International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

Heart Disease Prediction with PCA and SRP

Bandari Sai Santosh, Dharma Sahith Reddy, M Sai Vardhan, Dr. Shaik Subhani

Abstract: ***In this expeditiously modern world, it all depends on how effectively the data can be maintained and utilized for a suitable purpose. Handling such a large quantity of dynamic data is not at all an easy task. On the contrary, we can use classification techniques, which is purely for building the relationships among huge databases by easily predicting the outcomes by considering the type of relationship. This kind of techniques plays an essential role in every aspect of science and engineering, for example, human services, education, web-based businesses. In the Health maintenance industry, all the data mining techniques are most part utilized for malady prediction. The main goal in this work attempts deeply to anticipate the occurrence of coronary disease with reduced attributes in the dataset. In this case, 14 characteristics are associated with anticipating coronary illness. Following the process, Five classifiers like Classification by clustering, Support Vector Machine, Naive Bayes, Random Forest, Decision Tree are utilized to anticipate the diagnostic influence of heart disease once after reducing the range of characteristics.***

Keywords: Clustering, Decision Tree, Naive Bayes, Principal Component Analysis, Random Forest, Sparse Random Projection, Support Vector Machine.

I. INTRODUCTION

The coronary disease is a general term for a combination of diseases, conditions and disarranges that greatly impact the heart, furthermore, the veins. Side effects of coronary illness differ contingent upon the particular sort of coronary illness.
Inherent coronary illness alludes to an issue with the heart's structure and capacity because of irregular development of the heart even before the initial stage. A phenomenon of failure of the heart known as Congestive heart failure, which is the time at which the heart does not siphon adequate blood to different parts in the body. Coronary illness is a term that alludes to harm to the heart that happens on the grounds that its blood supply is diminished, it prompts the fatty deposits to develop on the linings of the veins that give the heart muscles with blood, bringing about them narrowing. This paper distinguishes the hazard dependencies for the diverse sorts of heart maladies. Nowadays, all the hospitals utilize a type of emergency clinic data structure to handle all the aspects related to healthcare or understanding data.

Manuscript published on 30 April 2019.
***** Correspondence Author (s)
Bandari Sai Santosh*, Student, Dept. of Information Technology, Sreenidhi Institute of Science and Technology(Autonomous), Hyderabad
Dharma Sahith Reddy, Student, Dept. of Information Technology, Sreenidhi Institute of Science and Technology(Autonomous), Hyderabad
M Sai Vardhan, Student, Dept. of Information Technology, Sreenidhi Institute of Science and Technology(Autonomous), Hyderabad
Dr. Shaik Subhani, Associate Professor, Dept. of Information Technology, Sreenidhi Institute of Science and Technology(Autonomous), Hyderabad

These systems ordinarily produce enormous amounts of data which appear as numerical, content, diagrams, and charts. Up to a greater extent, There is concealed data in this information that is legitimately more than required. Assuring both the Security and Authenticity of the health data.
Heart failure is the most common therapeutic issue which happens particularly in older patients in view of their eating mundane, non-steroidal calming drugs and which will lead indeed, even towards death. One such case happened regarding coronary illnesses is Cardiovascular malady. Hence it is exceptionally basic to foresee such illnesses through reasonable symptoms. There are different kinds of calculations which are available for the forecast of heart maladies which are Support Vector Machine, Naive Bayes, Random Forest and so forth. Lamentably all doctors don't have mastery in each strength and in addition, there is a lack of professional skill at specific spots. Subsequently, a customized therapeutic analysis system would almost certainly be exceedingly helpful for bringing the proficient along with precise outcome. Fitting Computerised data and decision supportive networks may help in accomplishing clinical examinations at a diminished expense. World Health Organization in the year 2003 revealed that around 30% of total worldwide deaths are due to Cardiovascular Disease. Before the current year's over, Cardio-Vascular Disease is relied upon to be the main source for deaths in developing nations because of progress in way of life, work culture and nourishment propensities. Thus, increasingly cautious and productive strategies for heart ailments and occasional examination are of high significance.

II. RELATED WORKS

For assessing the performance and efficiency, numerous experiments[1] are being carried out using the classification mining techniques. From the results observed till now shows us that Naive Bayesian Classifier can outperform very often and sometimes the techniques of Decision Tree. Adding upon that, an optimization system using reduction algorithms like Principal Component Analysis is also being implemented in order to diminish the number of characteristics without compromising on accuracy for improvising the coronary disease. Our contribution differs by diminishing the attributes from fourteen to six which are essentially needed for the diagnosis and had the capability to accomplish a similar efficiency and accuracy. Furthermore, Our approach towards the diagnosis is unique as we performed different algorithms on the attributes to acquire the maximum accuracy and least time, plotting it on charts for clear understanding through graphical representation. The remaining of the paper are sorted out in the following manner. Section 3 explains proposed work, section 4 explains implementation results and section 5 explains conclusion.

III. PROPOSED WORK

A. Problem Definition

Heart disease prediction utilizing data mining is a champion among the most interesting and testing undertakings. The lack of experts and high erroneous cases has required the need to build up a rapid and productive framework. The primary goal of this work is to recognize the patterns or highlights from the therapeutic data utilizing the corresponding classifier model. The properties that are progressively significant to coronary illness prediction can be seen. This can assist the professionals in understanding the main root cause of disease profoundly.

B. Experimental Data Set

14 attributes from therapeutic Data [2] have been considered to implement our classifier model. These fourteen characteristics are recorded in Table 1. For effortlessness, These attributes based on categories were utilized for all the classifier models. The total number of attributes in the data have been diminished to six using Principal Component Analysis and Sparse Random Projection. The data set which is reduced is passed on to the five models.

Table 1: Attributes of Heart Disease Data Set

S.no	Characteristics
1	Age
2	Gender
3	Type of Chest pain
4	blood pressure
5	Cholesterol count
6	blood sugar
7	Electrocardiographic results
8	Thalach
9	Induced Angina
10	Old peak
11	Slope
12	Thal
13	CA
14	Concept class

C. Naive Bayes

One of the Classification algorithms which is known as Probability Classifier is based on Bayes theorem. Naïve Bayes is a statistical classifier which describes not even a single dependency between characteristics. This classifier is essential, gainful and is having a better than average execution. Once in a while, it beats present day classifiers notwithstanding when the supposition of free indicators is far. This preferred standpoint is especially enunciated when the quantity of pointers is huge. A champion among the most basic obstacles of Naive Bayes is that it has a strong component of freedom suspicions. Perceptions frequently demonstrate that Naïve Bayes performs reliably after the decrease of various traits.

As indicated by the Bayesian hypothesis,

$P(A|B) = P(A)*P(B/A)/P(B)$,

Where $P(B|A)=P(A \cap B)/P(A)$

In light of the above condition, Bayesian classifier finds out the restrictive likelihood of an event having a spot with each class and reliant on such contingent likelihood information, the occurrence is delegated the class with the most shocking contingent likelihood.

D. Decision Tree

Decision trees are incredible and well-known apparatuses for prediction and classification. They have some rules, which can be comprehended by people and utilized in a learning framework, for example, the database. is a well-known classifier which is basic and simple to execute. Subsequently, it is increasingly suitable for exploratory knowledge disclosure. Regardless, it still experiences reiteration and replication. Along these lines, vital advances need to be taken to deal with reiteration and replication. The execution of decision trees can be upgraded with reasonable attribute selection. Correct determination of attributes segment data set into distinct classes. We use Decision tree for classification in our work. From the observations, we have found out that Decision Tree has the capabilities to outperform other classifiers often but takes more time to construct the model.

E. Classification via clustering

The Phenomenon in which segregation of sets of data into a set of sub-classes with some meaning is called Clustering and the resultant sets are termed as called clusters. Clustering is a classification technique which is purely unsupervised and has no predefined classes. This strategy might be utilized as a preprocessing step before sustaining the data to the classifying model. All the attributes must be subjected to normalization before clustering to avoid dominance of high-value attributes over low-value attributes. Further, classification is performed depending on clustering. From the observations, it is found that Classification via clustering performs inadequately even after reduction of various attributes when contrasted with the other techniques.

F. Support Vector Machine

Support vector machine is another fundamental calculation that each machine learning expert ought to have in his/her arsenal store. Support vector machine is very favored by numerous individuals as it produces huge precision with less calculation control. Support Vector Machine, otherwise called SVM can also be used for numerous reasons, for example, relapse and grouping errands. The objective of this algorithmic calculation is to discover a hyperplane in the space that particularly orders the information focuses. Support vectors are some particular information indicates that are very near to the hyperplane and impact the position and introduction of the hyperplane. Utilizing these support vectors, we endeavor to expand the edge of the classifier. Erasing the support vectors will change the situation of the hyperplane.

G. Random Forest

H. Important tasks such as Classification and Regression can be applied easily by using an algorithm like Random Forest. As the name proposes, this algorithm makes a forest with various trees. When all is said in done, the more trees in the forest the more strong the forest resembles.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Along these lines in this classifier, In a single forest, the more the number of trees will equally determine the high precision results. It is very much similar to creating multiple decision trees to convert into a forest. The specialty of the random forest algorithm is that it can handle missing values and can model categorical data. It doesn't overfit the model even if we have a number of trees.

IV. IMPLEMENTATION RESULTS

To improve the prediction of classifiers, algorithms like Principal Component Analysis and Sparse Random Projection are incorporated, this resulted in 6 attributes which are essentially more towards the diagnosis of the heart disease. The five classifiers such as Naive Bayes, Decision Tree, Classification via Clustering, Support Vector Machine, Random Forest are implemented for the diagnosis of patients with heart diseases. Reduced 6 attributes were fed to all the specified classifiers. Results are listed in Table 2. From the observations, we can exhibit that the Effective algorithm like Random Forest technique outperforms all other four data mining techniques but with high construction time. Decision Tree technique can provide high accuracy with considerably less construction time. On the Contrary, Before and post the diminishing the attributes, Naive Bayes Classifier can perform consistently with the same time for construction. Out of all, Classification via Clustering performs very poor with the least accuracy and high rate of construction time when compared to other techniques. Graphical Representation of the implementation results is shown in Figure 1 & 2.

Table 2: Comparison of five classifiers using Principal Component Analysis

Techniques	Accuracy	Construction Time
Random Forest	56.58%	0.470 s
Decision Tree	56.58%	0.000 s
Naive Bayes	53.95%	0.000 s
Support Vector Machine	53.95%	0.007 s
Classification via Clustering	47.37%	0.046 s

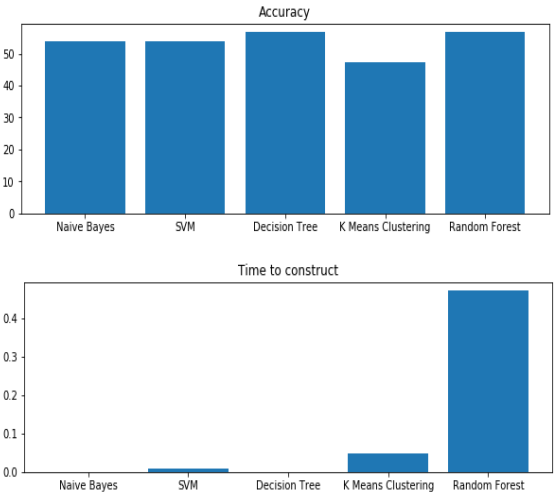

Figure 1: Graphical representation of the five classifiers using Principal Component Analysis

Table 3: Comparative Table of the five classifiers using Sparse Random Projection

Techniques	Accuracy	Construction Time
Random Forest	53.95%	0.453 s
Naive Bayes	53.95%	0.003 s
Support Vector Machine	51.32%	0.014 s
Decision Tree	51.32%	0.000 s
Classification via Clustering	27.63%	0.046 s

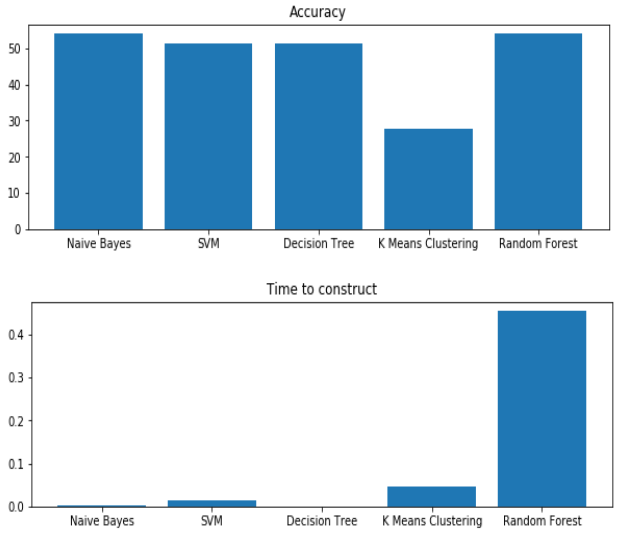

Figure 2: Graphical representation of the five classifiers using Sparse Random Projection

V. CONCLUSION AND FUTURE WORK

In this paper, we first reduce the attributes from 14 to 6, which makes a significant impact on the diagnosis by using reduction algorithms such as Principal Component Analysis and Sparse Random Projection. Once the reduction of attributes is carried out, We then compare all the five classification techniques for the prediction of a diagnosis of coronary disease patients. The Classifier techniques which we utilized are Naive Bayes, Decision Tree, Classification via clustering, Support Vector Machine and Random Forest. From the observations, we can exhibit that the Random Forest data mining technique outperforms all other four techniques but with high construction time. Decision Tree technique can perform with high accuracy with significantly less construction time. On the Contrary, Before and post the diminishing the attributes, Naive Bayes Classifier can perform consistently with the same time for construction. Out of all, Classification via Clustering performs very poor with the least accuracy and high rate of construction time when compared to other techniques. Irregularities and missing values were settled before model construction, yet progressively, that isn't the situation. We plan to work for assessing the intensity of the heart diseases and keep a check on it.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

ACKNOWLEDGEMENT

We would like to thank our Research Guide Dr. Subhani Shaik, Associate Professor in Department of Information Technology, SNIST, Hyderabad for their continue support and valuable suggestions throughout carried this work. Authors are also grateful to the reviewer for the renovation of manuscript. We would also like to thank the Department of information Technology providing us with the facility for carrying out the simulations.

REFERENCES

1. B.Venkatalakshmi, M.V Shivsankar:" Heart Disease Diagnosis Using Predictive Data mining", TIFAC-CORE, Pervasive Computing Technologies, Velammal Engineering College, Chennai, India.
2. UCI Machine learning Repository from https://archive.ics.uci.edu/ml/datasets/Heart+Disease.
3. Scikit package notes from https://www.geeksforgeeks.org/learning-model-building-scikit-learn-python-machine-learning-library/
4. Online source: https://scikitlearn.org/stable/modules/generated/sklearn.random_projection.SparseRandomProjection.html
5. Online source: https://www.programcreek.com/python/example/72587/matplotlib.ticker
6. Online source: https://matplotlib.org/api/_as_gen/matplotlib.pyplot.bar.html