International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 8, August 2020 ||

Prediction Classifier for Heart Disease Using Machine Learning Approach

Dhanush Kumar M V [1], Raghavendra Rao B G [2], Narendra G R[3]

P.G. Student, Department of Master of Computer Applications, Sir MVIT, Bengaluru, Karnataka, India[1]

Assistant Professor, Department of Master of Computer Applications, Sir MVIT, Bengaluru, Karnataka, India[2]

Project Manager, Gobal Quest Technologies, Judicial Layout, Yelahanka, Bengaluru, Karnataka, India[3]

ABSTRACT: Heart disease is considered as one of the major causes of death throughout the world. It cannot be easily predicted by the medical practitioners as it is a difficult task that demands expertise and higher knowledge for prediction. An automated system in medical diagnosis would enhance medical efficiency and also reduce costs. We will design a system that can efficiently discover the rules to predict the risk level of patients based on the given parameters about their health. The goal is to extract hidden patterns by applying data mining techniques, which are noteworthy to heart diseases, and to predict the presence of heart disease in patients where the presence is valued on a scale. The prediction of heart disease requires a huge size of data which is too complex and massive to process and analyze by conventional techniques. Our objective is to find out the suitable machine learning technique that is computationally efficient as well as accurate for the prediction of heart disease.

KEYWORDS: Support Vector Machine; Naive Bayes; Logistic Regression; UCI; Data cleaning; Feature Scaling; Factorization;

I. INTRODUCTION

Cardiovascular disease is determined to be a critical cause of death across the globe. These disease predictions require expert knowledge and huge medical amenities such that accurate results can be generated. As performing such a very critical act one has to be very conscious and predict accurate results during the diagnosis is made, this where the thought of automation comes which could help the medical practitioners to decide before the patient undergoes further treatment. An automated system in medical diagnosis can help in enhancing the efficiency of medical diagnostics and also reduce costs.

In this paper, a system is been design in such a way that discovering certain rules to predict the efficiency of risk levels of a person with given parameters of their current health directories will help in diagnostics. The aim is to detect those hidden patterns and extract them on applying some data mining techniques, which can be significant to identify the heart disease, and ease in classifying the presence of disease in patients with a scale of presence is evaluated. As the prediction requires voluminous data which is complex by its nature and it is massive in size to process and analyze by traditional techniques.

The main purpose of the system is to design a suitable and strong machine learning techniques with highly automated efficiently and predict the heart disease accurately. Heart diseases are intensely causing death all over several countries, since from a few decades. In earlier stages detection and under some clinical supervision one can reduce the loss of death. However, in all cases, the detection may or may not be accurate and respective consultant doctor is available or not since as it requires more time, expertise, and sapience in the field. An abnormality found in normal blood flow in a human heart leads to causes several types of heart disease which are referred to be cardiovascular diseases (CVD).

II. RELATED WORK

Heart disease is considered as one of the major causes of death throughout the world. It cannot be easily predicted by the medical practitioners as it is a difficult task that demands expertise and higher knowledge for prediction. An automated system in medical diagnosis would enhance medical efficiency and also reduce costs. We will design a system that can efficiently discover the rules to predict the risk level of patients based on the given parameters about their health.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 8, August 2020 ||

The goal is to extract hidden patterns by applying data mining techniques, which are noteworthy to heart diseases, and to predict the presence of heart disease in patients where the presence is valued on a scale. The prediction of heart disease requires a huge size of data which is too complex and massive to process and analyze by conventional techniques. Our objective is to find out the suitable machine learning technique that is computationally efficient as well as accurate for the prediction of heart disease.

Data mining combines statistical analysis, machine learning, and database technology to extract hidden patterns and relationships from large databases. The implementation of work is done on Cleveland heart diseases dataset from the University of California Irvine (UCI) machine learning repository to test on different data mining techniques. Machine learning algorithms were studied for predicting the Heart Disease Prediction. Research papers on Naïve Bayes, Logistic Regression classification referred.

EXISTING SYSTEM

In the existing system, to improve the classification accuracy k-Nearest Neighbour and genetic algorithm is used on the dataset provided by users.

- Genetic search is used as a measure to decide the redundant and non-relative attributes, and these attributes must be ranked in such a way that those must contribute more towards the classification process.
- The low ranked attributes must be flushed out and an algorithm for classification is designed based on evaluated attributes.
- Finally, obtained classifiers are trained into a model for classifying the data set as a person is healthy or sick.

Disadvantages

- In the existing classification model, most of the approaches tend to perform low on extreme data sets that are imbalanced. However, it has two issues namely data source issues and performance issues.
- Hence, another alternative model is required which as to be build to overcome such issues. To do this one can create an efficient algorithm that can handle huge and tedious health care information that doesn't require much manual processing.
- Finally, Health care data sets are tedious and imbalance in nature, so there may be inaccurate predictions made with the existing model.

PROPOSED SYSTEM

Proposed system Approach Combines SVM, Naïve Bayes, Logistic Regression classification.

Advantages

- To get knowledge of heart disease risk using some data mining model and machine learning approaches.
- The performance of the classification algorithm can be raised by considering more attributes.
- For the removal of irrelevant attributes fro data sets can increase in performance of classifier by using attribute selection method and logistic regression shows better performance in prediction.

III. METHODOLOGY

1 SUPPORT VECTOR MACHINE

A Support Vector Machine (SVM) is a supervised learning method that analyzes data used for classification and regression analysis. It is given a set of training data, marked as belonging to either one of two categories, an SVM training algorithm then builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. The points are separated based on hyperplanes that separate them. When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find the natural clustering of the data to groups, and then map new data to these formed groups. In the project, we have used this algorithm to classify the patients according to the risk posed to them based on the parameters provided.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 8, August 2020 ||

2 NAÏVE-BAYES ALGORITHM

In Machine Learning, we are often interested in selecting the best hypothesis (h) given data (d).In a classification problem, our hypothesis (h) may be the class to assign for a new data instance (d).One of the easiest ways of selecting the most probable hypothesis given the data that we have that we can use as our prior knowledge about the problem. Bayes' Theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge.

Bayes' Theorem is stated as:

P(h|d) = (P(d|h) * P(h)) / P(d)……..Eqn 1

Where

•P(h|d) is the probability of hypothesis h given the data d. This is called the posterior probability.

•P(d|h) is the probability of data d given that the hypothesis h was true.

•P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.

•P(d) is the probability of the data (regardless of the hypothesis).

You can see that we are interested in calculating the posterior probability of P(h|d) from the prior probability p(h) with P(D) and P(d|h).

After calculating the posterior probability for several different hypotheses, you can select the hypothesis with the highest probability. This is the maximum probable hypothesis and may formally be called the maximum a posteriori (MAP) hypothesis.

This can be written as:

MAP(h) = max(P(h|d))

or

MAP(h) = max((P(d|h) * P(h)) / P(d))

or

MAP(h) = max(P(d|h) * P(h))…….Eqn 2

The P(d) is a normalizing term which allows us to calculate the probability. We can drop it when we are interested in the most probable hypothesis as it is constant and only used to normalize.

Back to classification, if we have an even number of instances in each class in our training data, then the probability of each class (e.g. P(h)) will be equal. Again, this would be a constant term in our equation and we could drop it so that we end up with:

MAP(h) = max(P(d|h))

IV. EXPERIMENTAL RESULTS

The dataset we obtained was not completely accurate and error-free. Hence we first carried out the following operations on it:

- **Data Cleaning**

NA values in the dataset were the major setback for us as it was reducing the accuracy of the prediction profoundly so, we removed the fields which had NA value. We substituted it with the mean value of the column. This way, we removed all the NA values in the data set.

- **Feature Scaling**

Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without feature scaling. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be scaled so that each feature contributes approximately proportionately to the final distance. So we scaled the various fields to get them closer in terms of values.

Eg. Age had just two values i.e. 0,1 and cholesterol had high values like 100. So, to get them closer to each other, they were scaled to a value between -1, and 1.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 8, August 2020 ||

- **Factorization**

In this section, we assigned a meaning to the values so that the algorithm doesn't confuse them. For example, assigning meaning to 0 and 1 in the age section so that the algorithm doesn't consider 1 as greater than 0 in that section.

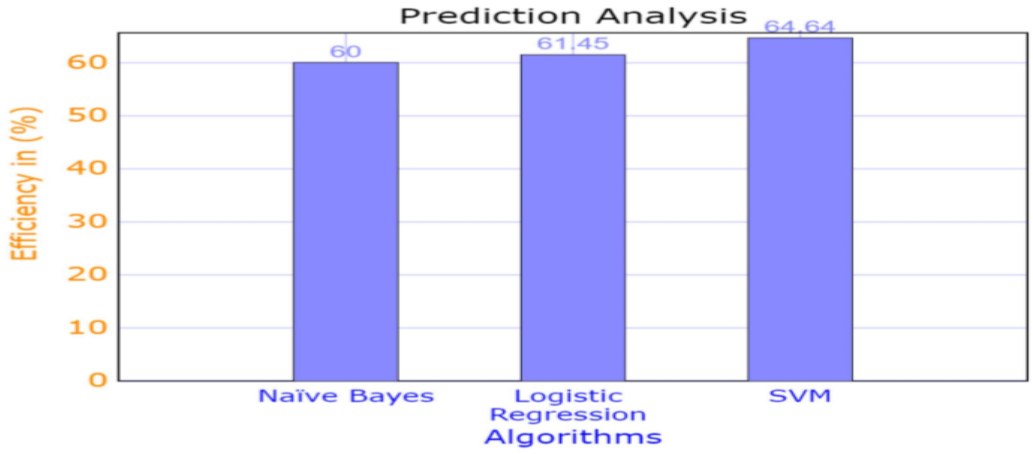

Fig 4.1 A Prediction analysis graph based on experimental results.

It was observed that: Naïve Bayes had 60% accuracy, Logistic Regression had 61.45% and SVM had 64.4%. Hence SVM was selected as the most efficient algorithm for the web application.

V.RESULTS AND DISCUSSION

- The accuracy of the system arrives from the collected attributes within the datasets. An increase in the accuracy of data set selection helps to develop a better automation system.
- The best algorithm can be chosen which can obtain better accuracy and legitimate in discovering the prediction model.
- A highly efficient algorithm that provides more accurate results is chosen in Disease prediction.

VI.CONCLUSION

In this application, the prediction of cardio disease involved with the various analyzing process with distinct Machine Learning algorithms developed in Python by providing an immediate mechanism for the user to use a build model in the Machine Learning Approach. Further Work includes advanced methods in applying these algorithms which can lands up with even better performance with high intensive data sets.

At first, the three algorithms were implemented. Datasets were trained for all the algorithms individually. After this, all of them were tested. The most efficient algorithm was to be selected based on various criteria. We found out that the SVM algorithm was the most efficient out of the three with an accuracy of 64.4%. Naïve Bayes and Logistic Regression had an accuracy of 60% and 61.4% respectively. Thus the SVM algorithm was further implemented using a better user interface in form of a System application. For this, the Tkinter GUI in Python was used. This would help the end users get a preliminary prediction about the condition of their hearts. Since heart diseases are a major killer in India and throughout the world, the application of a promising technology like machine learning to the initial prediction of heart diseases will have a profound impact on society. This will tell the user if they are at a risk and if they need to visit the doctor. This will help reduce the death rate due to heart attacks. Hence by using the above approach successful analysis of heart diseases of the individual was performed and the result was obtained which predicted the risk of heart disease based on the parameters provided by the user.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 8, August 2020 ||

REFERENCES

[1] Ricardo Buettner , Marc Schunter, An Efficient machine learning-based detection of heart disease, IEEE International Conference on E-health Networking, Application & Services (HealthCom), 2019

[2] Sneha Borkar, M. N. Annadate, An Supervised Machine Learning Algorithm for the Detection of Cardiac Disorders, Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018 .

[3] Marjia Sultana, Afrin Haider and Mohammad Shorif Uddin, Analysis of Data Mining Techniques for Heart Disease Prediction, 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 2016.

[4] Soodeh Nikan, Femida Gwadry-Sridhar, and Michael Bauer, Machine Learning Application to Predict the Risk of Coronary Artery Atherosclerosis, International Conference on Computational Science and Computational Intelligence (CSCI),2016.

[5] Sonam Nikhar, A.M. Karandikar "Prediction of Heart Disease Using Machine Learning Algorithms" in International Journal of Advanced Engineering, Management and Science (IJAEMS) June-2016 vol-2

[6] Deeanna Kelley "Heart Disease: Causes, Prevention, and Current Research" in JCCC Honors Journal

[7] Nabil Alshurafa, Costas Sideris, Mohammad Pourhomayoun, Haik Kalantarian, Majid Sarrafzadeh "Remote Health Monitoring Outcome Success Prediction using Baseline and First Month Intervention Data" in IEEE Journal of Biomedical and Health Informatics .