International Journal of Engineering &Technology, 7 (3.12) (2018) 750 -753

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Heart Disease Prediction

S.Vinothini[1], Ishaan Singh[2], Sujaya Pradhan[3], Vipul Sharma[4]

[1]Assistant Professor, Department of IT, SRM IST
[2,3,4,]Student, SRM IST, SRM University

Abstract

Machine learning algorithm are used to produce new pattern from compound data set. To cluster the patient heart condition to check whether his /her heart normal or stressed or highly stressed k-means clustering algorithm is applied on the patient dataset. From the results of clustering ,it is hard to elucidate and to obtain the required conclusion from these clusters. Hence another algorithm, the decision tree, is used for the exposition of the clusters of . In this work, integration of decision tree with the help of k-means algorithm is aimed. Another learning technique such as SVM and Logistics regression is used. Heart disease prediction results from SVM and Logistics regression were compared.

1. Introduction

In medical fields to discover some significant disease such as HIV ,cancer, heart disease which are the main cause of death throughout the world, machine learning can be used to solve these problems. And to predict these type of disease is of great consequence to research and application level.
In this paper, machine learning algorithm is used to detect heart disease by using patient's medical record. We use the Machine Learning Repository from UCI to get dataset for heart disease patient for both training & testing. We used x patient's record and which have 14 attributes age , sex ,chest pain ,resting blood pressure, serum cholesterol ,fasting blood pressure, . Result which we get are divided into 4 major parts.0 represents than the patient doesn't have heart disease ,1 represents that patient have small possibility of having heart disease ,2 and 3 are combined and represent that patient surely has heart disease he should visit doctor immediately.
Our objective is to cluster heart beat result using patient data by applying k means .Then we use this cluster as an input to decision tree algorithm to interpret result accurately. Decision tree indicates whether that particular patient is normal, stresses and highly stresses. SVM classification and logistic Regression model to predict heart disease severity. Finally accuracy result of SVM and logistic regression is compared.

2. Attribute & Preprocessing

Extricate Viable Columns

There are 76 columns & 76 features in the data set. We use 14 attributes from these 76 attributes .Remaining 62 columns are disregard. We filtered the rest 14 columns and use them in our project.

3. Existing System

Heart disease is a largest cause of death in majority of countries. In 2012, according to The World Health Organization (WHO) about 12 million passing has arised globally, each and every year due to the Heart diseases. In India about 1.7 million people have died due to heart diseases in 2016 according to the Global Burden of Disease Report which was released in september 2017. The methods to avoid heart attack is not accurate as there is no proper symptoms to it through which we can identify when a person is going to have a heartattack.

4. Models

4.1 K means Algorithm

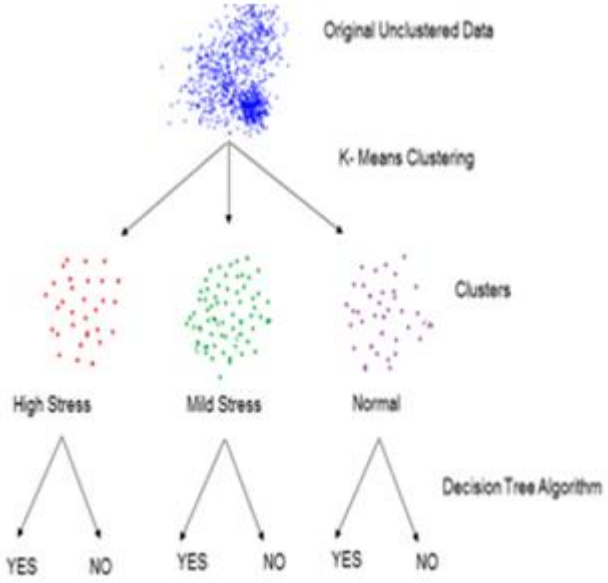
When it comes to clustering problems K means is the most straightforward algorithm which can be used. It is used for clustering the data set into k no. of clusters and then find centroid for each cluster.
Patient's data set is very large so to get the output accurately, we divide the data set into 3 clusters. We divided the data set into clusters on the basis of there stress value ,1st cluster contains the people which are normal, 2nd which are stressed and 3rd which are highly stressed .Further k means give us centroid for each cluster.

```
Centroids Value
[[1.92879630e+02 5.36296296e+01 6.57407407e-01 3.17592593e+00
  1.30074074e+02 2.47351852e+02 1.48148148e-01 9.53703704e-01
  1.48453704e+02 3.78370370e-01 7.81481481e-01 1.52777778e+00
  7.77777778e-01 4.52777778e+00 9.25925926e-01]
 [9.83437500e+01 5.82812500e+01 4.37500000e-01 3.15625000e+00
  1.37000000e+02 3.36218750e+02 2.18750000e-01 1.25000000e+00
  1.53968750e+02 3.12500000e-01 1.12500000e+00 1.56250000e+00
  6.56250000e-01 4.75000000e+00 8.75000000e-01]
 [6.38703704e+01 5.35555556e+01 7.68518519e-01 3.18518519e+00
  1.30861111e+02 2.29083333e+02 1.29629630e-01 1.11111111e+00
  1.51148148e+02 3.14814815e-01 1.27129630e+00 1.64814815e+00
  6.28370370e-01 4.87037037e+00 9.72222222e-01]]
```

Now we run decision tree algorithm on these 3 clusters.

4.2 Decision Tree Algorithm

After forming clusters, these clusters are the input for decision tree algorithm. It produces decision rules at the output.
Decision tree creates a tree structure to classified data as yes -> prone to heart disease or no ->not prone to heart disease. For each cluster decision tree classify data as yes or no.



4.3 Multiclass SVM

SVM is used to construct a set of hyperplane in a high dimensional space which is used in classification, regression an in other tasks.
In our project we use seaborn data visualization to provide a high level interface for drawing statistical graphics .PLT.SHOW() is used to show seaborn plots. To use seaborn with matplotlib SET_CONTEXT() and SET_PALETTER() is used. We use correlation matrix for SVM classification. SVM can give correlation coefficient for each of the column in stated matrix. For example ith entry measure correlation between ith column and the jth column of the given matrix. The inputs in the diagonal of correlation matrix are same as it is used to compute the correlation of themself. Reason of symmetric because the correlation between the ith and jth column is as the correlation between jth and ith column. In our code we used function called math.show() which has attributes called **corellation**,**vmax** and

Vmin with initialisation.
Different functions used in scatter plot –
Colourbar() : A function use with images and counter plots. Here we are using this for plotting correlation matrix with colour bar legend.
Colourbar(cax),cax : The axe instances in which colour bar is drawn.

Ticks=np.arange(0,9,1) : Return th x ticks as a list of location from 0 to 9th column.
Ax.setxticks(ticks),ax.setyticks(ticks),ax.setxticklabels(names),ax.setyticklabels(names) :These 4 ticks are using for plotting correlation matrix of heart patients.

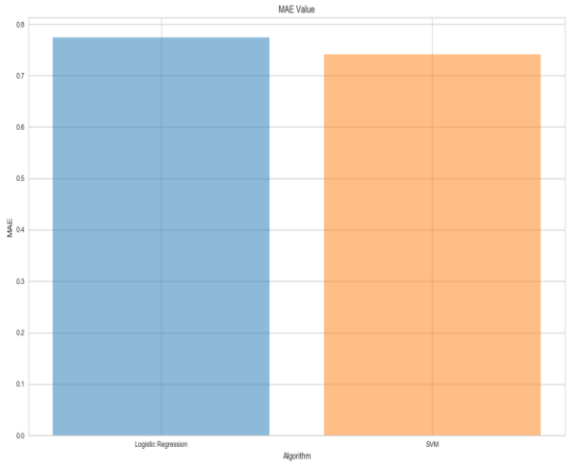
5. Logistic Regression

Linear SVMs and Logistic regression generally perform comparably in practice. There are two main ways to perform linear regression in python – with Stats models and scikit-learn
Stats models provides classes and functions for estimation of many different statistical models.
Predict(XX) predicts the y using the linear model we fitted.
Logistic regression is a S-shaped curve whose input is real value number and the output is between 0 and 1.

1/(1+e^-value)

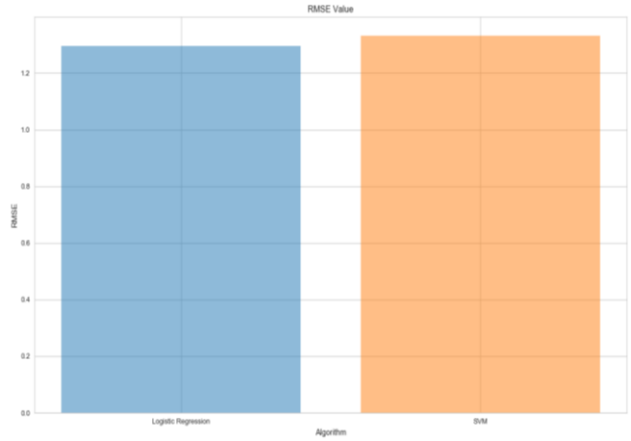
It is a probability function in which input belongs to any one of the various classes(classification).
To check the error in algorithm we used different features -:
Mean Absolute Error (MAE): MAE is used to calculate the mean magnitude of errors in a deck of predictions excluding there directions. If the absolute value is not taken, the mean errors becomes the Mean Bias Error (MBE) and is used to calculate mean model bias .

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

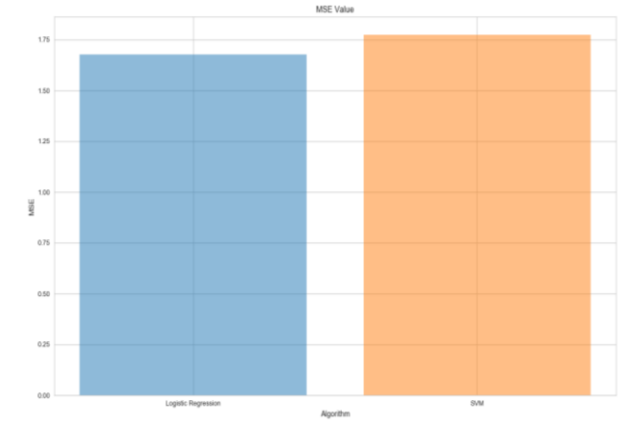


Root Mean Squared error (RMSE): RMSE is a quadratic numbering rule that also calculate the mean weight of error.
By using RMSE high weight can be given to large errors .RMSE functioning is more efficient when it comes to large errors. In our output also RMSE value is more accurate than MAE value as compare to predictive data. It is not necessary that with increase in error variance of RMSE increase. When the variance of frequency distribution of error magnitude increases RMSE also increases.

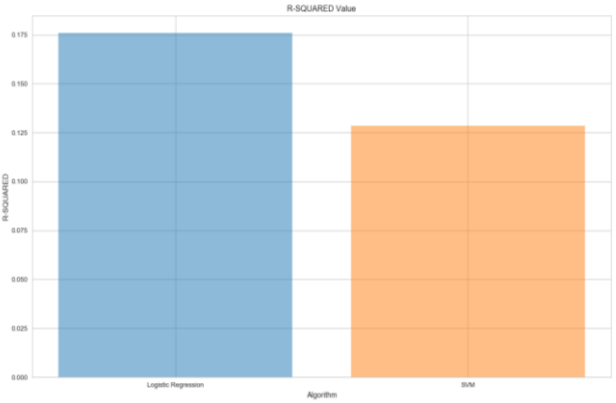
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$



Mean Squared Error (MSE): MSE tells the distance between regression line and set of points .It calculate the distance between points and regression line and than squaring them. The squaring is necessary to remove the negative signs. The smaller the MSE the closer you are to finding the best fit line. Value of MSE is always positive and the values which is closer to zero they are better Predictor and Eliminator are the two values of MSE

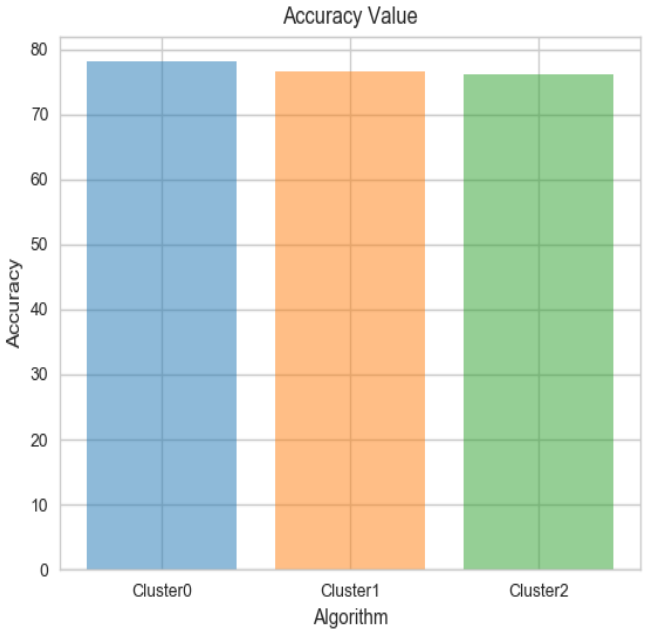


R-Squared: Function of r square is to know the distance between the regression line and the data points. Other name of r square is coefficient of regression or the coefficient of multiple regression.
R-squared=Explained variation /Total variation
The output value of R square is between 0% - 100%.Best model which firs your data has the highest R square value.

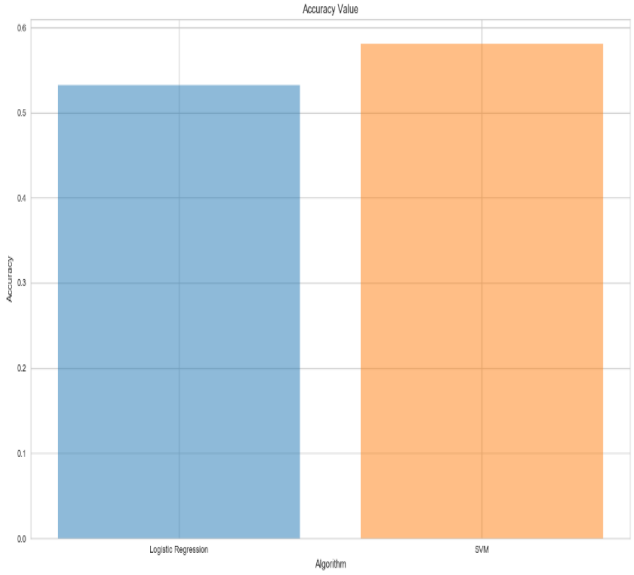


6. Feature Selection and Results

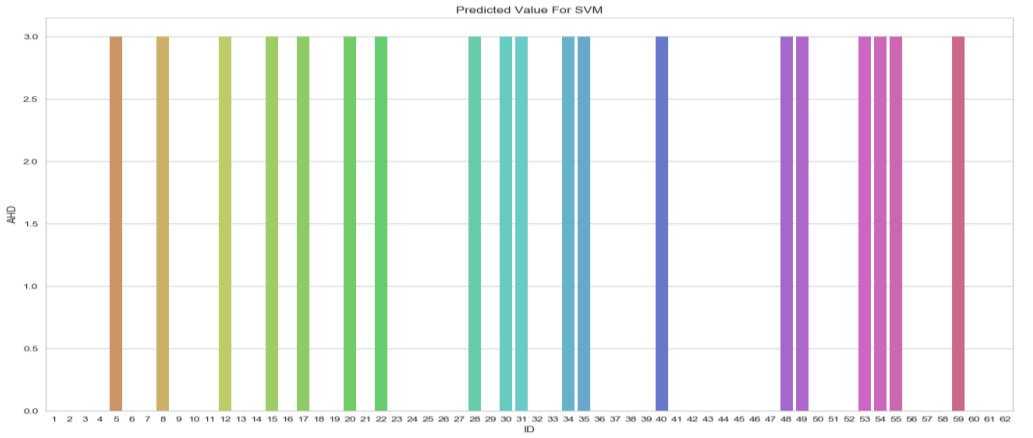
Accuracy is the vital role when we are going to predict something. In this project we use k means algorithm to make three clusters. Out of three clusters cluster0 has the most accuracy.

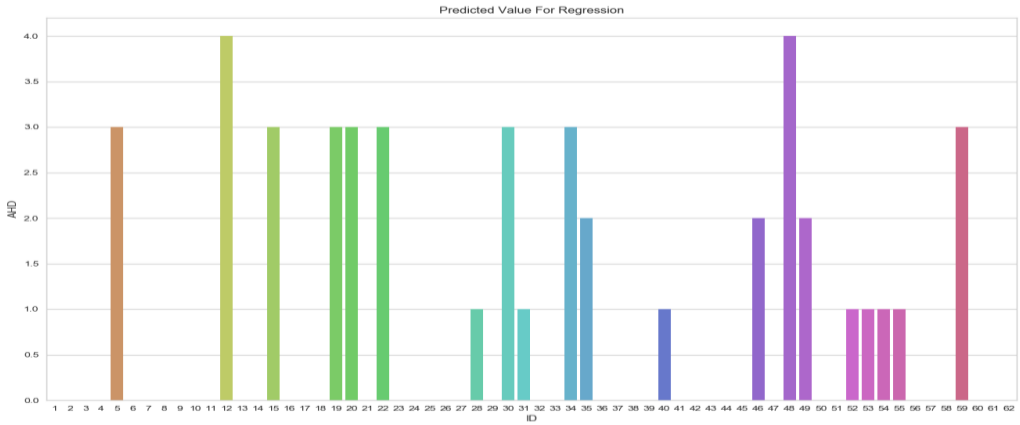


We also calculate accuracy between SVM and Logistic regression. By calculating it we get the output in form of graph in which the accuracy of SVM is greater than the accuracy of logistic regression.



The result which we get for heart disease prediction value using SVM and logistic regression are given in graphical representation below:





7. Future Enhancement

Further works involves development of the system using the mentioned methodologies and thus training and testing it. Other work is to develop a tool which can be used to predict probability of risk of a patient. This tool can be used further on other models such as neural networks, assemble algorithms, etc. In future, execution can be improved by genetic algorithm which used for AI which provide Solution for optimization and search problems for feature selection .Further this system can be enhanced using Swarm intelligence technique to decide the more weight age input parameters.

References

[1] Mackay,J., Mensah,G. 2004 "Atlas of Heart Disease and Stroke" Nonserial Publication, ISBN-13 9789241562768 ISBN-10 9241562765.

[2] Robert Detrano 1989 "Cleveland Heart Disease Database" V.A. Medical Center, Long Beach and Cleveland Clinic Foundation.

[3] Yanwei Xing, Jie Wang and Zhihong Zhao Yonghong Gao 2007 "Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease" Convergence Information Technology, 2007. International Conference November 2007, pp 868-872.

[4] Jianxin Chen, Guangcheng Xi, Yanwei Xing, Jing Chen, and Jie Wang 2007 "Predicting Syndrome by NEI Specifications: A Comparison of Five Data Mining Algorithms in Coronary Heart Disease" Life System Modeling and Simulation Lecture Notes in Computer Science, pp 129-135.

[5] Jyoti Soni, Ujma Ansari, Dipesh Sharma 2011 "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" International Journal of Computer Applications, doi 10.5120/2237-2860.

[6] Mai Shouman, Tim Turner, Rob Stocker 2012 "Using Data Mining Techniques In Heart Disease Diagnoses And Treatment" Electronics, Communications and Computers (JECECC), 2012 Japan-Egypt Conference March 2012, pp 173-177.

[7] Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H. Guppy, Stella Lee, Victor Froelicher 1989 "International application of a new probability algorithm for the diagnosis of coronary artery disease" The American Journal of Cardiology, pp 304-310.15

[8] Polat, K., S. Sahan, and S. Gunes 2007 "Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing" Expert Systems with Applications 2007, pp 625-631.