IJCRT.ORG ISSN : 2320-2882
INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

HEART DISEASE PREDICTION USING DIFFERENT MACHINE LEARNING TECHNIQUES

[1]Srijan Kumar Singh, [2]Siddharth Dhawan, [3]Tanmay Chordia, [4]Bindu Garg
[1]Student, [2]Student, [3]Student, [4]Professor
[1]Bharati Vidyapeeth Deemed to be University College of Engineering, Pune India,
[2]Bharati Vidyapeeth Deemed to be University College of Engineering, Pune India,
[3]Bharati Vidyapeeth Deemed to be University College of Engineering, Pune India,
[4]Bharati Vidyapeeth Deemed to be University College of Engineering, Pune India

Abstract: Heart-related diseases or cardiovascular diseases (CVDs) have been the main cause of a large number of deaths in the world over the last few decades and have emerged as the most life-threatening illness, not just in India but throughout the world. Therefore, a robust, effective and viable system is required to diagnose these diseases in time for proper medical care. Machine Learning algorithms and methods have been applied to different clinical datasets to mechanize the examination of enormous and complex information. In recent years, several researchers have used many techniques of machine learning to help the health care industry and the experts identify heart related diseases. This paper presents an overview of different models dependent on such calculations and procedures and analyze their performance. Models based on supervised learning algorithms such as, Gaussian Naïve Bayes, Decision Trees (DT), Support Vector Machines (SVM), Linear SVC, Random Forest (RF) are quite popular among the researchers.

I. INTRODUCTION

Heart is a central organ of the human body. Blood is pumped to every part of our anatomy. If it does not function properly, then the brain and various other organs will cease to operate, and the individual will die within a few minutes. Transition in lifestyle, work-related stress, and poor eating habits lead to elevated levels of multiple heart-related illnesses.

Cardiac disorders have emerged as one of the most common causes of death worldwide. Heart-related illnesses are responsible for claiming 17.7 million lives per year, 31 percent of all global deaths, according to the World Health Organization. Heart related diseases have also been the leading cause of mortality in India. According to the 2016 Global Burden of Disease Survey, published September 15, 2017, heart disease has killed 1.7 million Indians in 2016. Heart-related diseases increase health-care costs and can reduce an individual's productivity. The World Health Organization (WHO) figures indicate that from 2005 to 2015, India lost up to $237 billion due to infectious or cardiovascular diseases [2]. Therefore, it is very important to predict heart-related diseases correctly and practicably.

Health organizations around the globe are collecting data on various health related issues. Using various machine learning techniques these data can be used to obtain useful insights. But the data collected is very large, and the data can be very noisy at times. Such datasets, which are too complex for the comprehension of human minds, can be easily examined using various techniques of machine-learning. Therefore, in recent times, these algorithms have become very useful for reliably predicting the existence or absence of heart-related diseases.

II. DIMENSIONALITY REDUCTION

The method of reducing the number of random variables under consideration by acquiring a collection of key variables is in math, machine learning and knowledge theory. Approaches can be broken down into collection of features and extraction of features. The data required for a function or a problem that consist of many attributes or parameters, but not all of these attributes may influence performance equally. A large number of attributes, or features, that affect the complexity of the computation and may even result in over-fitting that results in poor results.

Feature Extraction

In machine learning and analytics, selection of features, also known as selection of component, selection of attributes or selection of component subsets, is the method of choosing a subset of specific features (variables, predictors) for use in model construction. Why is it necessary: -

- Simplifying the templates to make them easy for researchers / users to view.
- To reduce the complexity or training time.
- To avoid the curse of dimensionality phrase.
- Better generalization by minimized overfitting.

Feature Selection

In machine learning, pattern recognition and image processing, the extraction of features starts with an initial set of measured data and constructs derived values (faces) that are intended and non-redundant, enabling the corresponding learning and generalization phases, and in certain cases contributing to better human interpretations. Extraction of the function corresponds to reduction of the dimensionality. Extraction of features means to reduce the amount of resources used to define a large range of details.

III. Algorithms and Techniques Used

A. Decision tree

Decision tree is the most effective and common classification and prediction method. A Decision tree is a graph-like flowchart in which each internal node signifies a check on an element, each branch reflects a check outcome, and each leaf node (terminal node) carries a class name. This algorithm splits the population into two or more identical sets depending on the most important predictors. The Decision Tree algorithm measures each and every attribute's entropy first. The data set is then broken with the help of variables or predictors for maximum gain of knowledge or minimal entropy. Such two measures are achieved recursively for the rest of the attributes.

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

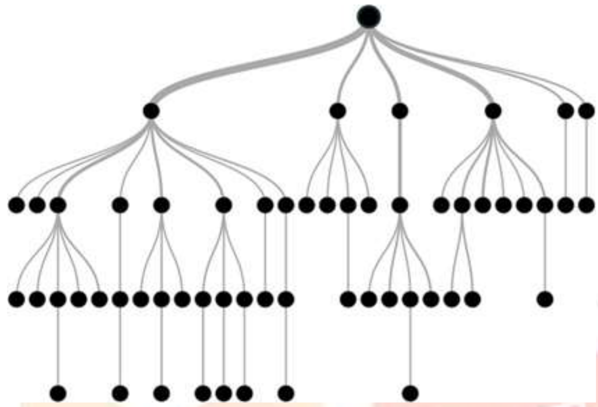

fig. 1: decision tree

Using this technique to classify our problem we got the accuracy of 77.05% using all the features in Heart Disease UCI dataset.

B. Random Forest

Random Forest is a popular supervised machine learning algorithm. This method may be used for regression and classification tasks but usually works best in classification tasks. Random forest classifier generates a collection of decision trees from randomly identified training group sub-set. It then aggregates the votes from various decision trees for the final class of the test item to be determined. Basic parameters for the Random Forest Classifier may be total number of trees to produce and parameters relevant to the decision tree such as minimum split, split criterion etc.

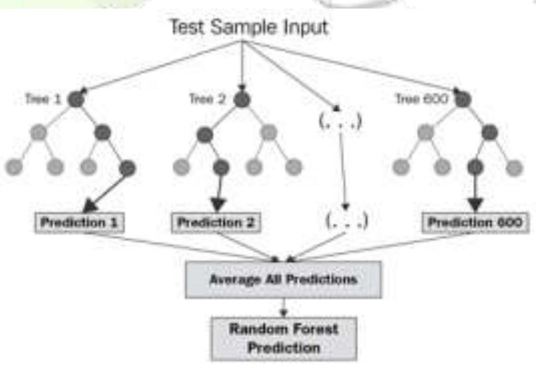

fig. 2: random forest prediction

Using this technique to classify our problem we got the accuracy of 80.33% using all the features in Heart Disease UCI dataset.

C. Support Vector Machine

"Support Vector Machine" (SVM) is a supervised algorithm for machine learning that can be used both for classification and regression problems. It is primarily used in classification issues, though. In the SVM algorithm, each data object is plotted as a point in n-dimensional space (where n is the number of features you have) with each field being the value of a particular coordinate.

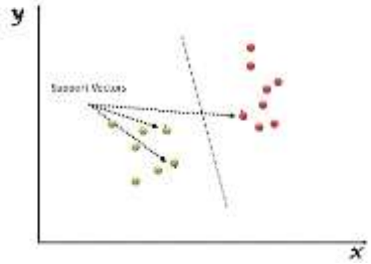

fig. 3: support vector machine

1. Gaussian radial basis function (RBF)

It is a general-purpose kernel, used when there is no prior knowledge about the data.

$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma\|\mathbf{x_i} - \mathbf{x_j}\|^2)$$

Using this technique to classify our problem we got the accuracy of 68.85% using all the features in Heart Disease UCI dataset.

2. Linear SVM

Linear SVM is the latest incredibly quick machine learning (data mining) algorithm to solve multiclass classification problems from ultra-large data sets that implements an original patented variant of a cutting plane algorithm to construct a linear vector support system.

Using this technique to classify our problem we got the accuracy of 81.97% using all the features in Heart Disease UCI dataset.

D. Gaussian Naive Bayes

Gaussian Naive Bayes methods are a series of supervised learning algorithms focused on the implementation of the theorem of Bayes with the "naive" presumption of conditional independence for each pair of characteristics given the class variable value. The theorem of Bayes notes the following relation, provided the variable class and dependent function vector x1 through x2.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Using this technique to classify our problem we got the accuracy of 85.25% using all the features in Heart Disease UCI dataset.

E. Logistic Regression

Logistic Regression is used where the target variable is categorical. Logistic regression name comes from the function used at computational level to compute probabilities; the logistic function also called sigmoid function.

$$P = \frac{1}{1 + e^{-(a+bX)}}$$

Statisticians have been established to explain the features of population growth in biodiversity, increasing increasingly and optimizing environmental efficiency. This is an S-shaped curve that can take any number evaluated and map it to a value between 0 and 1.

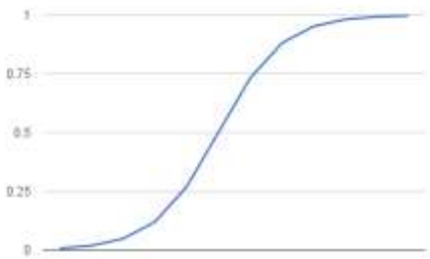

fig. 4: logistic regression

Using this technique to classify our problem we got the accuracy of 85.25% using all the features in Heart Disease UCI dataset.

IV. Conclusion

Based on the above analysis, it can be concluded that predicting cardiovascular diseases or heart-related diseases offer substantial scope for machine learning algorithms. Gaussian Naïve Bayes and Logistic Regression, performed exceptionally well with 85.25% of accuracy whereas Decision trees performed very poorly with mere accuracy of 77.07%. Random Forest (80.33%) and SVM (81.97%) models have performed moderately well, as they overcome the problem of overfitting by using multiple algorithms (Random Forest multiple decision trees). Models based on the Naïve Bayes classifier were very efficient in computational terms. Systems based on machine learning algorithms and techniques were very accurate in heart-related predictions.

References

[1] Ramadoss and Shah B et al."A. Responding to the threat of chronic diseases in India". Lancet. 2005; 366:1744–1749. doi: 10.1016/S0140-6736(05)67343-6.

[2] Global Atlas on Cardiovascular Disease Prevention and Control.Geneva, Switzerland: World Health Organization, 2011

[3] Dhomse Kanchan B and Mahale Kishor M. et al. "Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis", 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication.

[4] R.Kavitha and E.Kannan et al. "An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature Selection Technique in Data Mining ", 2016

[5] Shan Xu ,Tiangang Zhu, Zhen Zang, Daoxian Wang, Junfeng Hu and Xiaohui Duan et al. "Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework", 2017 IEEE 2nd International Conference on Big Data Analysis.

[6] Manpreet Singh, Levi Monteiro Martins, Patrick Joanis and VijayK. Mago et al. " Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map", 978-1-5090-0626-7/16/$31.00 c 2016 IEEE.

[7] Kanika Pahwa and Ravinder Kumar et al. "Prediction of Heart Disease Using Hybrid Technique For Selecting Features", 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON).

[8] Seyedamin Pouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez et al. " A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease", 22nd IEEE Symposium on Computers and Communication (ISCC 2017):Workshops - ICTS4eHealth 2017

[9] Hanen Bouali and Jalel Akaichi et al. "Comparative study of Different classification techniques, heart Diseases use Case.", 2014 13th International Conference on Machine Learning and Applications

[10] Seyedamin Pouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez et al. " A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease", 22nd IEEE Symposium on Computers and Communication (ISCC 2017):Workshops - ICTS4eHealth 2017

[11] Houda Mezrigui, Foued Theljani and Kaouther Laabidi et al. "Decision Support System for Medical Diagnosis Using a Kernel- Based Approach", ICCAD'17, Hammamet - Tunisia, January 19- 21, 2017.

[12] Dr.(Mrs).D.Pugazhenthi, Quaid-E-Millath and Meenakshi et al. "Detection Of Ischemic Heart Diseases From Medical Images " 2016 International Conference on Micro-Electronics and Telecommunication Engineering.

[13] J. Hodges et al. "Discriminatory analysis, nonparametric discrimination: Consistency properties," 1981.

[14] S.Rajathi and Dr.G.Radhamani et al. "Prediction and Analysis of Rheumatic Heart Disease using kNN Classification with ACO ", 2016.

[15] Puneet Bansal and Ridhi Saini et al. "Classification of heart diseases from ECG signals using wavelet transform and kNN classifier", International Conference on Computing, Communication and Automation (ICCCA2015).

[16] Simge EKIZ and Pakize Erdogmus et al. "Comparitive Study of heart Disease Classification", 978-1-5386-0440-3/17/$31.00 ©2017 IEEE.

[17] Renu Chauhan, Pinki Bajaj, Kavita Choudhary and Yogita Gigras et al. "Framework to Predict Health Diseases Using AttributeSelection Mechanism", 2015 2nd International Conference on Computing for Sustainable Global Development (INDIA Com).

[18] M.A.JABBAR , B.L Deekshatulu and Priti Chndra et al. "Alternating decision trees for early diagnosis of heart disease", Proceedings of International Conference on Circuits, Communication, Control and Computing (I4C 2014).

[19] Amir Hussain, Peipei Yang, Mufti Mahmud and Jan Karasek et al. "A Novel Cardiovascular Decision Support Framework for effective clinical Risk Assessment.", 978-1-4799-4527- 6/14/$31.00 ©2014 IEEE.

[20] Quazi Abidur Rahman, Larisa G. Tereshchenko, Matthew Kongkatong, Theodore Abraham, M. Roselle Abraham, and Hagit Shatkay et al. "Utilizing ECG-based Heartbeat Classification for Hypertrophic Cardiomyopathy Identification", DOI 10.1109/TNB.2015.2426213, IEEE Transactions on Nano Bioscience TNB-00035-2015.

[21] Ahmad Shahin, Walid Moudani, Fadi Chakik, Mohamad Khalil et al. "Data Mining in Healthcare Information Systems: Case Studies in Northern Lebanon", ISBN: 978-1-4799-3166-8 ©2014 IEEE.

[22] Tahira Mahboob, Rida Irfan and Bazelah Ghaffar et al. "Evaluating Ensemble Prediction of Coronary Heart Disease using Receiver Operating Characteristics", 978-1-5090-4815- 1/17/$31.00 ©2017 IEEE.

[23] Saba Bashir, Usman Qamar, M.Younus Javed et al. "An Ensemble based Decision Support Framework for Intelligent Heart Disease Diagnosis" International Conference on Information Society (i- Society 2014).

[24] Ammar Asjad Raja, Irfan-ul-Haq , Madiha Guftar Tamim Ahmed Khan and Dominik Greibl et al. "Intelligent Syncope Disease Prediction Framework using DM-Ensemble Techniques", FTC 2016 - Future Technologies Conference 2016.

[25] CI Education, Heart Disease Data Set [OL]. http://archive.ics.uci.edu/ml/datasets/Heart+Disease CHDD.

[26] T. Padmapriya and V.Saminadan, "Handoff Decision for Multi- user Multiclass Traffic in MIMO-LTE-A Networks", 2nd International Conference on Intelligent Computing, Communication & Convergence (ICCC-2016) – Elsevier - PROCEDIA OF COMPUTER SCIENCE, vol. 92, pp: 410-417, August 2016.

[27] S.V.Manikanthan and D.Sugandhi "Interference Alignment Techniques For Mimo Multicell Based On Relay Interference Broadcast Channel" International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN: 0976-1353 Volume- 7, Issue 1 –MARCH 2014.