Application of Machine Learning Algorithms in Predicting Coronary Heart Diseases

ABSTRACT

In recent years, research works have been done using data mining to predict the risk of developing diseases based on medical tests. The aim of this study is to predict coronary heart diseases (CHD) by applying a machine learning algorithms and comparing their performances. As a result, the accuracy reached 83.83%% for Random forest classifier and area under ROC curve was 76% for Logistic regression. Being able to predict cardiovascular diseases in early stage could make major impact on changing a lifestyle in high risk patents and reduce future complications of developing hearth diseases**.**

INTRODUCTION

Coronary heart diseases (CHD) has been leading cause for death worldwide. According to the World Health Organization (WHO) 12 million deaths occur worldwide every year due to heart diseases [1]. One third of all deaths in the United States and other developed countries in people older than 35 are due to cardio vascular diseases [2]. In recent years data mining techniques are being applied to clinical environments such as diagnostic radiology, diabetes, dermatology and much more [13] .In this work data analytics was applied on the Framingham cardiovascular study that was collected over period of time. The main focus is to perform in depth analysis to predict cardiovascular diseases that could happen in future based on risk factors.

RELATED WORK

Lot of studies involving the use of predictive analytics was used to observe patterns in coronary heart diseases. Some of the techniques are more complex than others and involve use of more than one dataset. In this section we will study various techniques used by different authors which will help answer questions such as: what techniques are used for data preprocessing and what are the classification techniques which have proven to be most effective in each study.

In one of the papers [9], authors used comparison between two algorithms: support vector machine (SVM) algorithm and artificial neural network(ANN) model to compare positive predicted value(PPV).ANN algorithm is simple has a high speed and ability to solve complex relationship between variables and extracting the non-linear relationship by means of training data. On the other hand SVM is most common machine learning algorithm that can be used in classifications, recognition and prediction in supervised learning [9].For both techniques same predictive variables were used: gender, occupation, place of residency, family history , smoking status, hypertension history , chest pain cholesterol, blood glucose level and it was broken down into 70% training and 30% test. PPV was calculated based by computing the area under ROC curve. The closer this value is to 1 the higher the PPV of the model. Findings showed that ANN model had high PPV when predicting CAD. Furthermore, even SVM model showed moderate power in predicting CAD patients it had higher PPV in classifying and predicting CAD and higher accuracy and presented better classification.

Next, the authors of [10] carry out their experiment using data mining tool Weka and applying two algorithms: multilayer perception neural network (MLPNN) with backpropagation (BP). The MLPNN is subdivided into one input layer, one or more hidden layers and one output layer[10]. The input nodes pass values to hidden layer and then nodes of first hidden layer pass values to second one and so on until product output presents. BP algorithm is used to calculate differences between real and predicted values which is circulated from output nodes backward to nodes in previous layer. Dataset contained around 300 records and were divided into training (40%)and test set (60%).Confusion matrix was created and contained information about real and predicted classifications and contained four different entries: TP(true positive) FP(false positive)FN(false negative) and TN(true negative). After applying algorithms on training dataset system predicted heart diseases with 100% accuracy [10].

Belcy, et al.[11] have performed work to develop a prediction system in order to predict the possibility to develop CHD by using three classifiers: Random Forest, Decision Tree and Naïve Bayes. Their research methodology includes evaluation using cross validation applying 10-fold cross validation methods and evaluation using percentage split. For percentage split the training and testing data is split up in percentage of data such as 80% and 20%,60%and40%,70%and 30% Based on their results random forest algorithm performed best with 81% precession for 10-fold cross validation.

In this paper five different algorithms were used and their metrics results were compared.

DATA

Description of data

Dataset was found on Kaggle website [12] and it is from current cardiovascular study on citizens of Framingham town in Massachusetts. It includes over 4000 records and 15 attributes which presents possible risk factors for coronary hearth diseases.

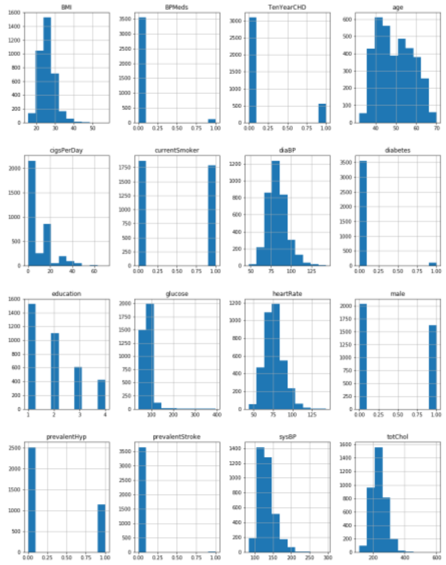
Attributes are divided into categories as: Demographic: sex of the patient (male or female) age and education. Behavioral: currentSmoker (if patient is current smoker or not) and cigsPerDay which presents an average number of cigarettes that patient smokes per day. Last category is Medical history of patient which includes: BPMeds: if patients is taking blood pressure medication; prevalentStroke if patent had stroke or not; prevalentHyp: if patient was hypertensive or not; diabetes: if patient was diagnosed with diabetes; totChol: presents total level of cholesterol:sysBP is patients systolic blood pressure; diaBP is patient diastolic blood pressure; heartrate: patients heart rate; glucose: patients glucose level. Variable to predict is 10 years risk of coronary heart disease CHD which was presented on binary form where “0” means No and “1” means yes for CHD.

Exploratory data analysis:

Exploratory data analysis was done to get insight of data using Python 3.7 Google Colab. The Framingham dataset consisted of 4240 observations:42.9 % corresponded to men and 57.1% to women. Average age was 49.6 years and percentage of patients who were smokers was 49.4% with average of 9 cigarettes per day. Average total cholesterol was 236.7, systolic blood pressure 132.4 and diastolic blood pressure 82.9. Average glucose level was 81.9 and total number of participants that suffered from coronary events was 15.2%.

The Figure 1. Presents Histogram of all columns and from it, was concluded that column 10years risk of coronary diseases is highly imbalanced with number of participants without CHD significantly higher.

Next, correlation between attributes was analyzed and its presented in Figure 2. Based on results column Education was excluded from future analyses since correlation factor was insignificant (0.0063).



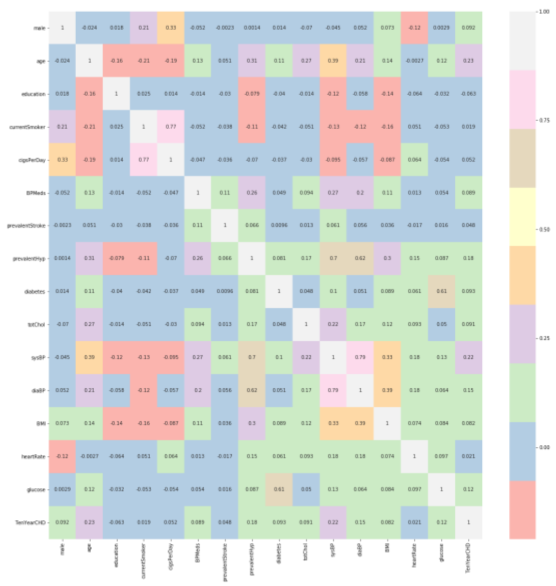


Figure 1. Histogram of Dataset

Figure 2. Correlation Heat Map

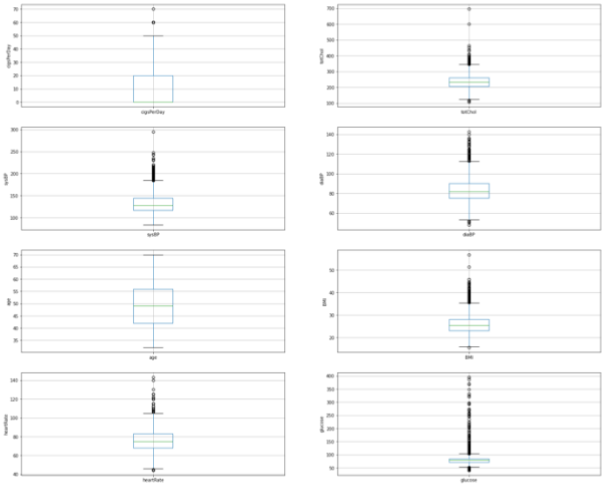


Figure 3. Boxplot for Outlier Estimation

Data preprocessing

As one of the primary steps of data preprocessing was checking if there are any missing values and handling them. Variables with missing data were education, number of cigarettes per day, blood pressure medication, total cholesterol, BMI index, heart Rate and Glucose which had larger number of missing values (388). Result show that total number of rows with missing values was 582. Since that was 14 percent of entire dataset rows missing values were replaced with mean from rest of non-missing values in each column.

Next, step in data preprocessing was to check for outliers which are observation points that are distant from other observations. Box plot in Figure 3. was used to see which columns have outliers and based on its results points that are outside of lower and upper quartiles were considered to be outliers. Total of 286 observations were removed from dataset.

Dataset was randomly split into 3 different sets: Train set 80% of data, Validation set 10 % of data and Test set remaining 10%. Numerically Training test contain 3202 observations while test validation and test set 356 and 396 observations respectively. The Validation set was used during model fitting to evaluate metrics. The Test set was completely unused during the training phase to avoid leaking any data during Training phase. Test set was used only at the end to evaluate how well the model generalize to new data. This is especially important with imbalanced datasets where overfitting is significant concern from the lack of training data.

From histogram, earlier, was concluded that target column TenYearCHD is highly imbalanced. Number of participants without CHD was 3391 and number of participants with CHD was 563 which is 6.02 : 1 ratio.(Figure 4.). In order to balance data, up-sampling method SMOTE(Synthetic Minority Oversampling Technique) was applied to oversample minority group. Oversampling was done only on the training data that none of the information in the test set is being used to create synthetic observations and there is no information that will bleed from test data into the model training. SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in feature space and drawing a new sample at point along that line. This procedure allows you to create as many synthetic examples for minority class as required. [14]

After applying SMOTE number of rows in training set increased to 5474 and ratio between 0 and 1 dropped to 1:1 with 2737 participants in each class .(Figure 5.)

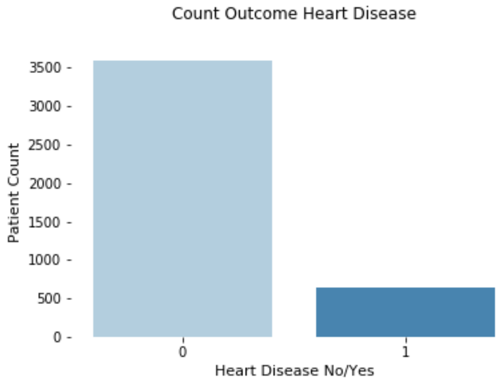


Figure 4. 10YearCHD before Normalizing

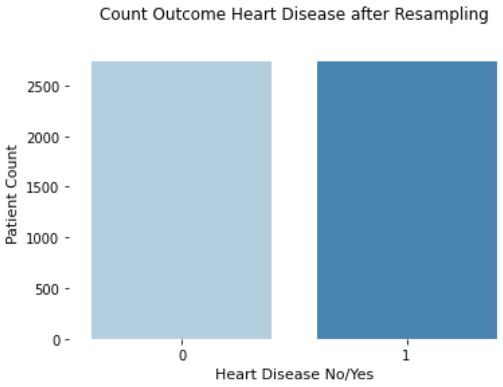


Figure 5.10YearCHD after applying SMOTE

METHODOLOGY

This work is aimed towards identifying the best classification algorithm for identifying the possibility of heart diseases. Comparative study and analysis was done using five different classification algorithms: Support Vector Machine, Logistic Regression, Decision Tree, Random Forest and K-Nearest Neighbor.

1.Support Vector Machine (SVM)is a supervised learning algorithm widely used for classification and regression. Objective of SVM is to find hyperplane that distinctly classify the data points but at the same time that hyperplane has maximum distance(margin) between data points of both classes.[4].I this research SVM model was built using kernel function ‘rbf’.

2.Logistic Regression is a classification learning algorithm widely used on the case of binary classification. In logistic regression the dependent variable is a binary variable that counts data code as 0 for no, failure and 1 for yes, success. It is based on predicting probabilities using Sigmoid function.[4]

3.Decision Tree classifier use decision tree to make a prediction about the value of target variable by learning simple decision rule inferred from the data [5]. The decision tree are functions that successfully determine the class that the input needs to be assigned. An input is tested against specific subsets of data, determined by the splitting criteria or decision function [6]. In this research ‘gini’ was used for impurity.

4.Random forest classifier generates multiple decision trees on different sub-samples of the data while training and then predicts the accuracy or loss score by taking a mean of this values. This helps to control overfitting that might happen when a single decision tree is used[7]. In this research work, parameter for Random forest are max_depth :100,max_features:auto,min-samples leaf:1 and n_estimator:1000.

5.K-Nearest Neighbor (KNN) – classify the data into one or the many categories by taking a majority vote of its neighbors. The label is assigned depending on the most common of categories among its neighbors [8]. In this research value for of 2 was used for k.

To obtain optimal parameters for each algorithm Grid Search Cross Validation was applied on Train set and K fold for cross validation was set to 5. With 5 fold cross-validation data training data was split into five folds and each was containing 20% of training data. The average of 5 metrics values was use as final value. When evaluating the model it is very important to do it on held out samples that were not seen during grid search process, therefore dataset was split to train-validation-test. Once parameters were obtained there were first applied to validation set and then test set was used at the last to evaluate how well the mode generalize to new data.

Different metrics were obtained in order to evaluate between machine learning algorithms. *Accuracy* presents ratio of numbers of correctly classified examples divided by total number of classified examples.[4] Problem with accuracy arises when cost of misclassification of the minor classes are very high that is case in imbalanced data. In our case, the fail to diagnose the disease of sick person is much higher then the cost of sending a healthy person to do more tests.

Confusion Matrix is table that summarize how successful the classification model is at predicting examples belonging to various classes.[4]There are 4 important terms :

True positive are the cases predicted YES and actual output was YES.

True negative are the cases predicted NO and actual output was NO.

False positive are the cases predicted YES and the actual output was NO.

False negative are the cases predicted NO and the actual output was YES.

Precision is the ratio of correct positive predictions to overall number of positive predictions(TP/TP+FP).

Recall is the ration of correct positive predictions to overall number of positive examples (TP/TP+FN)

Area under ROC curve are functions of the sensitivity and specificity for each value of the measure or model. The sensitivity of a test is the probability of a positive test result, or of a value above a threshold, among those with disease (cases). The specificity is the probability of a negative test result, or a value below a threshold, among those without disease (noncases).

ROC curve is combination of true positive rate (recall) and false positive rate. The AUC ranges from 0.5 to 1. Higher the area under curve (closer to 1) better classifier. [14]

RESULTS AND DISCUSSION:

After running different classifier models Metrics results are presented in Figure 6. Measure of accuracy, precision, recall and F1 score were used to measure performance of algorithms.

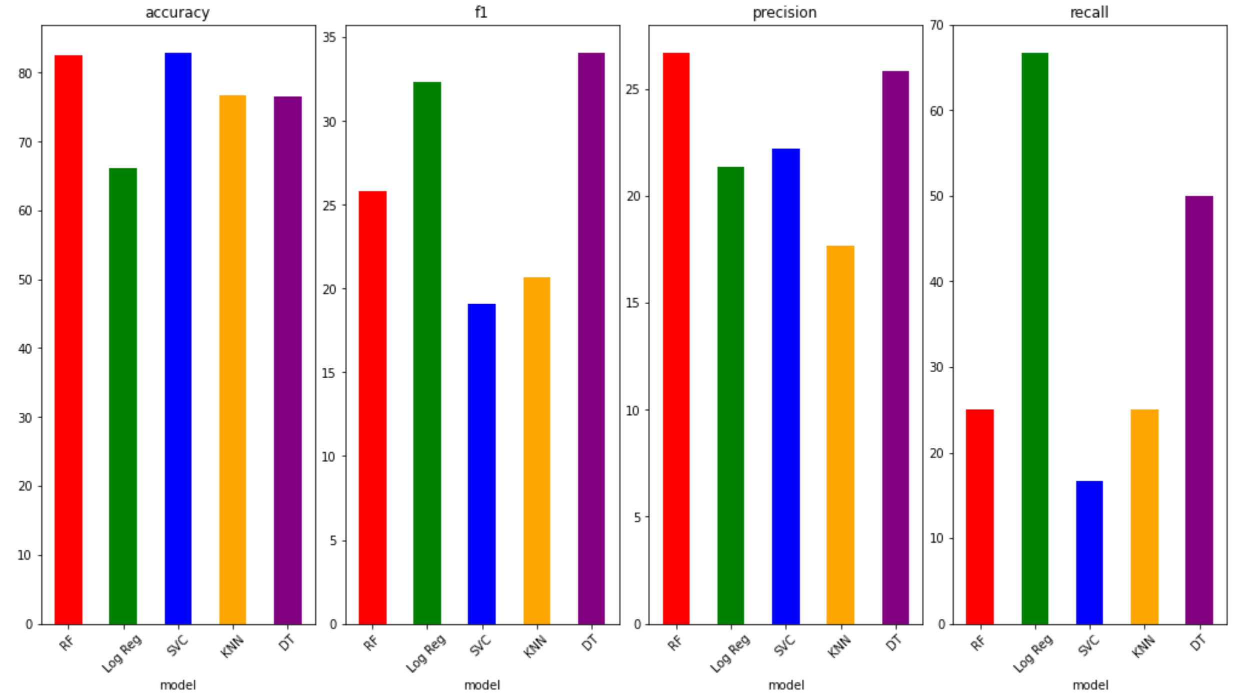


Figure 6. Metrics summary for each algorithm

When comes to Accuracy, Random Forest performed with highest accuracy score of 83.3% and was followed by Support Vector Classifier which performed with 82.83 % accuracy. We have to keep in mind that dataset was highly imbalanced and accuracy can give us false results. Same classifiers gave us highest results for precision as well. Random Forest was 29.5 % precise and Support vector 22.22%. Looking at F1 score Decision Tree and Logistic regression performed with higher scores: Logistic regression 66.67% and Decision Tree 58.33%. F1 scores were highest for the same algorithms: Logistic regression 32.32% and Decision tree with f1 score 31.28%.

We can see that algorithms that performed high with accuracy and precision gave lower result for recall and f1 score.

Importance in this research is to correctly classify patients that have hearth diseases and we want FN to be lowest as possible. Looking at Confusion metrics (Figure 7.) 0 presents patients without heart disease and 1 presents patients with no heart diseases. Logistic regression performed with lowest misclassification cases for FN patents of (16 patients) but it also have highest number of misclassification for FP(118 patients)

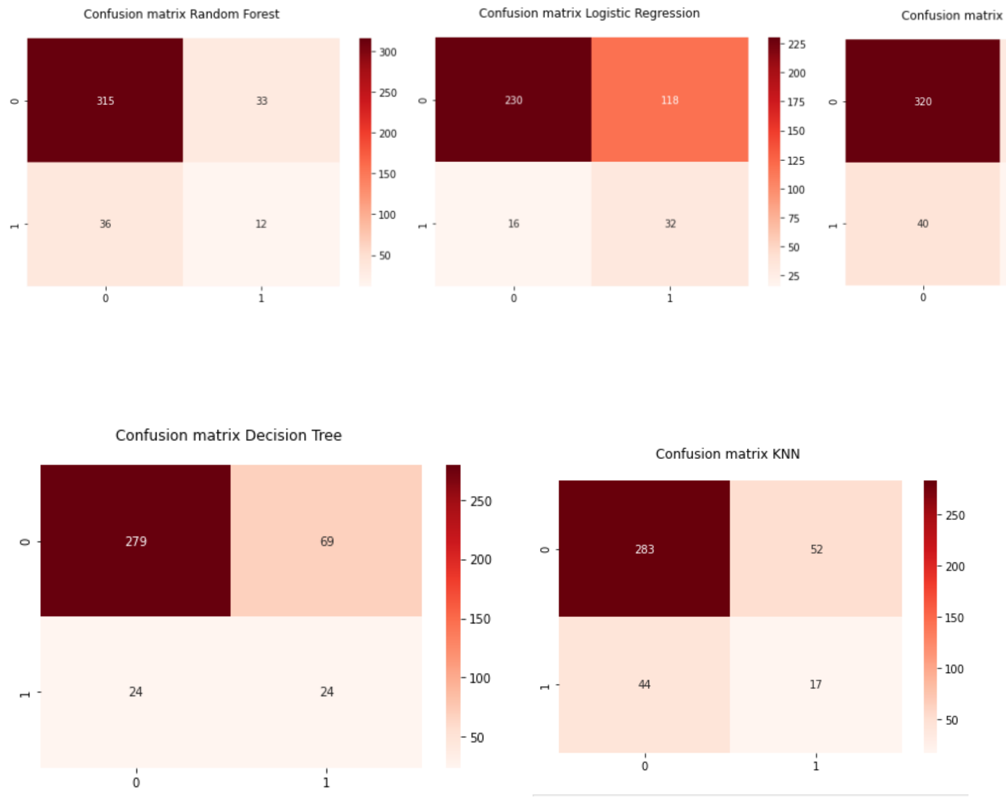


Figure.7 Confusion Matrix summary

Since dataset was imbalanced, we use area under the ROC curve as primary measure of the classification performance of the scaler. This curve plots the positive predictive values against negative predictive values. AUCs can be interpreted as following: if individual is randomly selected from the group labeled with high risk and another is randomly chosen from the group labeled lower risk the AUC is the probability that higher-risk individual has CHD risk greater than that of the lower -risk individual[14]. AUC is correct classification of those who had or did not have events. Looking at graph from Figure 8. we can conclude that Logistic regression preformed with higher AUC which is correctly classify of 76% of cases and Random forest AUC of 74%

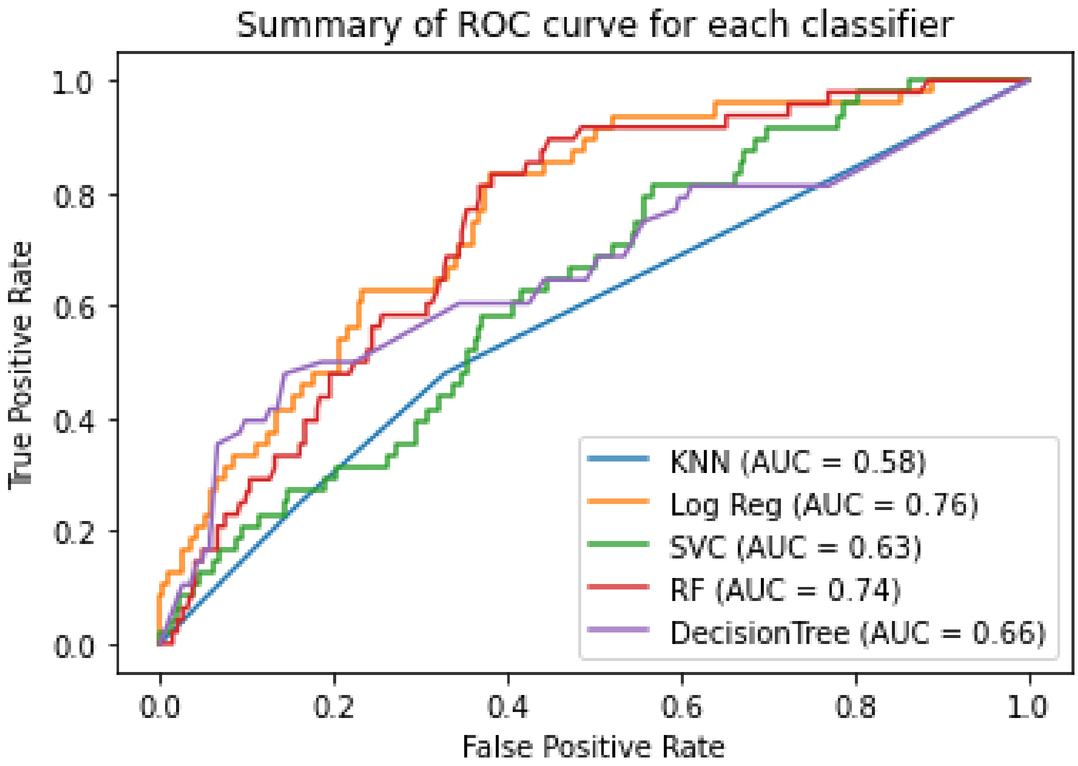


Figure 8. Summary of ROC for each classifier

Overall results were statistically significant with over 83% of accuracy and 76% AUC of predicting CHD based on risk factors from Framingham dataset on cardiovascular studies. Logistic regression was performing classifier for imbalanced data with 76% AUC and lowest misclassification cases for FN patients.However, model could be improved with more data samples or adding additional attributer that have high correlation with hearth diseases such are:LDL Cholesterol, family history of CHD, physical activity.

References :

1.Sanchis-Gomar,Fabian,Carme Perez-Quilis,Roman Leischik,& Alejandro Lucia. ‘Epidemiology of coronary hert diseases and acute coronary syndrome.*’Annals of Translational Medicine*[Online],4.113(2016):n.pag.Web 15 April.2020

2. Rosamond W, Flegal K, Furie K, et al. Heart disease and stroke statistics--2008 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Circulation 2008;117:e25-146

3.Orphanou, Kalia & Dagliati, Arianna & Sacchi, Lucia & Stassopoulou, Athena & Keravnou, Elpida & Bellazzi, Riccardo. (2018). Incorporating Repeating Temporal Association Rules in Naïve Bayes Classifiers for Coronary Heart Disease Diagnosis. Journal of Biomedical Informatics. 81. 10.1016/j.jbi.2018.03.002.

4.Burkov,Andriy. *The Hundred-Page Machine Lerning Book*.Andriy Burkov,2019.

5. Amarbayasgalan T, Park KH, Lee JY, Ryu KH (2019)” Reconstruction error based deep neural networks for coronary heart disease risk prediction”. PLoS ONE 14(12): e0225991. https://doi.org/10.1371/journal.pone.0225991

6. Song, Yan-Yan, and Ying Lu. “Decision tree methods: applications for classification and prediction.” *Shanghai archives of psychiatry* vol. 27,2 (2015): 130-5. doi:10.11919/j.issn.1002-0829.215044

7.Breiman,L.Random Forest.*Machine Learning* 45,5-32(2001) doi:10.1023/A:1010933404324

8. Kuśmirek, Wiktor et al. “Comparison of kNN and k-means optimization methods of reference set selection for improved CNV callers performance.” *BMC bioinformatics* vol. 20,1 266. 28 May. 2019, doi:10.1186/s12859-019-2889-z

9. Ayatollahi, Haleh et al. “Predicting coronary artery disease: a comparison between two data mining algorithms.” *BMC public health* vol. 19,1 448. 29 Apr. 2019, doi:10.1186/s12889-019-6721-5

10. Singh, Poornima et al. “Effective heart disease prediction system using data mining techniques.” *International journal of nanomedicine* vol. 13,T-NANO 2014 Abstracts 121-124. 15 Mar. 2018, doi:10.2147/IJN.S124998

11. Kamaraj, K.Gomathi & Priyaa, D.Shanmuga. (2016). Heart Disease Prediction Using Data Mining Classification. International Journal for Research in Applied Science & Engineering Technology (IJRASET). 4. 60-63.

12. Aman Ajmera, Framingham Heart study dataset [online dataset]. Kaggle Inc; URL <https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset>

13. Beunza, Juan-Jose, et al. “Comparison of Machine Learning Algorithms for Clinical Event Prediction (Risk of Coronary Heart Disease).” *Journal of Biomedical Informatics*, vol. 97, 2019, p. 103257., doi:10.1016/j.jbi.2019.103257.

14. Artigao-Rodenas LM, Carbayo-Herencia JA, Divisón-Garrote JA, Gil-Guillén VF, Massó-Orozco J, Simarro-Rueda M, et al. (2013) Framingham Risk Score for Prediction of Cardiovascular Diseases: A Population-Based Study from Southern Europe. PLoS ONE 8(9): e73529. https://doi.org/10.1371/journal.pone.0073529

15. Brownlee,Jason.’SMOTE for Imbalanced Classification with Python.*’ Machine Learning Mastery*,7 Apr.2020, machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.