Analysis of Heart Patients Disease Using Data Mining Tool Orange

Danish Umer

June 24, 2020

Analysis of Heart Patients Disease Using Data Mining Tool Orange

Data Science, (Professor)
RIPHAH University, Dept. of Information Technology (IT), and Computer Science (CS) Lahore Township
Danish Umer, (Student)
Dept. of Information Technology (IT), Computer Science (CS) RIPHAH University,
Lahore Township.

***Abstract*— in the study of health care is very important now a days in human life. In medical science and their related areas are health concern business has become a notable field in the wide spread area. Health sector are generating lot of information and data which help to understand and need to be analysis, theses data are must convert into meaningful data. To use these patient's information, make future decision and achieve effective decisions, these decisions help to overcome patients to admit hospital and use expensive treatment. Use patient's data and implement data mining techniques to learn patients' patterns and find solution to overcome these diseases. Be that as it may, there is an absence of examining instrument as per furnish compelling test results together with the covered-up data, so and such a framework is created utilizing information digging calculations for characterizing the information and to recognize the heart illnesses. In Healthcare problems data mining provide solution. For heart diseases patient there are 4 algorithms which help to find patterns and solution these are Random Forest, SVM, KNN, Logistic Regression and Naïve Bayes algorithm which help to diagnosis heart issues. In this research paper I am using data mining tool Orange and analyzes parameters and find prediction on heart patients' diseases and along these lines proposes a heart ailment forecast framework (HDPS) put together aggregate with respect to the data mining approaches.**

Keywords*— Data Mining, Data Mining Classification Techniques, SVM, KNN, Random Forest, Naïve Bayes, Orange Tool, Heart Disease.

I. Introduction (*Data Mining*)

Data Mining is the process of finds patterns in huge data sets to include methods to the intersection database systems , ML (Machine Learning) and Statistics. DM Data Mining is the steps of analysis KDD [1] or the (Knowledge Discovery in Databases). Data Mining (DM) is extract useful data from huge datasets and implement techniques to convert visualization data. Extracting of data from huge data set and make patterns and visualization of their pattern to make decisions this result is diagnosis of diseases is very important. DM can stand use to extracting knowledge by predicting and analyzing of some diseases. Health caring department have their large number of potentials according to find the hidden visualization patterns among datasets in medical field. Data Mining plays an important role of Heart Disease prediction. In heart patient need more or more test to diagnoses and predict future plans but if we implement some data mining techniques, they can reduce the number of patient's tests. They reduce test cost and plays significant role in time and performance also reduce cost. In Health field Data Mining is a very important because through data mining allows the doctors to see which attributes of patients are important to diagnoses some are: patient's symptoms, their age and weight etc. These things help to doctors they easily diagnose the disease much efficiently. Discovery of knowledge in raw data is the method of extract useful data, patterns and convert into useful information mining. After extracting information, its makes use of different algorithms use and implement and derived patterns bye the knowledge world in database method in Figure 1.

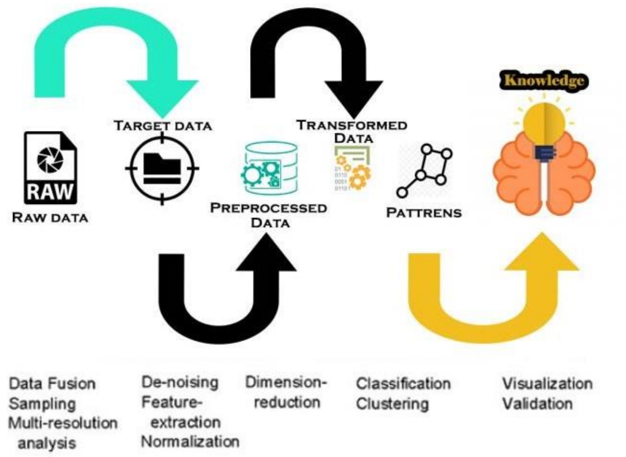

Figure 1: KDD (Knowledge Discovery within Databases) [1]

Different stages concerning information revelation of databases technique are portrayed as chases. In stage of selection of that the different data resources. Different stages concerning information revelation of databases technique are portrayed as pursues. In stage of selection of that the different data resources. Wanted, blank and noisy data are removing is in preprocessing stage and outfitted the clean information which executes the design in understanding counting a regular configuration of change organize. After that data mining techniques are applied at long last, into the between the connotation arrange, that will display the outcome after end-client in an important way.

II. *Data Mining* Techniques

There many types of techniques but popular techniques are:

1. **Classification learning**
2. **Numeric prediction**
3. **Association rule mining**
4. **Clustering**

Classification learning: In Statistics, ML Machine Learning is a supervised learning technique in which they learn program of computer where data is input given to it after that they use to learn the classify new observations.

Numeric prediction: is tied in with getting the following an incentive in the stream and here one isn't trying to forecast the class rather the worth or abilities.

Association rule mining: ML machine learning is Rule Based learning which is discovering relations between variables in large DB and it is intended to identify very strong rule occurred in DB.

Clustering: is grouping the set of objects which are same size that the objects are in same group is called clustering.

The four type of learning methods are need to select which one is better to other and performs much better to the others. Data mining methods are depending on which data are most suitable for and which techniques are use and solving issues in data mining on these types. Data is to stand and use the selection of the choice of information mining procedure which is generally reasonable for the information utilized.

III. *ML(Machine Learning)*

Ml (Machine learning) is data science method which is computer program is learn from very past experience and develop the algorithms that helps to learn by own without little or no human interfere. In ML there is huge amount of task they do like decision making, classification and prediction.

AI originates from man-made consciousness examine and has become a fundamental part of information science. ML starts with input so a preparation data index. In this stage, the Machine Learning calculation utilizes the preparation dataset in the wake of gaining from the information and structure designs. The Phase in learning the model which is train and after that we apply test data on it. Test data are generated from or datasets and make it apply on training sets to check accuracy which they gave results and after that we are analysis on it.

The general extraction in favors to the test dataset shows the model's capacity incompatibility with playing out its assignment against information. ML gives out past a statically coded set with respect to announcements into articulations, so a great deal are increasingly created based as respects the info information.

IV. *OPEN SOURCE SOFTWARE*

Open source software's are usually are freely use and available on internet they easily use and modification are available. Open source has, in the brains in regards to many, come to be synonymous with free software (Walters, 2007). If anyone know about develop extension then they customize it and use it open source software. Charging an expense for specific activities is regularly disallowed by utilizing an open permit understanding whereby any alterations to the source code consequently become open space. Most of people around the globe can develop open source software to helps peoples.

V. *HEART DISEASES*

Heart issues are very common in now a day. So, it is very important to find the prediction in initial stage, to overcome death rate if we find heart issue in early stages, but it is to difficult to find heart issues in initial stage and most of the doctors are not properly find situation because heart patients are having my symptoms. So, doctors adopted many methodology and scientific technologies to identify. These technologies lead to diagnosing not only fetal diseases but also many diseases. The Successful patient's treatment leading to accurate diagnosis & continually attributed to right. Sometimes the doctors are failed to find accuracy and his assuming is not right and his decisions are not accurate and it is very dangerous for heart patients. Ml are used to find accurate decisions for heart patients [8].

VI. *HEART DISEASE DATASET*

The datasets are used for working is ML Repository which heart patient's data sets which have some patients test records the file is used for orange tool is .tab extension. These data sets are freely having in kaggel. These datasets have 300 instance and 70 attributes. We use only 14 attributes. In my case study 14 attributes are enough to use and implement ML algorithms they also help to lead to find solutions. 14 attributes are used to find factors for prediction of heart disease [8]. Data sets have missing values and we prepresses it and convert into tool using extension.

VII. *OVERVIEW OF DATA MINING TOOLS*

There are huge number of data mining applications which is use in marketing some of paid tools and some or open source tools. These applications are going from promoting and publicizing about products, Al research, biological sciences and products there is also use in Crime reports to predict crime scene in High level Govt intelligence. Due to huge use of application of DM and creating information application of DM and vast number of free tools are use and these tools are creating to manage data mining. Every tool has their disadvantage and advantages.[6] In market and internet there are open source tools are available they also dividing into groups have been developed by a RC (Research Community). The model of open source development model is use to not a necessarily supported not only single society but its internationally contributed team to developed open source developed model for use. These works is lead to help to extract data from raw databases and many techniques and tool are used for this purpose. DM Data mining tools are used to find behavior, visualization or patterns and trends to proactive knowledge to make decisions on it. DM like problem need to very powerful tools and the development concerning DM. To growing tool and their selection tool options are increasingly difficult because there are many tools in market. [7] There are many tools but some of our explain in below.

There are many tools in the market and internet but i use orange tool to use to perform DM data mining techniques. This tool is available in open source and they are user friendly tool. The selecting tool is the first methodology which is number of available open source DM tools in with being tested. There are many free tools are available on internet and after many searches I found Orange tool for my case study. They are very powerful tool of python.

VIII. About Orange Tool

If we talk about Ml open source tool then orange is best open source tool which is use in my case study. This data mining tool which is use in ML Machine Learning Technology. Orange Tool is user friendly tool and they help to drag and drop option in his interface. They also provide visualization of data [9]. Numerous examinations is conceivable by means of its visual programming interface that is move related with gadgets and numerous visual apparatuses will in general be upheld, for example, bar graphs, trees, scatterplots diagram, den programs and heat maps. Orange is frequently a part organized information mining just as AI programming suite made in the python language. Orange is regularly a very skilled open-source perception just as a gathering of information mining (Data Mining) instruments alongside an easy to use [10],[11].

IX. Comparative Study

Our Comparative Study is to get raw data from any free source and then Appling data mining algorithms and collecting data after we are tested and select which data sets are use. After selecting data sets next is according to our test, we select classification Algo test tools and performance. That the overall research which i am followed in shown in Figure 2.

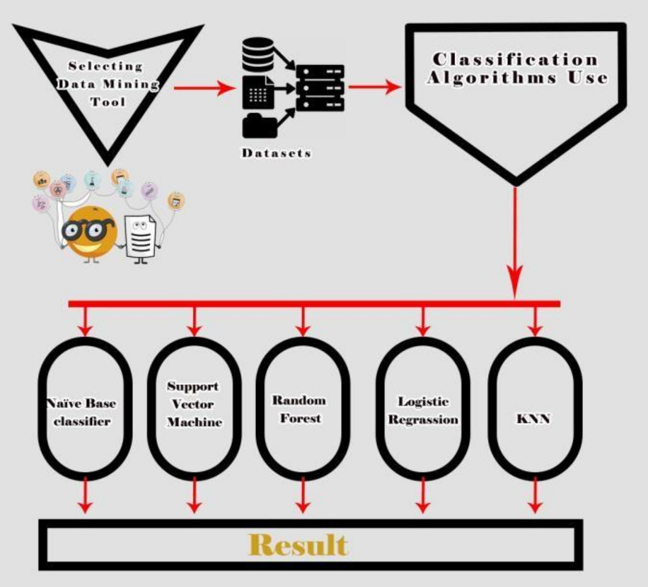

Figure 2: Tools Implementation Methodology [2]

X. Research Question

In my case study is based on prediction on heart patients' diseases with using Orange tool and identify patterns and find which classification algorithms are best and have more accurate then other my research questions are:

- **Which Classification algorithm are best?**
- **Using Orange tool which classification algorithm are accurate in heart patients?**

Recall and Precision

a) Precision

It is known as predictive positive value. They provide Avg probability of relevant retrieval of data.

Precision is denoted as = P

Number of true positives denoted as = P+ve

False positives denoted as = FP-ve

Formula: P= P+ve /(P+ve + FP-ve)

b) Recall

Recall It is characterized as is average probability of complete retrieval this is called Recall.

Recall denoted by = R

True Positions by = Tp

False negative by =F-ve

Formula: R=Tp/(Tp+F-ve)

c) Naïve Bayes

Naïve Bayes Classifier method are suited particularly when dimensionality of the input’s values is high. Naïve Bayes are not suitable when all attributes are independent on each other, it is also hard to understand and debug [2] decision tree perform poorly. This Algorithms are used in computer version and robotics. For Our data sets heart disease comparative analysis of precision and recall analyzing Orange Tool Precision 0.828

and Recall 0.828, accuracy 0.907.

d) Support Vector Machine

Support Vector Machines (SMO) proves that this classification algorithm is much better than other and they proved it most people mostly prefer to use this algo. This supervised learning technique in medical side they are use very vastly and commonly they plotting disease predicting attributes regarding the multidimensional hyperplane and in two data clusters they optimally by creating approach. if we compare other this algo have high accuracy in use of nonlinear features which is called kernels. For Our data sets heart disease comparative analysis of precession and recall analyzing Orange 0.828 and Recall 0.828, accuracy 0.772

e) Random Forest

RF Classification tree is an ensemble of unpruned classification tree they give good performance in practical problem. If data sets are not noisy, they work excellent and it is not subject to overfitting. Its works fast, generally it’s more accurate than other tree base algorithms. RM (Random forests) predictions are based on number of trees and each tree are train within isolation. We have three main choices to performed random tree are constructed. For Our data sets heart disease comparative analysis of precession and recall analyzing Orange 0.832 and Recall 0.832, accuracy 0.772

f) Logistic Regression

Logistic regression relapse is a suitable relapse investigation to direct when the needy variable is dichotomous (parallel). The strategic relapse is a prescient examination. Logistic regression is utilized when the area variable (target) is clear cut. Logistic regression relapse is another method obtained by AI from the field of insights. It is the go-to technique for twofold characterization issues. For Our data sets heart disease comparative analysis of precession and recall analyzing Orange 0.849 and Recall 0.848, accuracy 0.885

g) KNN (K-nearest neighbor)

KNN is advanced method for classification which find the group of object k's in trading data sets and closed into test values and In the event that equivalent class is partaken in different of K-closest neighbors, at that point per-neighbor loads of as class are included or the subsequent weighted aggregate is utilized as the probability score of that class as for the test record. Its accuracy is depending on k's values and use nearest neighbor classifier [9]. Selecting K’s can be complete experimentally and where a number about shapes taken out from the training are set KNN. Our data sets heart disease comparative analysis of precession and recall analyzing Orange 0.676 and Recall 0.677, accuracy 0.706

Result Analysis:

In our dataset we implement 5 classification algorithms and we see Naïve Bayes have more accurate then other.

Classification Algorithms	Precession	Recall	Accuracy
KNN	0.676	0.677	0.706
SVM	0.828	0.828	0.882
RF	0.832	0.832	0.882
NB	0.828	0.828	0.907
LR	0.849	0.848	0.885

Table 1: Classification Algorithms with orange tool result

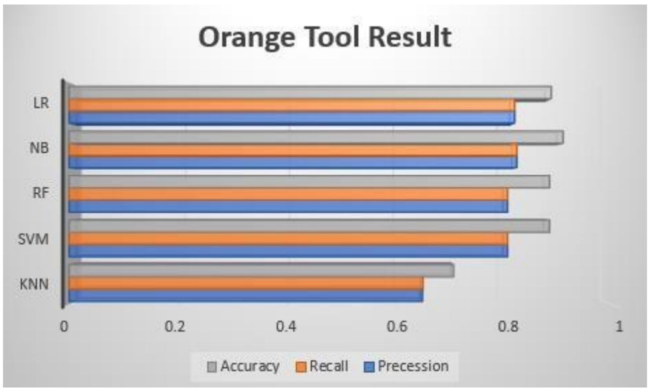

Figure 3: Visualization of Classification Algorithms with result

Our Working on orange tool implementing 5 classification algorithms in Figure 4.

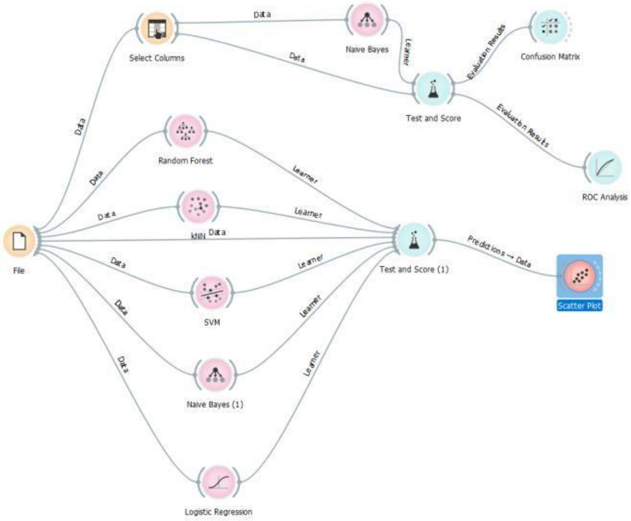

Figure 4: Orange tool working interface

In my case study I am comparing classification algorithms which are figure 5, 6,7,8,9. Pictorial reorientation in figure which comparing these classification

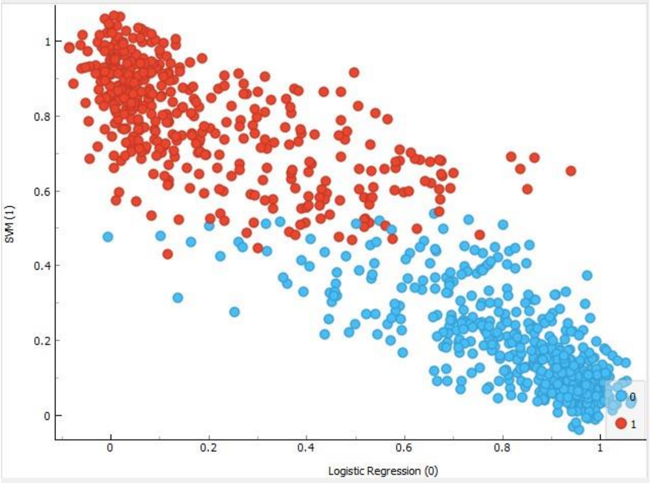

Figure 5: Comparing classification algorithms in LR Algo vs SVM Algo

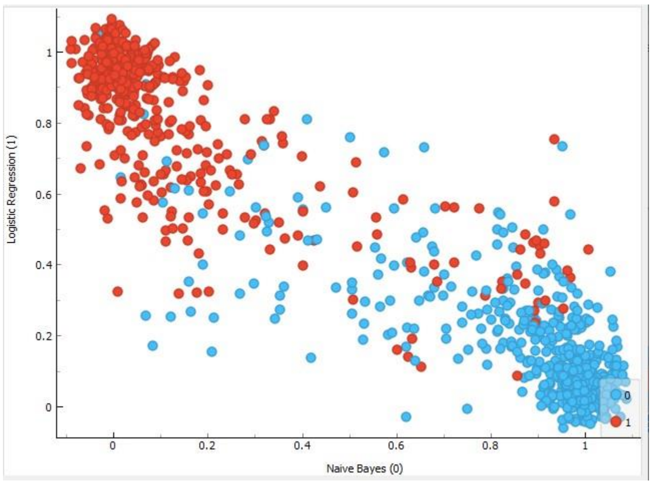

Figure 6: Comparing Naïve Bayes Algo vs LR Algo

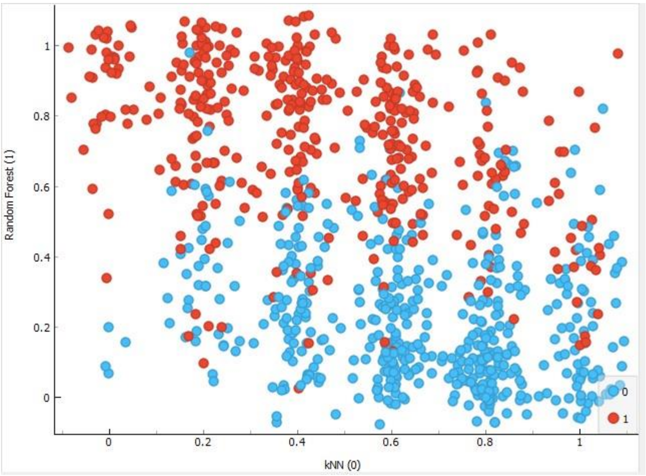

Figure 7: Comparing RF Algo vs KNN Algo

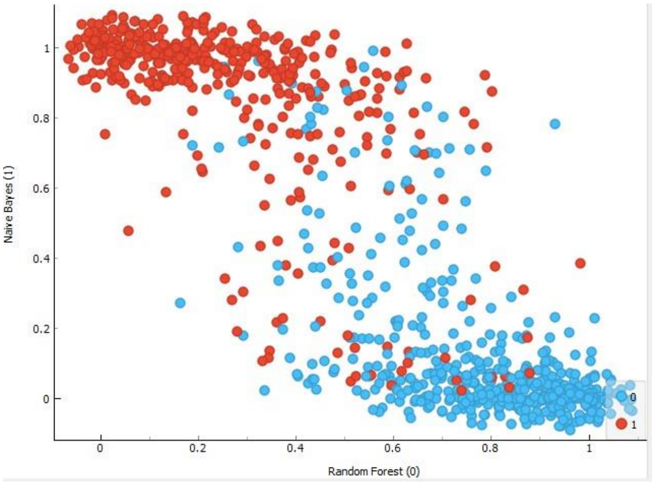

Figure 8: Comparing RF Algo vs Naïve bayes Algo

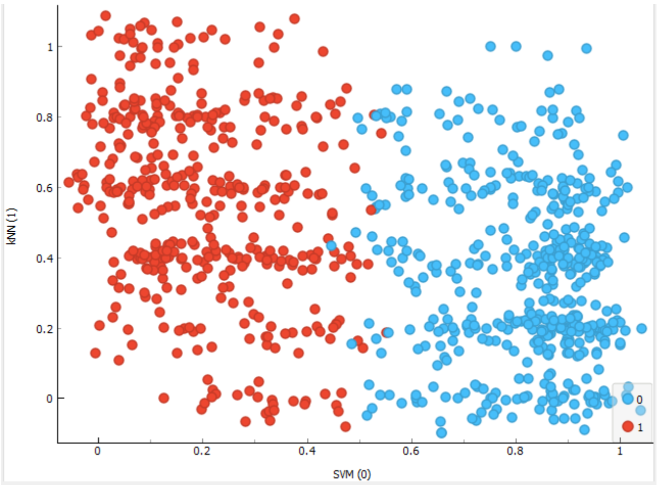

Figure 9: Comparing SVM Algo vs KNN Algo

XI. DISCUSSION

Orange tool is open source free tool they provide user friendly tool. Heart patients are need care before they diagnose his issue Doctors are need to first proper guess about his patients with previous reports of this type of patients. So, doctors adopted many methodology and scientific technologies to identify.

They also provide visualization of data [12]. Numerous examinations are conceivable by means of its visual programming interface that is move related with gadgets and numerous visual apparatuses will in general be upheld, for example, bar graphs, trees, scatterplots diagram, den programs and heat maps.

XII. CONCLUSION AND FUTURE SCOPE

Data mining Techniques help to find out best classification techniques to help the predict heart diseases its patterns and behavior. Classification algorithms are help to find out data mining methods to unclassified cases or label class. As my result in orange tool I am implementing 5 algorithms of classification and our result says that Naive Bayes is more accurate. the main objective is to compare 5 classification techniques and check which one is accurate. This Algorithms are used in computer version and robotics. For Our data sets heart disease comparative analysis of precision and recall analyzing Orange Tool Precision 0.828

and Recall 0.828, accuracy 0. 907.

Socio-statistic and life organize based division approach give client inclinations towards the items dependent on their statistic and life organize patterns. Naive Bayes, multiplayer recognition and Bayes net more tasteful are applied to different sections to watch item inclination expectations. Ten times cross approval is utilized to assess the presentation of the classifiers. The aftereffects of naive Bayes classifier show improved exactness in arranging occasions of significant portions. Naive Bayes classifier delivered better outcomes in foreseeing and controlling the clients of various life sections towards item inclinations contrasted with other two methodologies.

Naïve Bayes are proven that this algorithm are calculate more accuracy then other we work on in future to make decision on it and make ML to get perdition in early basses. We make machine to which have to pass just data sets and make all possible prediction on it and the predictions help to overcome heart diseases.

References

[1] G. KDD '18: Proceedings of the 24th ACM SIGKDD International Conference .November 16, 2018. 24th ACM Conference on Knowledge Discovery and Data Mining - KDD 2018.

[2] Data Science from Scratch Book by Joel Grus. understanding data science. Originally published: 2015Author: Joel Grus. Data Science from Scratch Book by Joel Grus. understanding data science. Originally published: 2015Author: Joel Grus.

[3] Data Mining online user-generated content: using sentiment analysis technique to study hotel service quality.Proceedings of the 46th Hawaii International Conference on System Sciences.

[4] Data Science from Scratch Book by Joel Grus. understanding data science. Originally published: 2015Author: Joel Grus.

[5] Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine, Single Volume: Expert Consult Premium Edition Originally published: February 25, 2011.

[6] Iyer A, Jeyalatha S, Sumblay R. Diagnosis of diabetes using classification mining techniques. IJDKP. 2015; 5(1):1–14.

[7] Gosain, A.; Kumar, A., "Analysis of health care data using different data mining techniques," Intelligent Agent & Multi-Agent Systems, 2009. IAMA 2009. International Conference on , vol., no., pp.1,6, 22- 24 July 2009.

[8] Reclaiming Liberalism (ISBN 1-86197-797-2) is a book written by a group of prominent British Liberal Democrat politicians and edited by David Laws and Paul Marshall in 2004.

[9] Competitions, Kaggle Kernels, Kaggle Datasets, Kaggle Learn, Jobs Board.Owner Alphabet Inc.On 8 March 2017, Google announced that they were acquiring Kaggle.

[10] Storytelling With Data: A Data Visualization Guide for Business Professionals .Book by Cole Nussbaumer Knaflic.

[11] Orange Data Mining, ‘Orange Data Mining Library Documentation Release 3’.

[12] Iyer A, Jeyalatha S, Sumblay R. Diagnosis of diabetes using classification mining techniques. IJDKP. 2015; 5(1):1–14.