = 0.36, P (Yes) = 9/14 = 0.64

Now, P (Yes | Sunny) = 0.33 * 0.64 / 0.36 = 0.60, which has higher probability.

Naive Bayes uses a similar strategy to predict the likelihood of various classes based on various attributes. At the point when there are issues with many classes, this approach is essentially utilized in message grouping.

• Class of test informational collection forecast is speedy and straightforward. Also, it succeeds at multi-class expectation.

• An Innocent Bayes classifier performs better when the supposition of freedom is valid than different models, like strategic relapse, and requires less preparation information.

• It performs better with categorical input variables than it does with numerical input variables. For mathematical factors, ringer bend, which is major areas of strength for a, addresses the ordinary conveyance.

Applications of Naive Bayes Algorithms

Real-time Prediction: Naive Bayes is a quick classifier that is eager to learn. As a result, it might be applied to real-time prediction.

This algorithm is very widely renowned for its ability to predict many classes. Here, we can forecast the likelihood of several target variable classes.

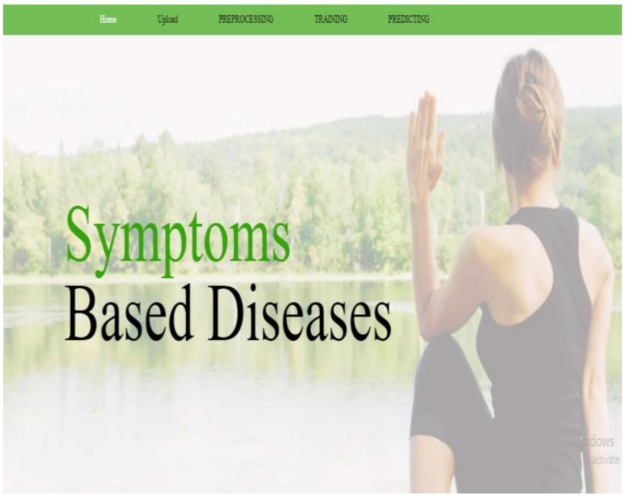
Message characterization, spam separating, and feeling examination: Gullible Bayes classifiers, which perform better in multi-class circumstances and follow the freedom basis, are more fruitful than different calculations in message order, spam sifting, and opinion examination. Along these lines, it is oftentimes utilized in Feeling Examination (in web-based entertainment investigation), to distinguish good and pessimistic shopper opinions, and Spam Sifting (to recognize spam email).

Suggestion Framework: Guileless Bayes Classifier and Cooperative Separating cooperate to make a proposal framework that channels entrepreneurial data and gauges regardless of whether a client will find a particular asset engaging.

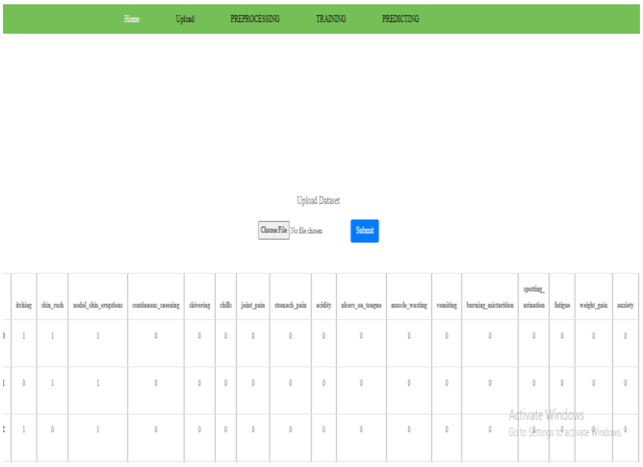
V. Results and Discussion

The following screenshots are depicted the flow and working process of project.

HOME PAGE:



UPLOAD PAGE:



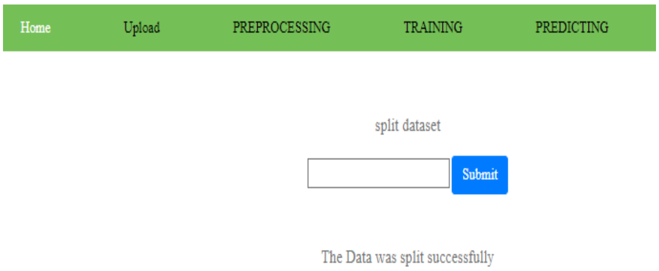
Upload Dataset

No file chosen

	itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	muscle_wasting	vomiting	burning_micturition	spotting_ urination	fatigue	weight_gain	anxiety
0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Activate Windows
Go to Settings to activate Windows.

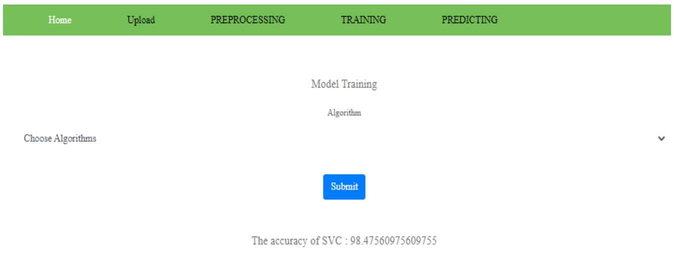
PREPORCESSING PAGE:



split dataset

The Data was split successfully

MODEL TRAINING PAGE:



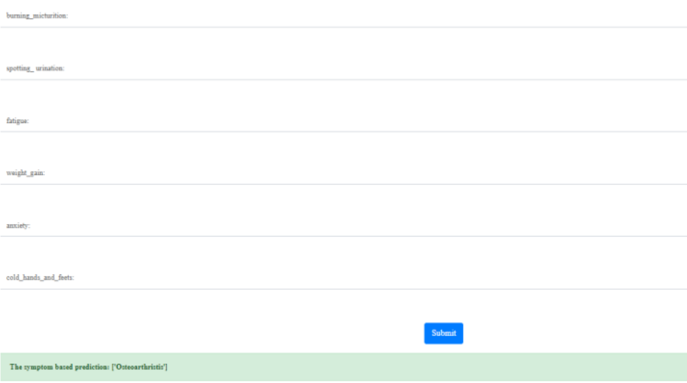
Model Training

Algorithm

Choose Algorithms

The accuracy of SVC : 98.47560975609755

PREDICTION PAGE:



burning_micturition:

spotting_ urination:

fatigue:

weight_gain:

anxiety:

cold_hands_and_feets:

The symptom based prediction: ['Osteoarthristis']

VI. Conclusion

Therefore, I've arrived to the conclusion that machine learning can be effectively used to track our health. We can maintain our health by periodically getting a free health check. The machine learning model was built, deployed using Flask (a Python web framework), and will eventually be made publicly available by turning that domain into a website. The customer merely needs to go to the appropriate page and select 5 to 8 diseases for our model to forecast the best result. The user will receive insight into their health and, if necessary, contact the relevant doctor after receiving the prediction. Anyone on our planet has the potential to become healthy.

VII. REFERENCES

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

[2]. Pingale, K., Surwase, S., Kulkarni, V., Sarage, S., & Karve, A. (2019). Disease Prediction using Machine Learning.

[3]. Aiyesha Sadiya, Differential Diagnosis of Tuberculosis and Pneumonia using Machine Learning (2019)

[4]. S. Patel and H. Patel, "Survey of data mining techniques used in healthcare domain," Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60, March, 2016.

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

[6]. Dhenakaran, K. Rajalakshmi Dr SS. "Analysis of Data mining Prediction Techniques in Healthcare Management System." International Journal of Advanced Research in Computer Science and Software Engineering 5.4 (2015).

[7]. Maniruzzaman, M., Rahman, M., Ahammed, B. and Abedin, M., 2020. Classification and

prediction of diabetes disease using machine learning paradigm. Health information science and systems, 8(1), pp.1-14.

[8]. Chen, M., Hao, Y., Hwang, K., Wang, L. and Wang, L., 2017. Disease prediction by machine learning over big data from healthcare communities. Ieee Access, 5, pp.8869-8879.

[9]. Haq, A.U., Li, J.P., Memon, M.H., Nazir, S. and Sun, R., 2018. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems, 2018.

[10]. Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., Warang, M. and Mehendale, N., 2020. Disease prediction from various

Cite this article as :

Revuru Ravi Teja, Mr. T. Saravanan, "Symptoms Based Disease Prediction Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 4, pp. 277-284, July-August-2023.