Using Machine learning to detect cardiovascular diseases

[1]Prabhleen kaur saini
[1]Student
[1]iet davv

Abstract - **The intersection of medical science and technology is the need of the hour. Deaths due to cardiac problems has sky rocketed over the past few years. Hence, monitoring the proper functioning of our heart becomes necessary. Although clinical testing is available but it fails to utilize large amount of patient records and data generated. Leveraging medical data, we can create trustworthy systems that can make detection of cardiac problems easier and economical. Machine learning helps to achieve this goal, it enables us to analyse patient records and infer patterns that helps to conclude whether someone has cardiovascular disease or not. In this paper we aim to find the best machine learning model by comparing different models as well as find parameters that are of utmost importance in deciding whether or not someone has cardiovascular diseases.**

keywords - **cardiovascular diseases, machine learning, logistic regression**

I. INTRODUCTION

According to the world health organisation cardiovascular diseases account for 17.9 million deaths each year, which is about 31% of total deaths worldwide. More than 75% of these deaths occur in developing countries. The major reason behind this is lack of proper medical facility in these countries where ratio of doctors to patients very low. The tell-tale symptoms of cardiovascular diseases can easily be diagnosed at any primary health facility. Identifying these signs at an early stage can be a boon in saving lives. Moreover, these days people are adopting lifestyles that make them more susceptible to cardiac problems. Having a tool that can help detect whether or not a person is prone to cardiovascular diseases can be of great benefit for the medical staff to provide results quickly and easily. There will be an increased awareness if these tools can be easily accessed by the general public.

In this paper we are going to explore if we can find a some key feature that can play a pivot role in deciding whether or not someone has cardiovascular diseases.

II. DATASET

The data used for modelling is obtained from Kaggle. It consists of 14 attributes that are used to classify whether a patient has heart disease (target =1) or does not has heart disease (target = 0). For further details about the data set refer to the reference section.

Description of the attributes is as follows:

INDEX	FEATURE NAME	MEANING
1.	**Age**	Age in years
2.	**Sex**	1 = male; 0 = female
3.	**cp**	chest pain type • 0: Typical angina: chest pain related decrease blood supply to the heart • 1: Atypical angina: chest pain not related to heart • 2: Non-anginal pain: typically esophageal spasms (non heart related) • 3: Asymptomatic: chest pain not showing signs of disease
4.	**trestbps**	resting blood pressure (in mm Hg on admission to the hospital) anything above 130-140 is typically cause for concern
5.	**chol**	serum cholesterol in mg/dl • serum = LDL + HDL + .2 * triglycerides • above 200 is cause for concern
6.	**fbs**	fasting blood sugar > 120 mg/dl

		(1 = true; 0 = false) '>126' mg/dL signals diabetes
7.	**restecg**	resting electrocardiographic results • 0: Nothing to note • 1: ST-T Wave abnormality ▪ can range from mild symptoms to severe problems ▪ signals non-normal heart beat • 2: Possible or definite left ventricular hypertrophy ▪ Enlarged heart's main pumping chamber
8.	**thalach**	maximum heart rate achieved
9.	**Exang**	exercise induced angina (1 = yes; 0 = no)
10.	**oldpeak**	ST depression induced by exercise relative to rest looks at stress of heart during exercise unhealthy heart will stress more
11.	**slope**	the slope of the peak exercise ST segment • 0: Upsloping: better heart rate with excercise (uncommon) • 1: Flatsloping: minimal change (typical healthy heart) • 2: Downslopins: signs of unhealthy heart
12.	**ca**	number of major vessels (0-3) colored by flourosopy • colored vessel means the doctor can see the blood passing through • the more blood movement the better (no clots)
13.	**thal**	thalium stress result • 1,3: normal • 6: fixed defect: used to be defect but ok now • 7: reversable defect: no proper blood movement when excercising
14.	**target**	have disease or not (1=yes, 0=no)

III. DATA EXPLORATION

Following dataset gives a brief overview of the values present in the dataset.

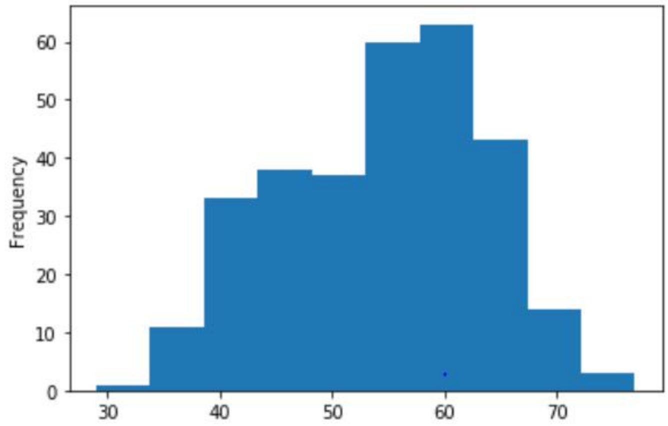
- Target column: Final value we want to predict or classify.
- Features refer to those columns that are used to find specific patterns that can be then used to predict a value for the target column.

	age	**sex**	**cp**	**trestbps**	**chol**	**fbs**	**restecg**	**thalach**	**exang**	**oldpeak**	**slope**	**ca**	**thal**	**target**
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000

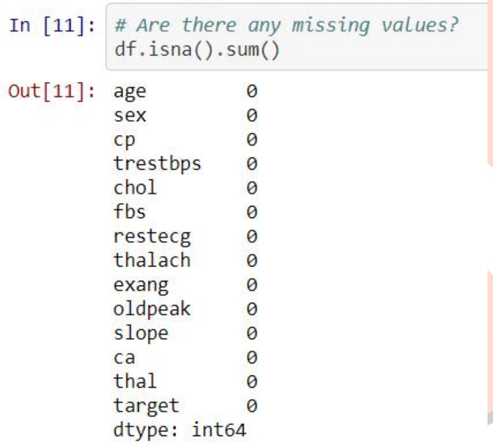
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

There is a total of 303 patient records with age ranging from 29 to 77 years. Average age is 54 years. About 163 people were found to be positive for cardiovascular disease and 135 negatives.

Following graph shows the distribution of age in the dataset:



1. **Missing values.**



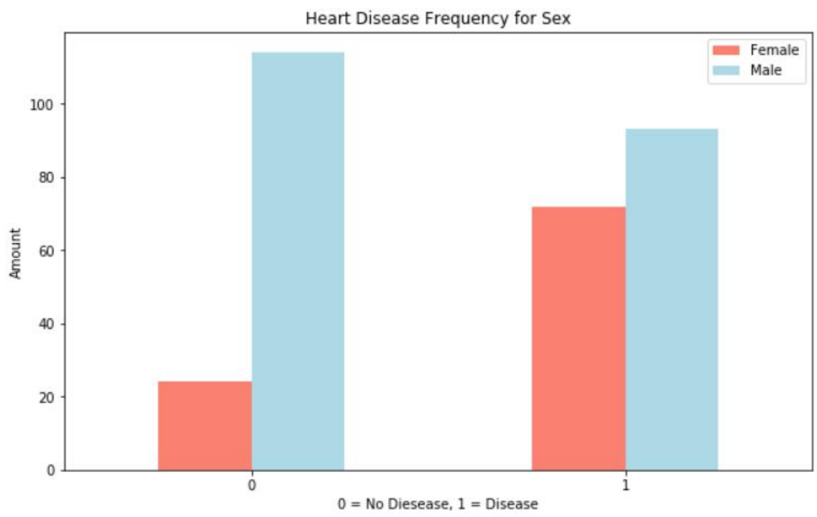
```
In [11]: # Are there any missing values?
         df.isna().sum()

Out[11]: age         0
         sex         0
         cp          0
         trestbps    0
         chol        0
         fbs         0
         restecg     0
         thalach     0
         exang       0
         oldpeak     0
         slope       0
         ca          0
         thal        0
         target      0
         dtype: int64
```

Clearly, there are no missing values in the dataset.

2. **Cardiovascular diseases vs sex**

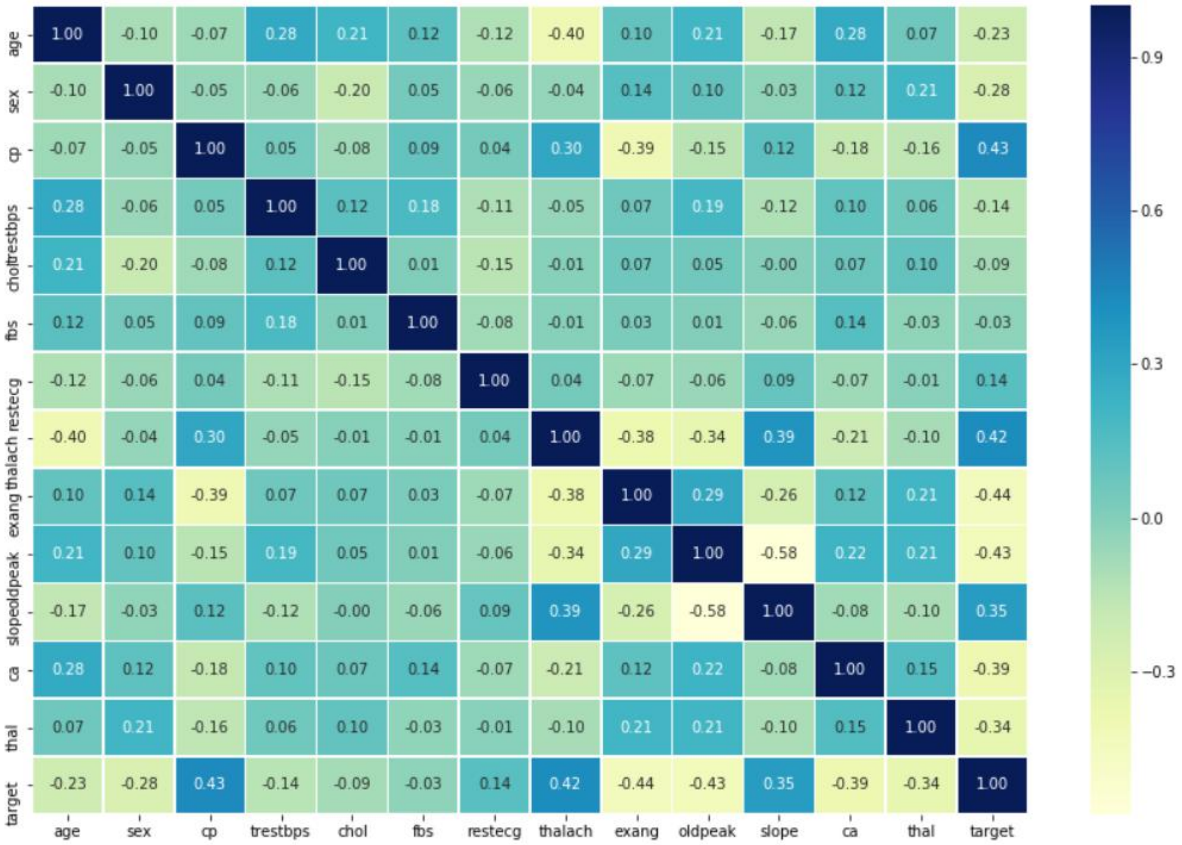
Let’s check how does chances of having cardiovascular diseases varies with sex:



Clearly, this shows that given data set is biased towards women. Which means that about 75% of total women were found positive for cardiovascular disease compared to 15% of total men.

3. Co-relation matrix

We can use a co-relation matrix to find further insights on how different attributes affect the target column.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.00	-0.10	-0.07	0.28	0.21	0.12	-0.12	-0.40	0.10	0.21	-0.17	0.28	0.07	-0.23
sex	-0.10	1.00	-0.05	-0.06	-0.20	0.05	-0.06	-0.04	0.14	0.10	-0.03	0.12	0.21	-0.28
cp	-0.07	-0.05	1.00	0.05	-0.08	0.09	0.04	0.30	-0.39	-0.15	0.12	-0.18	-0.16	0.43
trestbps	0.28	-0.06	0.05	1.00	0.12	0.18	-0.11	-0.05	0.07	0.19	-0.12	0.10	0.06	-0.14
chol	0.21	-0.20	-0.08	0.12	1.00	0.01	-0.15	-0.01	0.07	0.05	-0.00	0.07	0.10	-0.09
fbs	0.12	0.05	0.09	0.18	0.01	1.00	-0.08	-0.01	0.03	0.01	-0.06	0.14	-0.03	-0.03
restecg	-0.12	-0.06	0.04	-0.11	-0.15	-0.08	1.00	0.04	-0.07	-0.06	0.09	-0.07	-0.01	0.14
thalach	-0.40	-0.04	0.30	-0.05	-0.01	-0.01	0.04	1.00	-0.38	-0.34	0.39	-0.21	-0.10	0.42
exang	0.10	0.14	-0.39	0.07	0.07	0.03	-0.07	-0.38	1.00	0.29	-0.26	0.12	0.21	-0.44
oldpeak	0.21	0.10	-0.15	0.19	0.05	0.01	-0.06	-0.34	0.29	1.00	-0.58	0.22	0.21	-0.43
slope	-0.17	-0.03	0.12	-0.12	-0.00	-0.06	0.09	0.39	-0.26	-0.58	1.00	-0.08	-0.10	0.35
ca	0.28	0.12	-0.18	0.10	0.07	0.14	-0.07	-0.21	0.12	0.22	-0.08	1.00	0.15	-0.39
thal	0.07	0.21	-0.16	0.06	0.10	-0.03	-0.01	-0.10	0.21	0.21	-0.10	0.15	1.00	-0.34
target	-0.23	-0.28	0.43	-0.14	-0.09	-0.03	0.14	0.42	-0.44	-0.43	0.35	-0.39	-0.34	1.00

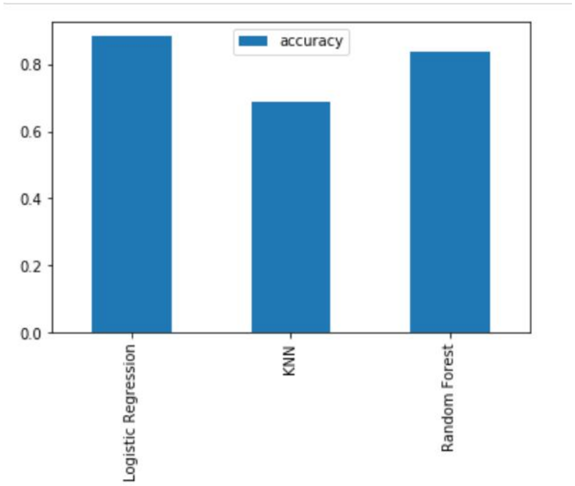
- A value of 1 means complete dependency whereas a value of zero means that the two variables are independent.
- A positive value means that as the value of feature variable increases the value of target variable increases slightly.
- A negative value means the as the value of feature variable increases the value of target variable decreases slightly.

IV. Modelling

We'll divide the given that into features and target column and then split it in training and test set.

- The training set is used by the model to learn patterns in the data.
- Test set is used to check performance of the patterns learnt by the model. They help the model to generalise as well as in hyperparameter tuning.

We'll be using 3 different models: Logistic regression, k-nearest neighbors (KNN) and Random forest classifier.
Training these models for the first-time yields following accuracy:
'Logistic Regression': 0.8852459016393442,
'KNN': 0.6885245901639344,
'Random Forest': 0.8360655737704918

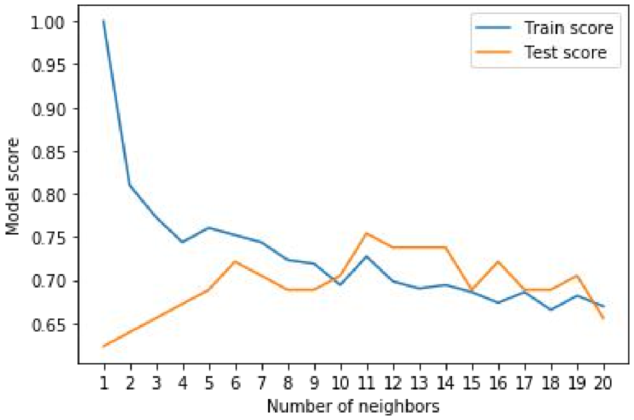


Logistic regression model has maximum accuracy of 88.52% whereas the KNN model has minimum accuracy of 68.88%.

V. HYPERPARAMETER TUNING

Let’s try to improve the accuracy of ‘knn’ model.

- **Hyperparameter tuning by hand:**
  We trained the ‘knn’ model for different values of n_neighbors between the range 0-20. Maximum KNN score obtained on the test data: 75.41%



As the maximum accuracy achieved is less than the other two models we’ll not consider the ‘knn’ for further discussion.

- **Hyperparameter tuning using randomized search cv:**
  Logistic
  We’ll change the value of parameters “C” and “solver”. After fitting the model on 150 different combinations we get a maximum accuracy of 0.8852 and the best combination of parameter value 'solver' = 'liblinear' and 'C' = 0.2335
  Random Forest Classifier
  We’ll change the value of the parameters “n_estimators”,” max_depth”, “min_samples_split” and “min_sample_leaf”. After fitting the model on 100 different combinations we get a maximum accuracy of 0.86.88
  And the best combination of parameter values
  'n_estimators'= 210,
  'min_samples_split = 4,
  'min_samples_leaf'= 19,
  'max_depth'= 3

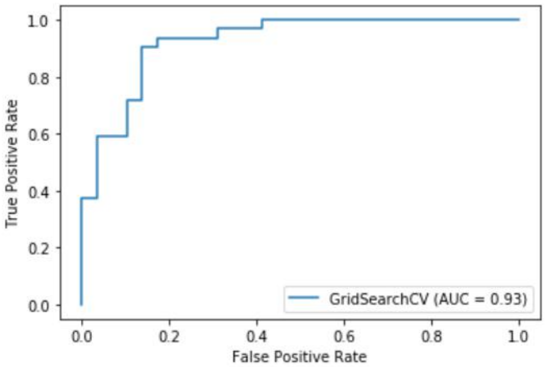
Since the Logistic Regression model provides the best scores so far, we'll try and improve it again using GridSearchCV.

- **Hyperparamter Tuning with GridSearchCV**
  By using grid search cv we are find the best possible combination of params *to be* 'C'= 0.20433597178569418, 'solver'= 'liblinear' and accuracy equal to 0.8852.

VI. Evaluation

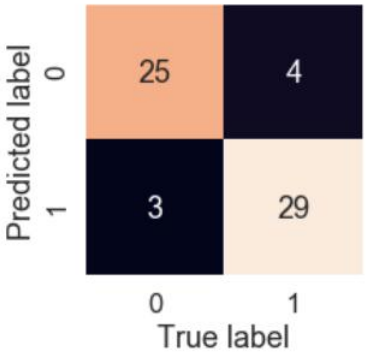
- **ROC curve**
  Roc curves are comparison between model’s true positive rate versus a model’s false positive rate. The curve is plotted between these two values, and for an ideal curve the area under the curve is 1.
  True positive: model predicts 1 when truth is 1.
  False positive: model predicts 1 when truth is 0.
  True Negative: model predicts 0 when truth is 0.

False negative: model predicts 0 when truth is 1.

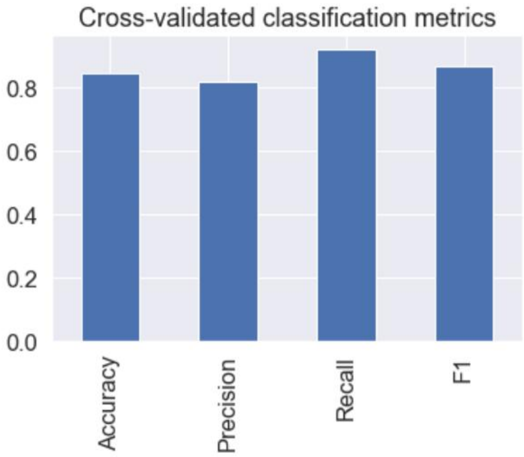


The area under the curve for our model is 0.93 which is close to the ideal score of 1. This implies, our model predicts less false positives and more true positives.

- **Confusion Matrix:**
A confusion matrix is a comparison between the labels a model predicts versus the labels it was supposed to predict. In sum, it gives an idea of where the model is getting confused.

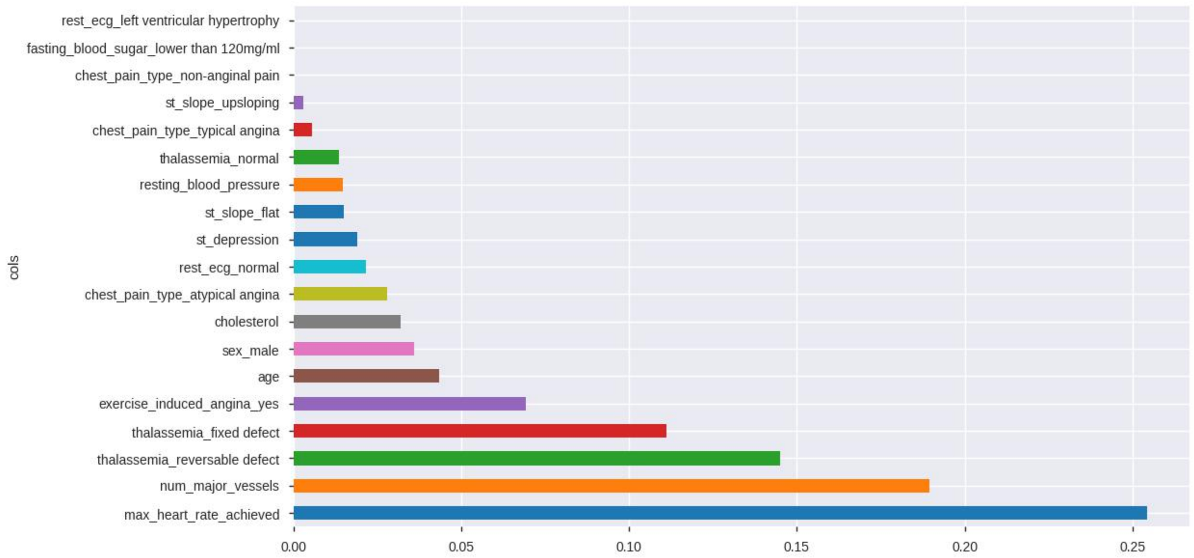


- **Classification report**
Classification reports are particularly useful when there is a class imbalance in samples i.e. when one class has much more samples than the other class. Other than accuracy it compares where other metrics like:
Precision: Indicates the proportion of positive identification which were actually correct.
Recall: Proportion of actual positive that were correctly classified.
F1 Score: A combination of precision and recall.
Cross validation: The technique of splitting the data into different train and test set, so that the results are not just based on one luck split. By using cross validation, the results are more trustworthy.
Values of different cross validated evaluation metrics:
accuracy – 84.47%
precision – 82%
recall - 92%
f1 score – 86.73%



VII. Feature Importance

Some attributes or labels are more important than others while determining the value of target column. The following graph highlights the contribution of various features in determining the value of target column.



VIII. Conclusion

Logistic Regression is the best suitable model with an accuracy of 84%. Clearly maximum heart rate, age, clotting in blood vessels proves to be a significant indicator of cardiovascular disease. These results are in line with the medical science theories. Hence, accurate prediction of these values using machine learning can help detect cardiovascular disease at an early stage.

IX. Acknowledgements

Principal investigator responsible for the data collection:
1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D.

X. References

[1] Kaggel heart diseases dataset https://www.kaggle.com/ronitf/heart-disease-uci

[2] North BJ, Sinclair DA. The intersection between aging and cardiovascular disease. Circ Res. 2012;110(8):1097-1108. doi:10.1161/CIRCRESAHA.111.246876

[3] UT Southwestern Medical Center. (2018, January 8). Proper exercise can reverse damage from heart aging. ScienceDaily. Retrieved May 29, 2020 from www.sciencedaily.com/releases/2018/01/180108090132.htm

[4] S. Bashir, U. Qamar and M. Younus Javed, "An ensemble based decision support framework for intelligent heart disease diagnosis," International Conference on Information Society (i-Society 2014), London, 2014, pp. 259-264, doi: 10.1109/i-Society.2014.7009056.

[5] American Heart Association. (2015). Know Your Target Heart Rates for Exercise, Losing Weight and Health. Retrieved May 30, 2020, from https://www.heart.org/en/healthy-living/fitness/fitness-basics/target-heart-rates

[6] Sandvik, L., Erikssen, J., Ellestad, M., Erikssen, G., Thaulow, E., Mundal, R., & Rodahl, K. (1995). Heart rate increase and maximal heart rate during exercise as predictors of cardiovascular mortality: a 16-year follow-up study of 1960 healthy men. Coronary artery disease, 6(8), 667–679.