Predicting heart Disease with ML

By Yousef Taheri
March 2020

Literature Review

- Y. Alp Aslandogan, et al. (2004)
- Niti Guru, et al. (2007)
- Resul Das, et al. (2009)
- Mai Shouman, et al. (2012)
- Chaitrali S Dangare, et al. (2012)
- Sudha Vijiyarani, et al. (2013)
- Jaymin Patel, et al. (2015)

Danger of heart disease

-Heart disease is the biggest killer of both men and women around the world.

-WHO analysed that twelve million deaths occurs worldwide due to Heart diseases.

-Heart attacks, strokes and other circulatory diseases account for 41% of all deaths (European Public Health Alliance 2010).

-In almost every 34 seconds the heart disease kills one person in world.

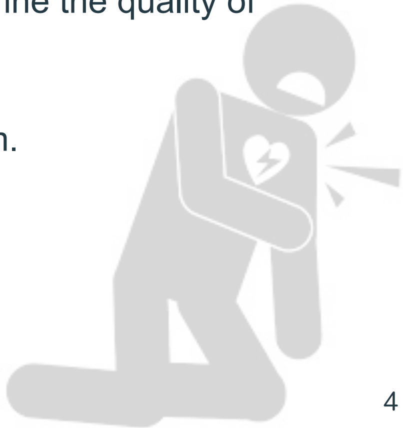
-Different person body can show different symptoms of heart disease which may vary accordingly (Naganna Chetty et al. 2015).

Hard to diagnose

-Different person body can show different symptoms of heart disease which may vary accordingly (Naganna Chetty et al. 2015).

-Diagnosing the disease correctly & providing effective treatment to patients will define the quality of service(K Sudhakar, et al. 2014).

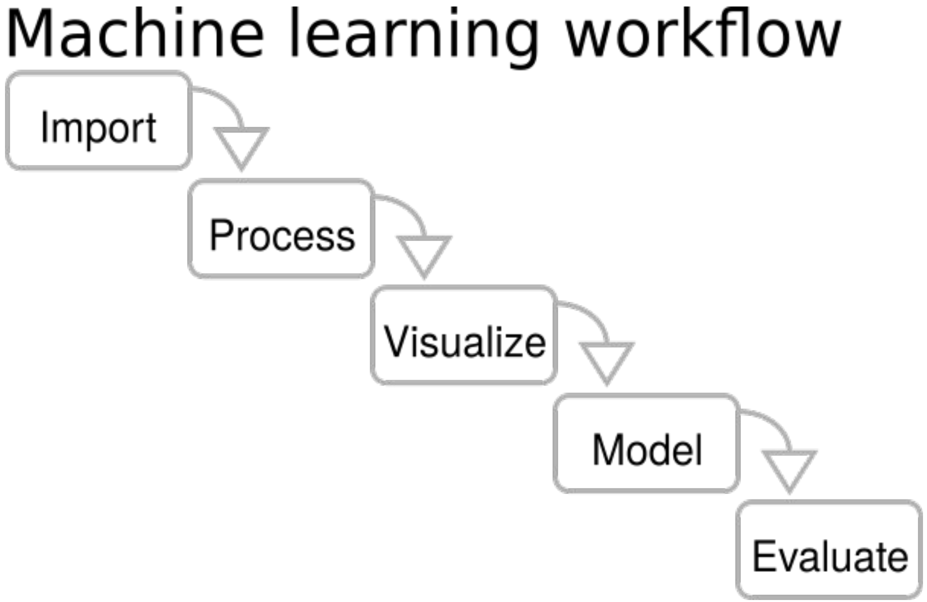
-Heart expert’s create a good and huge record of patient’s database and store them.



The rule of Machine Learning

-Researchers make use of several ML techniques to help the specialists identify the heart disease.

-ML is the use of software techniques for finding patterns and consistency in sets of data.



Commonly used procedures

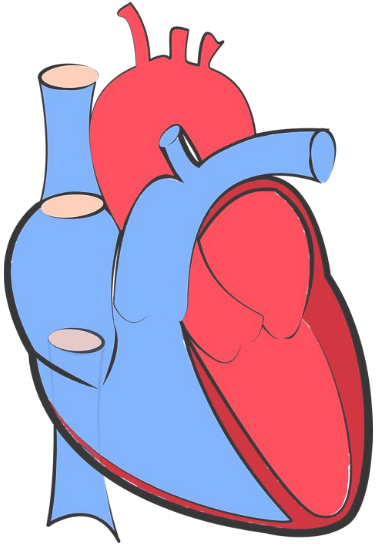
- decision tree
- K-nearest
- Naïve Bayes
- bagging algorithm
- neural networks
- SVM (Support Vector Machine)

Heart disease dataset

-Cleveland dataset from UCI repository is used, which is available at: http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

-The dataset has 14 attributes and 303 records.

-The only dataset that has been used by ML researchers to this date.



The dataset attributes

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	0 = female, 1 = male
Cp	Discrete	Chest pain type: 1 to 5
Trestbps	Continuous	Resting blood pressure
Chol	Continuous	Serum cholesterol
Fbs	Discrete	Fasting blood sugar>120 mg/dl: 1-true 0=False
Exang	Discrete	exercise induced angina (1 = yes; 0 = no)

The dataset attributes

Name	**Type**	**Description**
Thalach	Continuous	Maximum heart rate achieved
Restecg	Discrete	resting electrocardiographic results: 0, 1, 2
Oldpeak	Continuous	ST depression induced by exercise relative to rest
Slope	Discrete	the slope of the peak exercise ST segment
Ca	Continuous	number of major vessels (0-3) colored by flourosopy
Thal	Discrete	3 = normal; 6 = fixed defect; 7 = reversable defect
Num	Discrete	diagnosis of heart disease (angiographic disease status): 0, 1

Preprocessing steps

Preprocessing includes feature selection, imputing missing values and transforming data

Since number of features is not a problem in our case I didn’t perform a feature selection method

my preprocessing steps include:

-imputing missing values by mode or mean

-converting categorical features to dummy variables

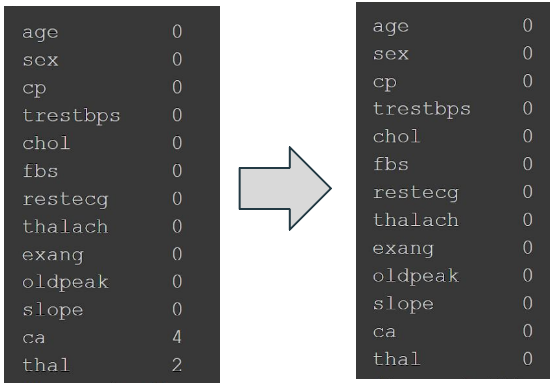
-Transforming data by removing the mean and scaling to unit variance

Preprocessing
Imputing missing values

-In our data dataset there are 4 missing values in “ca” and 2 missing values in “thal” features.

-Since ‘ca’ is continuous variable I impute it with it’s mean

-And because ‘thal’ is categorical variable I impute it with columns mode



```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          4
thal        2
```

```
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
```

Number of missing values for each feature before and after imputing

Preprocessing
Handling dummy categorical features

-In this step I convert each categorical variable to dummy variables i.e indicator vectors with length equal to the number of categories.

-Categorical variables in the data set are :`['sex','cp','fbs','restecg','slope','exang','thal']`

-After this transformation data dimension increase to 18.

Preprocessing scaling

-Scaling data usually helps classifier to perform better, especially for Neural Networks with tanh activation functions.

-For this dataset I scaled the data by removing the mean and dividing by variance

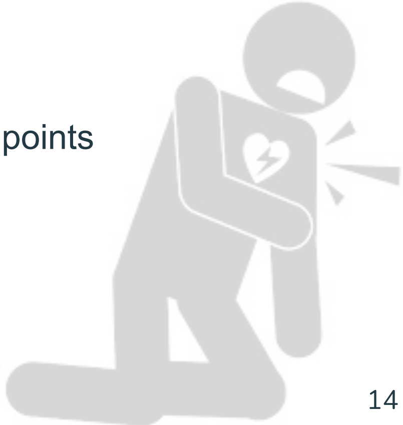
Visualizing data

Visualization helps to get a notion of how data points are scattered in space, and get a sense of noise, since our data set is basically in a 14 dimensional space we have to map the points to lower dimensions to be able to plot them,

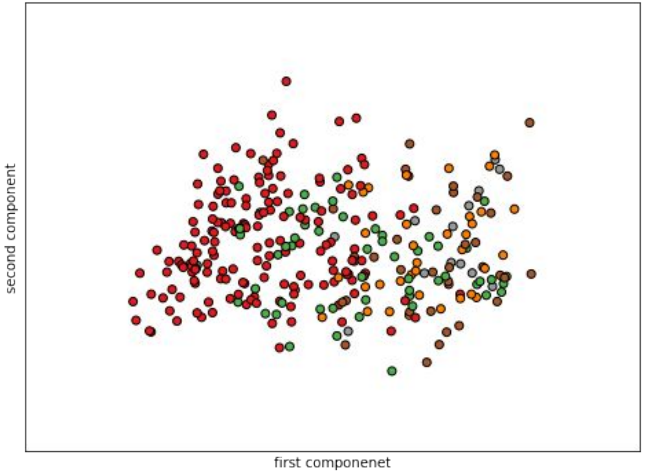
-Dimension reduction techniques like PCA can help us to do it.

So first we use PCA to map the points to lower dimensions 2 or 3 and then plot the points

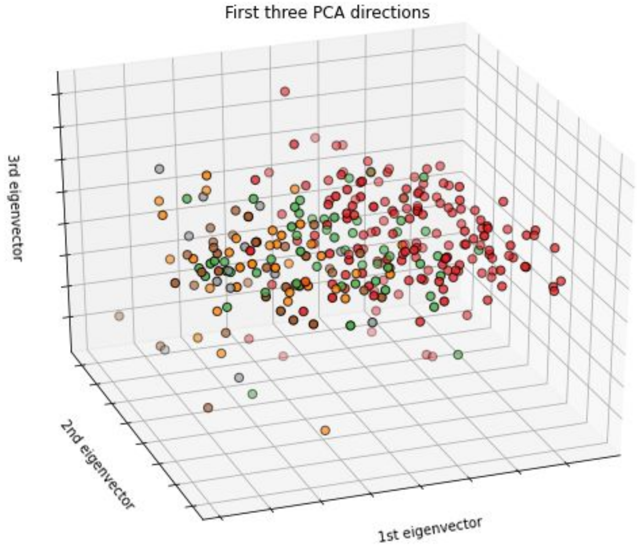
-points color are based on their class



Visualizing data



Data points in 2 dimensions



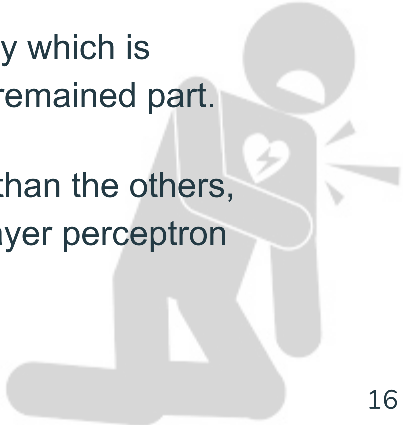
Data points in 3 dimensions

ML classifiers

-There are plenty of ML classifiers which can tested on this dataset. For this data set I tried many different classifiers including: Decision Tree, Random Forest, Gradient Boosted Trees, Linear SV, KNN and MLP.

-To compare the performance of the models I used 10 fold cross-validation accuracy which is average model accuracy on dataset by being trained on 9 parts and testing on the remained part.

-Among all the tried models I chose to only present 4 models which perform better than the others, these models are Gradient Boosted Trees, Linear SVC,Random Forest and Multi-layer perceptron



ML classifiers
Tuning parameters

-To find the best set of parameters for each classifier I used Grid search and manual parameter tuning.

-The accuracy for each model is based on the best set of tuned parameters

ML classifiers performance

Here I compare the models with classification accuracy

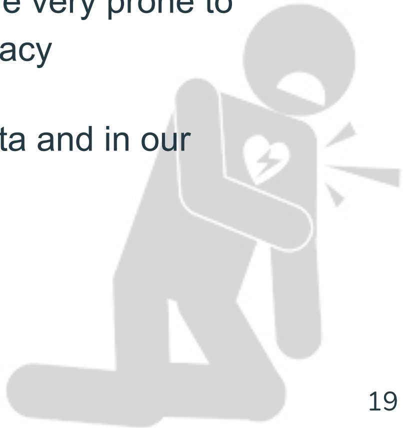
Model	**Parameters**	**CV accuracy**
XGBoost	max_depth=1, min_samples_leaf=1, min_samples_split=2, n_estimators=60	0.607
Linear SV	C=1, penalty="l2"	0.594
Random Forest	criterion='gini', min_samples_leaf=1, min_samples_split=2, n_estimators=100	0.583
MLP	activation='tanh', batch_size=50,max_iter=2000, hidden_layer_sizes=(20,10)	0.567

ML classifiers results

-XGBoost performed better than all the other models, this model is pioneer model especially in Kaggle competitions

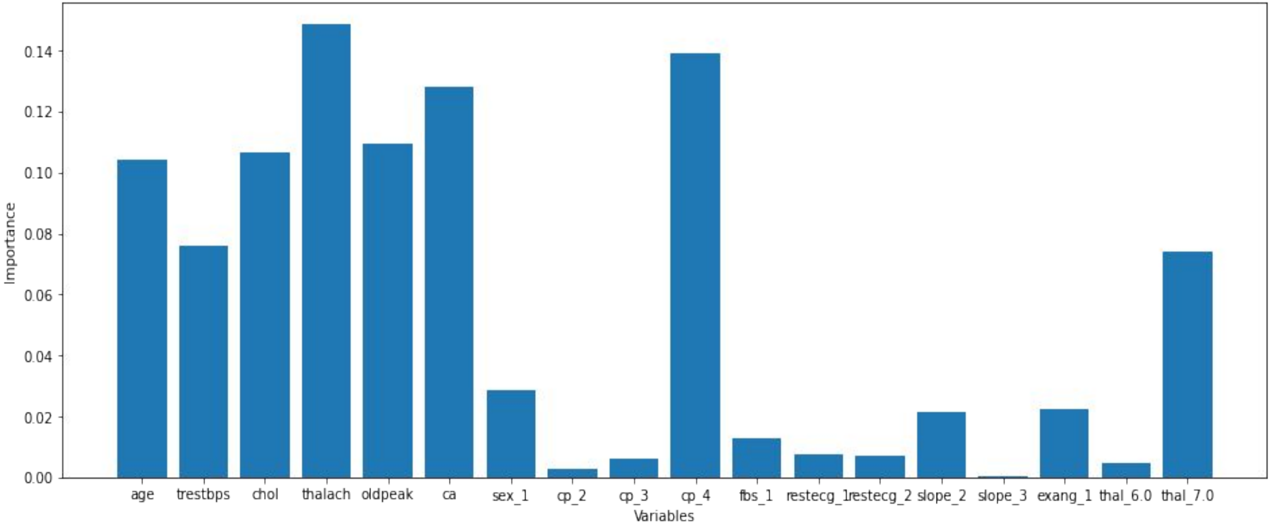
-Linear SV performed very well comparing to others, this show that other models are very prone to overfit on this dataset and with regularization methods we can achieve better accuracy

-MLP didn’t performed well, because usually deep learning models need a lot of data and in our case there are no much records in the data set



Feature importance

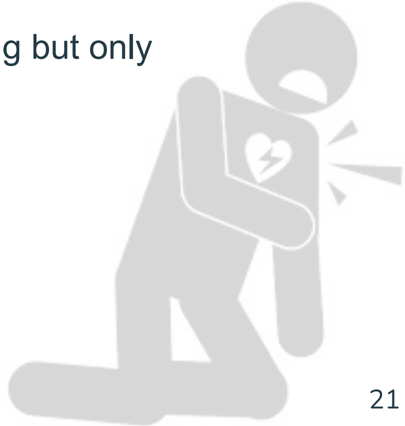
To get a sense of importance of each feature we can take a look at feature importance of XGBoost model



Clustering methods

-Since this each record in this data is related to patient, it might be interesting to see how similar or close are the patients with respect to the recorded attributes. To answer this question we can implement clustering methods and compare them to see which one is a best fit for this data set.

-I tried Kmeans, Spectral clustering, Gaussian mixture and Agglomerative Clustering but only compare the first 3 of them by the homogeneity score



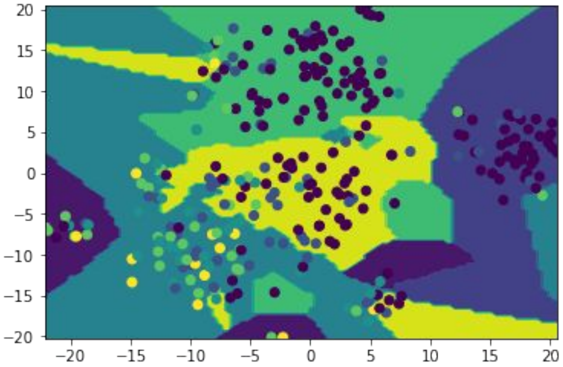
Clustering methods comparison

Here I compare all 3 models with respect to homogeneity score

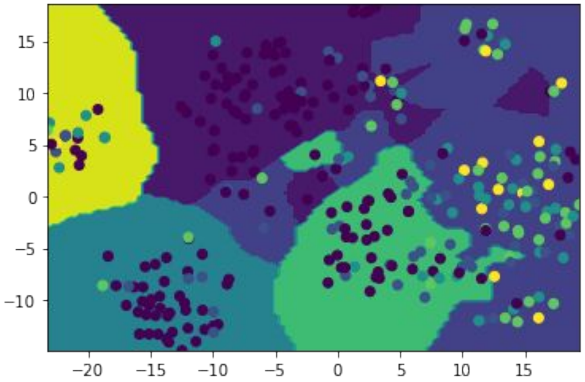
Model	Parameters	Homogeneity score
Kmeans	n_clusters=5, init='k-means++', max_iter=100	0.189
Spectral clustering	n_clusters=5, eigen_solver='arpack', affinity="nearest_neighbors"	0.207
GM	n_components=nClust, covariance_type='spherical'	0.17

Clustering
Visualizing clusters in 2D

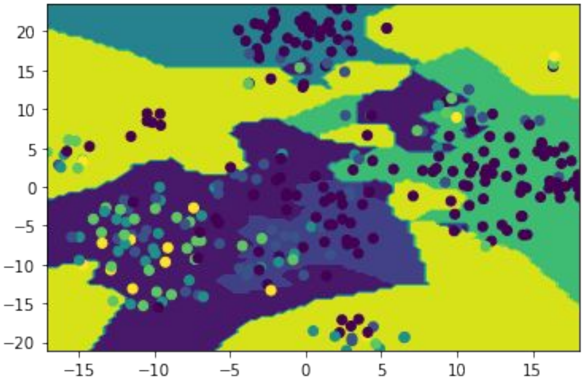
Here I visualized the clusters in 2 dimensions



k-means



Spectral clustering



Gaussian Mixture

Clusterin Results

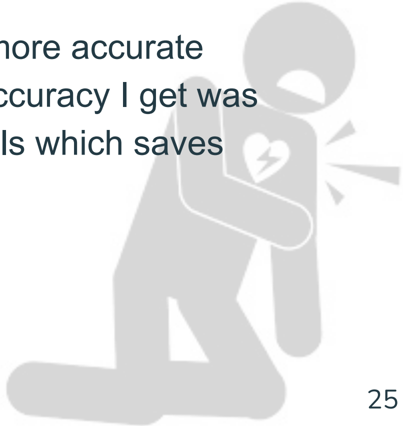
-Clustering result shows that Spectral clustering finds better clusters which are more close to the labels

-In other words it shows that, it clusters the data better based the presence of heart disease in patients.

Conclusion

-In conclusion, I believe only a marginal success is achieved in the creation of predictive model for heart disease patients and hence there is a need for combinational and more complex models to increase the accuracy of predicting the early onset of heart disease.

-Also with pregresses in IoT we can monitor patients body in real-time and collect more accurate data which help us to train better models. Here I only used 14 variables and best accuracy I get was around 60% but in the future and with more data we can build more accurate models which saves thousands of lives



References

- Aslandogan YA, Mahajani GA. Evidence combination in medical data mining. International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. 2004 Apr 5 (Vol. 2, pp. 465-469). IEEE.
- Chetty N, Vaisla KS, Patil N. An improved method for disease prediction using fuzzy approach. In2015 Second International Conference on Advances in Computing and Communication Engineering 2015 May 1 (pp. 568-572). IEEE.
- Dangare CS, Apte SS. Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications. 2012 Jun;47(10):44-8.
- Das R, Turkoglu I, Sengur A. Effective diagnosis of heart disease through neural networks ensembles. Expert systems with applications. 2009 May 1;36(4):7675-80.
- Dangare CS, Apte SS. Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications. 2012 Jun;47(10):44-8.
- Guru N, Dahiya A, Rajpal N. Decision support system for heart disease diagnosis using neural network. Delhi Business Review. 2007 Jan;8(1):99-101.

- Patel J, TejalUpadhyay D, Patel S. Heart disease prediction using machine learning and data mining technique. Heart Disease. 2015 Sep;7(1):129-37.
- Shouman M, Turner T, Stocker R. Using data mining techniques in heart disease diagnosis and treatment. In2012 Japan-Egypt Conference on Electronics, Communications and Computers 2012 Mar 6 (pp. 173-177). IEEE.
- Sudhakar K, Manimekalai DM. Study of heart disease prediction using data mining. International journal of advanced research in computer science and software engineering. 2014 Jan;4(1).