CSE3505 Foundation of Data Analytics

Project Title:

Heart Disease Analysis and Decision Tree to Predict Pathological Suspects

Abstract:

With the new and sedentary lifestyle which a large majority of the population across the world is being exposed to, there has been a significant increase in heart diseases like those of heart failure, pericardial diseases, heart stroke etc. In order to analyze and portray the severity of this issue, we have decided to conduct heart disease data analysis and also build a classification and regression tree using cardiotocographic data using essential parameters like fasting blood sugar, serum cholesterol, max heart rate, etc. Through this analysis, we will be able to identify if a person is prone to have a heart disease or not using logistic regression for gaining accurate and precise results. Apart from this, we will also be curating a decision tree using classification and regression to classify patients as pathological and non-pathological suspects of heart diseases. The motive behind choosing this problem statement for data analysis is to create awareness on the severity of heart diseases today and also provide a substantial solution in predicting if a person is likely to have a heart ailment or not.

Literature Survey:

S. No	Title	Journal/Year of Publication	Dataset Used	Methodologies Used	Metrics used	Interpretation of Results	Reference Link
1	Identification of significant features and data mining techniques in predicting heart disease, Telematics and Informatics	Volume 36, 2019, Pages 82-93, ISSN	UCI Machine Learning Repository - Cleveland dataset and UCI Statlog Heart disease dataset	k-NN, Naïve Bayes, Vote, Support Vector Machine and Neural Network	Accuracy, F-measure and Precision. Accuracy is the percentage of correctly predicted instances among all instances. F-measure is the weighted mean of the precision and recall. Precision is the percentage of correct predictions for the positive class.	Precision and accuracy of prediction was highest for K-NN, followed by Vote, Naïve Bayes, SVM, Neural Network	https://www.sciencedirect.com/science/article/abs/pii/S0736585318308876
2	Prediction of coronary heart disease using machine learning: An experimental analysis	Proceedings of the 2019 3rd International Conference on Deep Learning Technologies	South African Heart Disease which is a subset of a larger dataset. It contains 462 instances (observations) and 10 attributes in all (shown in Table 1), of	Decision Tree, Naïve Bayes Algorithm, SVM	The performance of the classification models derived by the ML is measured using the confusion matrix. The confusion matrix is a contingency table that displays the number of instances assigned to each class thus allowing us to calculate the classification	NB achieved the highest accuracy amongst the three models. SVM and DT J48 outperformed NB with a Specificity rate of 82% but proved to have an unacceptable Sensitivity rate of less than 50%. While NB Algorithm didn't reach the threshold of 80% Specificity and Sensitivity rate, it did turn out to be	https://dl.acm.org/doi/pdf/10.1145/3342999.3343015?casa_token=EQ63fwch8XoAAAAA:L6EK91Udq48GClSOWPgHDdAnE1-PTEZd6ZZ416bs4dF67u92qLlSxUanXUaRd1ZuJlFd4nyto8MGyw

			which 9 are independent factors and 1 variable, i.e. CHD is the dependent variable or labelled class. The dataset is a retrospective sample of males in a heart-disease high-risk region of the Western Cape in South Africa-KEEL [28] where the labelled class CHD has two predictive outcomes: positive (1) and negative (0).		accuracy, sensitivity, specificity, true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) among others	the best classifier for the considered dataset as its predictive rate is better that those of J48 and SVM algorithms at least on the considered dataset.	

3	Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease	2007 International Conference on Convergence Information Technology	The dataset consisted of 1000 consecutive patients who underwent coronary angiography for known coronary atherosclerosis at Anzhen hospital, capital University of Medical Sciences, Beijing from August 2005 to December 2005.	SVM, Artificial Neural Networks (ANN)	Three performance metrics were employed: accuracy, sensitivity and specificity. A distinguished confusion matrix is obtained to calculate the three measures. Confusion matrix is a matrix representation of the classification results. the upper left cell denotes the number of samples classifies as true while they were true (i.e., true positives), and lower right cell denotes the number of samples classified as false while they were actually false (i.e., true false). The other two cells (lower left cell and upper right cell) denote the number of samples misclassified.	The survey was done on a data of 1000 CHD cases with 11 attributes. In this research, they defined survival as any incidence of CHD where person is still alive after 6 months from the date of diagnosis. Theyused a binary categorical survival variable, which was calculated from the variables in the raw dataset, to represent the survivability where survival is represented with a value of ‘‘1’’ and non-survival is represented with ‘‘0’’. The aggregated results indicated that the SVM performed the best with a classification accuracy of 92.1%, the ANN model (with multi layered perceptron architecture) came out to be second best with a classification accuracy of 91.0%. The results showed here make clinical application more	https://ieeexplore.ieee.org/abstract/document/4420369

						accessible, which will provide great advance in healing CHD.	
4	Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques	International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012	Cleveland Heart Disease database consists of 303 records & Statlog Heart Disease database consists of 270 records	ANN, Naïve Bayes	A confusion matrix is obtained to calculate the accuracy of classification. It shows how many instances have been assigned to each class. In their experiment they have two classes, and therefore they have a 2x2 confusion matrix.	Three data mining classification techniques were applied namely Naive Bayes & Neural Networks. From results it has been seen that Neural Networks provides accurate results as compare to Naive Bayes.	http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.258.8158&rep=rep1&type=pdf
5	Early Prediction of Heart Disease Using Decision Tree Algorithm	International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST) Vol.3, Issue.3, March 2017	UCI Repository Dataset	Decision Tree (C4.5), Naïve Bayes	Accuracy, Sensitivity and Specificity are computed on the basis of the confusion matrix that is curated.	From results, it has been seen that Decision trees provides accurate results as compare to Naive Bayes. This system can be further expanded. It can use more number of inputs.	https://www.researchgate.net/profile/Safish-Mary/publication/315023624_Early_Prediction_of_Heart_Disease_Using_Decision_Tree_Algorithm/links/58c84b57aca2723ab16eba60/Early-Prediction-of-Heart-Disease-Using-Decision-Tree-Algorithm.pdf

6	Coronary Heart Disease Diagnosis Through Self-Organizing Map and Fuzzy Support Vector Machine with Incremental Updates	International Journal of Fuzzy Systems, Springer, 2020	UCI Repository for datasets of Cleveland and Statlog	Self Organizing Map, SVM	The experimental setup of our experiment on the dataset with missing values is as follows. The dataset is divided into two subsets, 70% of the dataset as training subset and 30% of dataset as test subset. Then we applied the imputation procedure through hot-deck and k-NN on the training set for missing value imputation. We then applied PCA on each cluster which does not include the missing values. In the final stage, the classification models were constructed by Fuzzy SVM. To obtain the classification accuracy, the Fuzzy SVM classification model was evaluated on the test set.	The results revealed that the dataset imputation has a positive relationship with the accuracy of the Fuzzy SVM classifier. In addition, we found that the methods which rely on PCA provide better accuracy in relation to the other methods. In fact, it was found that, in the medical dataset the multicollinearity can significantly affect the predictive accuracy of the classifiers. Our experimental findings on two datasets also showed that the use of the methods with incremental techniques can have advantages on enhancing the computation time of disease prediction.	https://link.springer.com/article/10.1007/s40815-020-00828-7

7	Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm	International Journal of Biological and Medical Sciences 3:3 2008	UCI Cleveland heart-disease database	The CANFIS model integrates adaptable fuzzy inputs with a modular neural network to rapidly and accurately approximate complex functions. In order to improve the learning of the CANFIS, quicker training and enhance its performance, we use genetic algorithms to search for the best number of MF for each input, and optimization of control parameters such as learning rate, and momentum coefficient. This	Mean Square Error	The performances of the CANFIS model were evaluated in terms of training performances and classification accuracies and the results showed that the proposed CANFIS model has great potential in predicting the heart disease.	http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.451.9421&rep=rep1&type=pdf

				approach also is useful to select the most relevant features of the training data which can produce a smaller and less complicated network, with the ability to generalize on freshly presented data, due to the removal of redundant variables.			

8	A Reliable Feature Selection Algorithm for Determining Heartbeat Case using Weighted Principal Component Analysis	2016 International Conference on System Science and Engineering (ICSSE) National Chi Nan University, Taiwan, July 7-9, 2016	MIT-BIH Arrhythmia Database	Weighted Principal Component Analysis (WPCA) method	Sensitivity, speed, reliability, detection	This study has presented a simple and reliable WPCA method to analyze ECG signals for diagnosing cardiac arrhythmias. The proposed method has the following advantages: (1) good detection results: sensitivities are 95.29%, 93.35%, 92.29%, 79.98%, 91.55% and 90.07% for heartbeat cases NORM, LBBB, RBBB, VPC, APC and PB, respectively; (2) simplicity: complicated mathematical computations are unnecessary; (3) high speed: the average time required for processing a 30-minute record of ECG data is less than 1 minute; and (4) high reliability: total classification accuracy approximates 93.19%. Therefore, the proposed WPCA is an efficient, simple and fast method for diagnosing cardiac arrhythmia based on ECG signals.	https://ieeexplore.ieee.org/abstract/document/7551594

9	Hybrid Classification Model of Correlation-based Feature Selection and Support Vector Machine	International Conference on System Science and Engineering (ICSSE), 2016	5 high dimensional datasets like Breast_2, Colon, DLBCL, Leukaemia and Prostate as shown in Table 1. All datasets are of binary classes (only two classes). The numeric values 1 and -1 are taken to represent classes.	CORR SVM hybrid model for Classification	Classification accuracy	This paper presents a Hybrid of Supervised Correlation method and Support Vector Machine for classification of high dimensional datasets. First each feature's absolute correlation value with respect to class is calculated and keep it into an array call array0. Then sort array0 in descending order of values and then sort features according to sorted array0 call this list as list1. Then select top K (a user defined number) features from list1 which forms reduced dataset. Then calculate classification measures with various options as presented in the literature	https://ieeexplore.ieee.org/abstract/document/7567338