ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

An Analytical Approach for Health Data Analysis and finding the Correlations of attributes using Decision Tree and W-Logistic Modal Process

Rohit Raturi, Abhishek Kumar

Associate Director, Regional Development Center, Enterprise Solutions Company, KPMG USA LP, Montvale, NJ, USA

Assistant Professor, Department of Computer Science & Engineering ACERC, AJMER, India

ABSTRACT: Medical domain is the wide range of information for the machine learning and deep learning methodologies and all the information we provided to the models can identify some of the insights of the features, which we are unable to identify with our knowledge. This research will speak on the diabetes dataset which was used for the identification of correlation between different attributes in the dataset. The machine learning will help to identify different things with respect to the models and all the information we are providing to the patients in the manner of reports and the treatment. This information is classified and this is the matter of security. But what if we can identify some new things from that information which is a classified, this comes to the picture for machine learning and this research work which we are mentioning in this article is about identifying the correlation between the samples of the information. This information is in the form of features and all the information we pass to the model can help to some extinct of identifying the apt feature and to identify the correlation of it with the next feature in the same model.

KEYWORDS: Machine Learning, Health care, deep learning, dataset, prediction.

I. INTRODUCTION

Health care is the major area for finding new information related to new diseases and the information and the clusters we find in health care data is more than any other kind of information. Health care is the most invested domain for the prediction model and the machine learning models using on health care information can identify some new things in each sector of the data available[1-5]. The information retrieved from the repositories of the health care will help use to insight the requirements of the implementations of the machine learning and deep learning methodologies. The implementation takes with the help of reputed algorithms and in general we call those as the models and the models like random forest, decision trees etc. the implementation of the decision trees was explained in this article with the sample outputs we acquired[6-8]. The implementation dealing with the correlations of the different attributes related to the different diseases. In this paper we discussed regarding the diabetes and the patient information related to the diabetic. This is some sort of chronic disease and the information related to the patients is stored in the form of the structured. This information is retrieved and subjected to the pre-processing. The pre-processing consists of removing the fake and unwanted information from the data set and the data after the process is used for the modeling.

Decision tree is the algorithm or we call it as a model used for identifying the output which is the final verdict based on the rules we impose on the data with the model. The model can consists of many kinds of information like, multi class representation using which we can build a multi class classification or regression mode[9-12]. The difference between classification and regression differs in identifying the data and which kind of data we are gathering. For example we

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

can consider the data like house rent prediction which consists of the information like square feet, number of rooms, distance from reputed locations etc. this is the static information which cannot be changed time to time. This kind of data can be subjected to classification. For example we are considering stock market as the prediction domain and the information we gathered can be the kinetic one. Because the information will change from time to time. This can be considered as the regression problem and we might not have any labels for the identification. The health care representation consists of some of the other research methodologies like neural networks, CNN, RNN, KNN etc. the most of the cases in the medical domain or any such kind of complex domains majorly work on CNB classifier. CNB classifier is the highest accuracy rate algorithm or model which can be helped for the identification of the features with the highest probability ratio.

The highest probability can be achieved the feature selection related to the domain. For example we have thousands of records in a dataset with some samples of 100 features. We cannot use all of the above features which can reduce the accuracy of the model. These are the some of issues will be discussing the lateral part of the section and all the information we provide in the literature survey has collected from the reputed medical device and medical third part organizations. Most of the literature review will be on identifying the kind of data and the next section will explain the proposed wok. Later section will explain the sample outputs we generated with our model based on the rule generation next we conclude the section with the apt information related to the title[13-14].

II. LITERATURE SURVEY

The literature survey mostly will discuss about the information related to the corporate which are being used for the development different medical related products which are used for the information retrieval. This information retrieved from different sources consists of the features which are as inputs. These inputs from the devices further need to be processed in the form of prediction models. Medical applications are used for the major implementations of the machine learning. To identify the importance of the knowledge base and the importance of the knowledge representation, we need to design the models with the sample rules. The rules are generated by the humans are subjected to the code. The code generated has to learn from the past experiences and the experiences we provide to the code are based on the human intelligence. The intelligence can be achieved by the group of people which can be related to the same domain of implementation. For example if the people we gathered are related to diabetics. They can give complete information about the disease the alternative treatments, medication and what are the things to be learned for the self treatment in the case of emergency. In this instance we need to identify the features which are correlated with each other. The features which are correlated are subjected to be the best part of implementation of the machine learning methodologies. The methodologies which are considered in this instance will lead you to the best part of the prediction model and its implementation. The implementation will run with related to the every part of the models and the every part of the feature we are considering.

i. MEDENT

MEDENT is the electronic health record application designed to make a better platform for the medical information and better utilization of the software which can be used for the medical applications. We have an user interface to the user requests for the medical services and all the requests are carried to the repository and based on this we have a conclusion for storing the medical records and providing the medical services.
Figure 1 explains the MEDENT chart central form which will carry the information related to the patients in the form of fields and all the information is carried to the server and the information provided will be considered to be the best part of understanding the correlation between the features.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

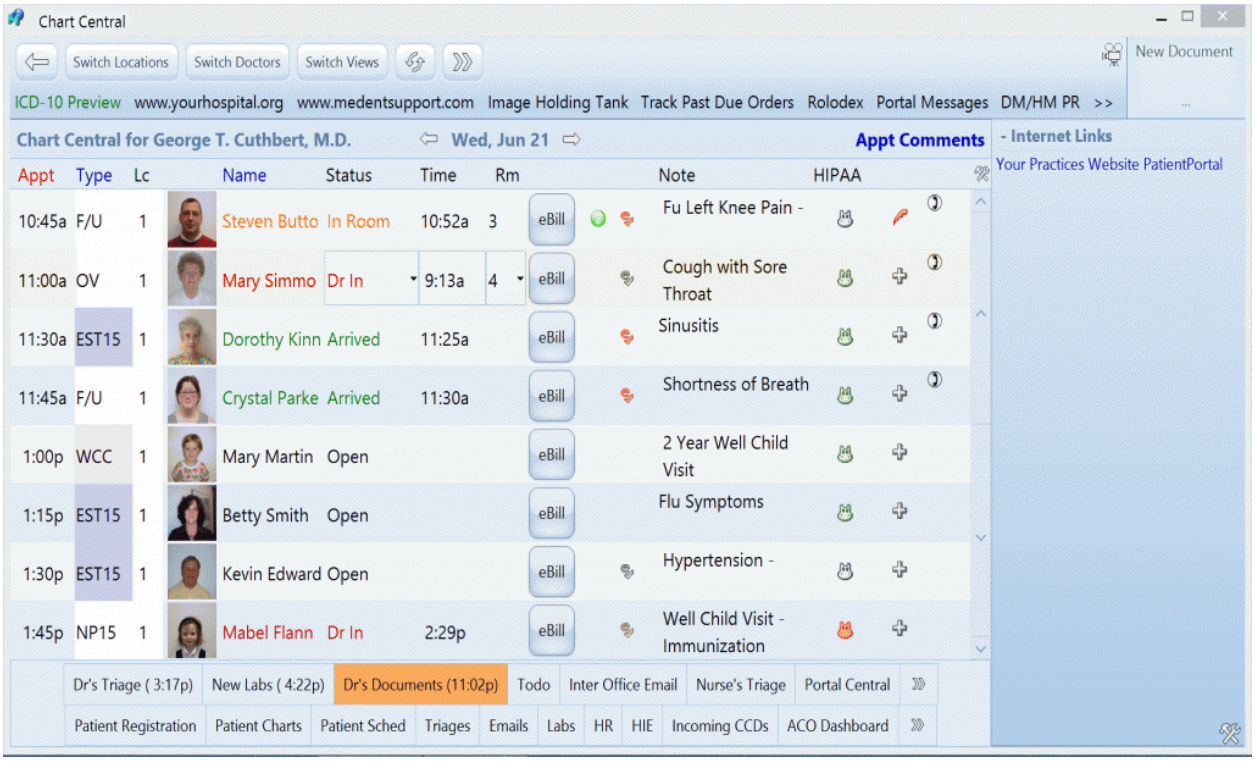

Fig: 1: MEDENT chart central form

ii. Salesforce Health Cloud

The implementation of the cloud computing platform is the major assert for the implementations of different health care related task. These tasks are used to find the way to help the poor for better implementation of the resources of the medical applications. These applications are used for the implementation of the best models. The models which are launched on the cloud can be used for any kind purpose related to health and the information and the service can be accessed at any time [15-17].

Machine learning application like Lybrato, MediPro is the other examples for the health care applications.

The most of the scenarios are being used in the machine learning applications is, they are being used for the data storage and access. If they are used in the full fledged manner they can be used for any kind of applications.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

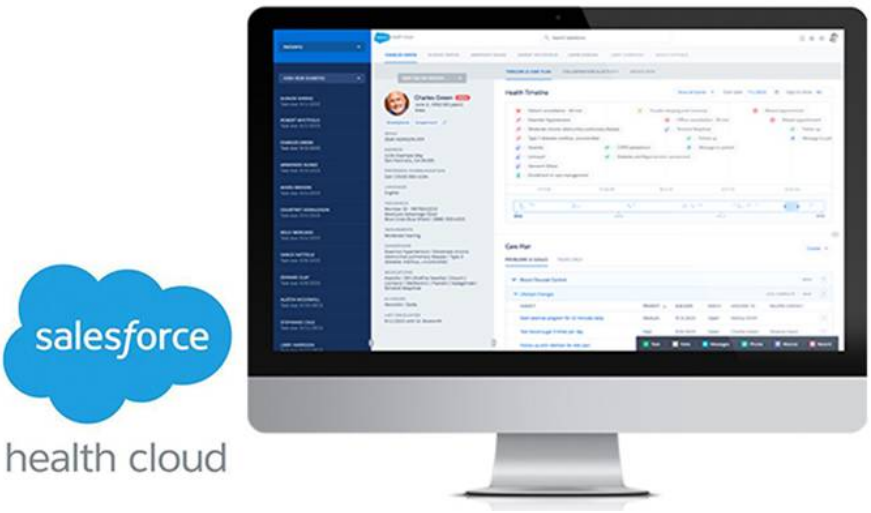

Fig: 2: Saleforce Health Cloud Dashboard

III. PROPOSED WORK

In our Study, we'll use the Diabetes patient Data in which we'll analyze the each attributes statistics, and find the correlations between attributes to another attribute. This will help us to find the best predicting attribute for the final prediction as outcome. In previous studies there was lot of work done in the health field to find the best technique or predicting the Disease Status. But in our study, we'll analyze the data more interactively, will find the co-relations between dataset attributes, will proceed to find the final result as prediction of disease and at the end of study we'll also complete the comparative study of evaluations[18].

Diabetics is the most common thing in this present generation and this application and the research are used for analyzing the importance of the features and the importance of the correlation between the features. These are the considerations which have to be followed and the information retrieved from the different sources can also affect the modeling scenario.

i. **Implementation:** We have a dataset for Diabetic Patients with the Disease symptoms as attributes. We have some screen captures of dataset indicating some important statistical information of data values.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

Label			Min	Max	Average
Outcome	Integer	0	0	1	0.349
Pregnancies	Integer	0	0	17	3.845
Glucose	Integer	0	0	199	120.895
BloodPressure	Integer	0	0	122	69.105
SkinThickness	Integer	0	0	99	20.536
Insulin	Integer	0	0	846	79.799
BMI	Real	0	0	67.100	31.993
DiabetesPedigreeFunction	Real	0	0.078	2.420	0.472
Age	Integer	0	21	81	33.241

Fig: 3 Dataset Attributes

In above figure we have a list of attributes contained in the dataset. Each attribute has some statistical information in the figure i.e. the data type of the attribute such as integer/real/Polynomial etc. , missing values of the specific attribute, Minimum value of the attribute in whole dataset, Maximum value of the attribute in whole dataset, the Average of the attribute value.

Graphical View of all attributes: The given figure represents the graph position of each attribute according to the variance of the values and indicated by the color variation [19].

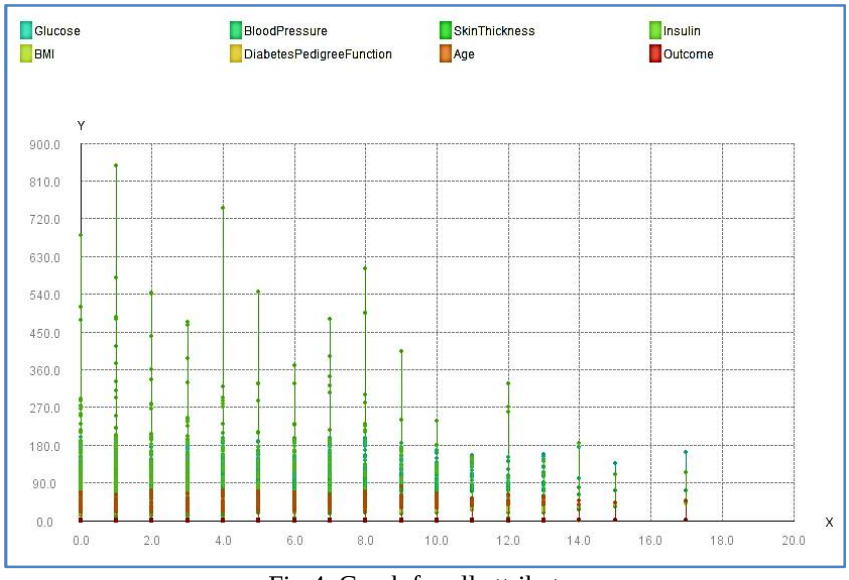

Fig 4: Graph for all attributes

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

Finding Co-relation among the attributes: In this section we'll find the co-relation among the various attributes of the dataset. Because each attribute in our dataset plays a role of a particular symptom of the disease, the co-relation shows the dependency of the one symptom to another symptom. The Co-relation also denotes the major cause of the disease as having maximum weight. Finding co-relation minimizes the extra effort for finding the outcome and the processing cost. If we could identify the major causes of the disease, then the size of dataset can be reduced and the processing efforts and cost can be minimized.

To find the co-relation amog the dataset attributes, we'll use an operator (technique) named as weight by corelation. This operator process each attibute of the given dataset and find the weight of each attribute comparetively[20].

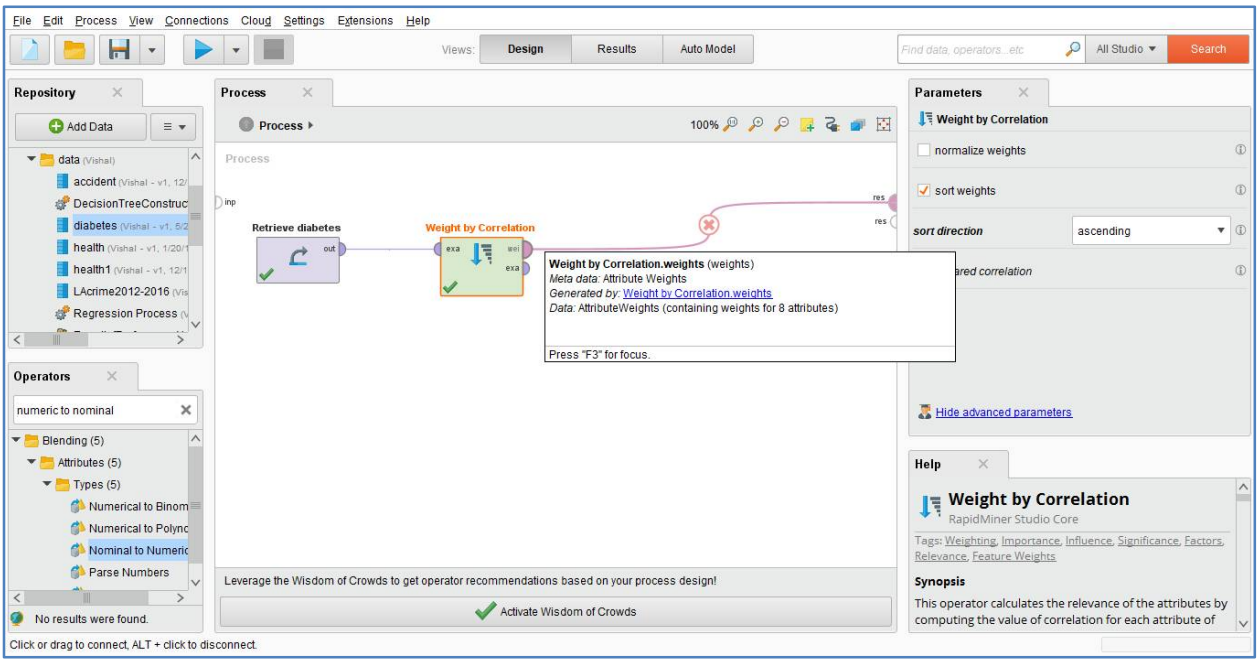

Fig 5: Corelation Process

attribute	weight
BloodPressure	0.004
SkinThickness	0.006
Insulin	0.017
DiabetesPedigreeFunction	0.030
Pregnancies	0.049
Age	0.057
BMI	0.086
Glucose	0.218

Fig 6: Attributes with their respective weights

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

In above figure there is a table containing list of attributes of dataset and their respective weights according to the attributes values. The maximum weight denotes the maximum dependency of that particular attribute on the prediction variable or attribute. And the least attribute weight represents the minimal dependency on the decision attribute. The the conclusion of the story is that the least weight attributes can be ingnored during the prediction process.

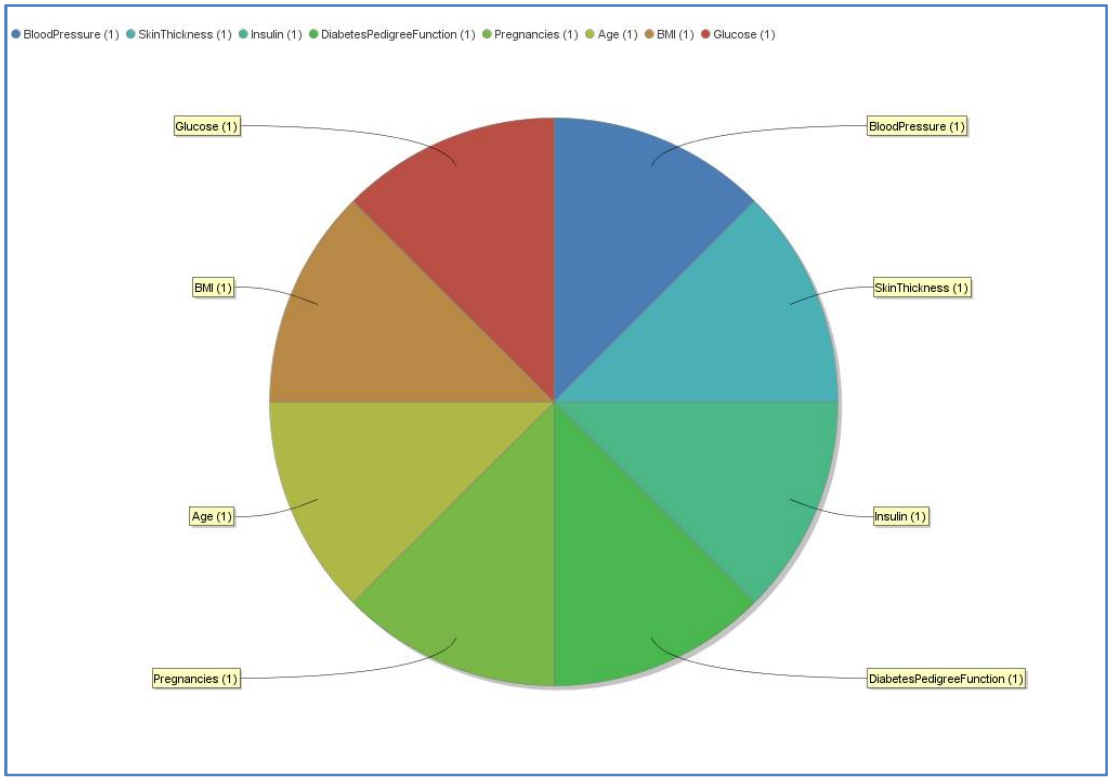

Fig:7 Graphical presentation of the test result

In above figure, we have a pictorial representation of the Co relation test result through pie-chart. Each section in pie chart represents the weight of attribute.

Predicting the Disease using Decision Tree Model Process: Decision Tree is one of the practicle categoriztion technique to understand the basic scenario of the dataset. We'll use this technique to understand the basic catogrization of the dataset attributes and shall use this process model to predict the major causes of the disease. To use this technique we'll use the Decision Tree operator and provide the dataset with major attributes found in the corelation test i.e. use those attributes that had maximum weights[21].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

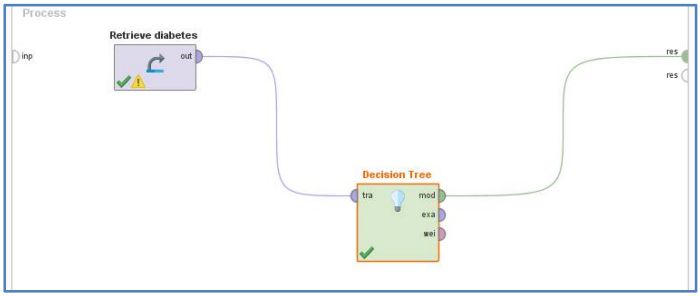

Fig 8: Decision tree Model Process

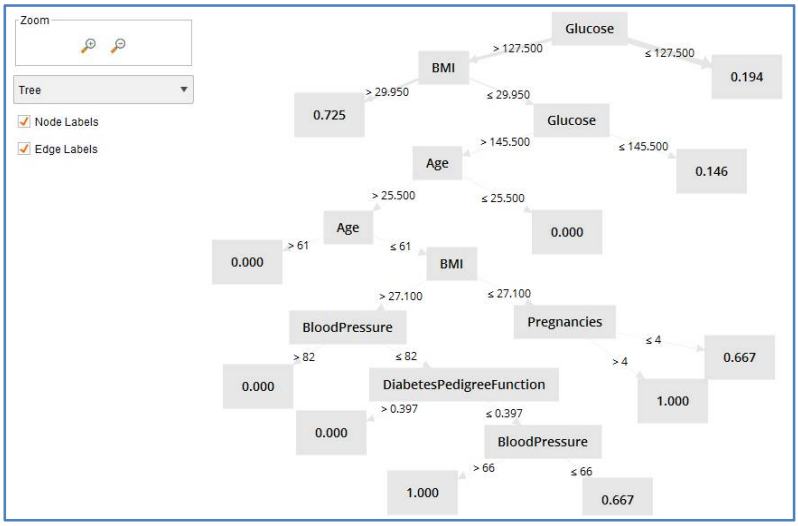

Fig 9: Result of Decision Tree (Graphical View)

In above figure, we have a graphical view of decision tree construction. We can see each node with its predicting statistical measure. And according to its discreet values the splitting was done. The categorization of nodes represents the dependency of the attributes on its parent attribute. Like Glucose have two children i.e. BMI and AGE. The Top most Node represents the Major cause of the disease and so on.

Basic Rule generation using Decision Tree: In this section of study, we are generating some important rules for disease causes with their important statistical measures i.e. confidence and support. These rules are generated from the above decision tree process. It will indicate the whole process result or decision tree construction rules. By going through these rules we can easily identify the major causes of the disease in our case. We can see that the Glucose, BMI, Age have maximum support in comparison of other attributes.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

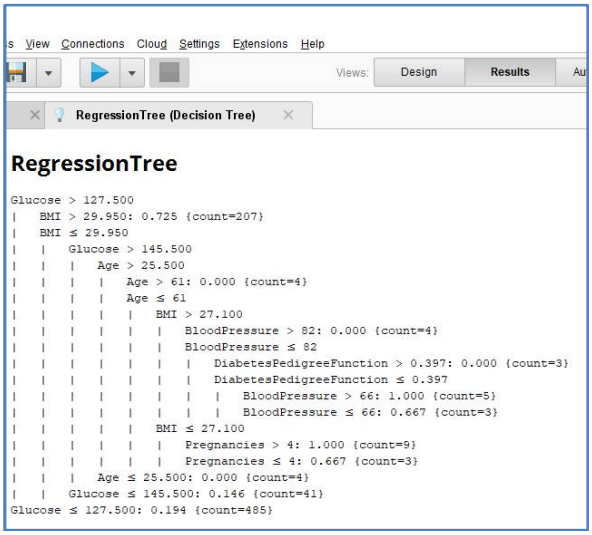

Fig 10: Rules generated from decision tree

W-Logistics approach for finding results: The W-Logistic is regression process in data analytics. The W denotes the extension i.e. this is the Weka Extension. Weka is a open source tool for data mining and data analytics. It is java based open source tool containing some of the major data mining techniques in the form of algorithms. regression examination could be a heap of factual procedures for assessing the connections among variables. It incorporates various strategies for displaying and breaking down a couple of variables, once the eye is on the association between a dependent variable and a minimum of one free variables. All the additional expressly, regression examination encourages one see however the run of the mill estimation of the dependent variable changes once any of the autonomous variables is fluctuated, whereas the opposite free variables area unit control fastened[22].
Now we'll apply the W-Logistic Operator on the given dataset and find the final result.

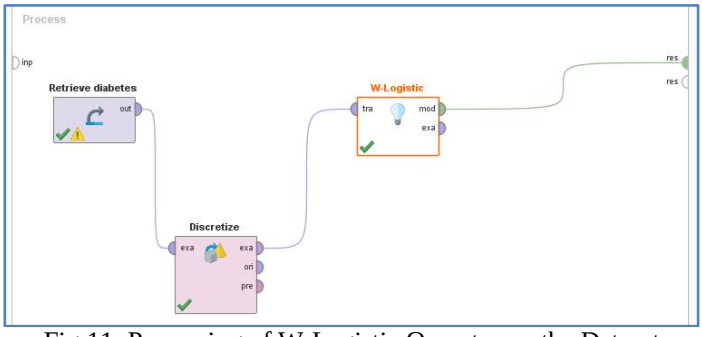

Fig 11: Processing of W-Logistic Operator on the Dataset

Now we have the result of the above process. The Result contains the coefficients and odds ratios of the Variables with their respective classes.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

W-Logistic

```
Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
                               Class
Variable                       first
=================================
Pregnancies                        0
Glucose                            0
BloodPressure                      0
SkinThickness                      0
Insulin                            0
BMI                                0
DiabetesPedigreeFunction           0
Age                                0
Intercept                    -24.893


Odds Ratios...
                               Class
Variable                       first
=================================
Pregnancies                        1
Glucose                            1
BloodPressure                      1
SkinThickness                      1
Insulin                            1
BMI                                1
DiabetesPedigreeFunction           1
Age                                1
```

Fig 12: Final result of the W-Logistic Operation

Performance Evaluation of the used Techniques: In this section of study, we'll evaluate the comparative performance of the both techniques i.e. Decision tree and W-Logistic.

i. **Decision Tree Performance Evaluation:** We have some performance evaluation measures in data analytics to evaluate the performance of the technique. These are Accuracy confusion matrix, AUC curves, True Positive Rate and precision Matrix. We'll test each of them step by step. This is will done by using the performance operator in the tool. The screen capture of complete process is given in the figure placed below:

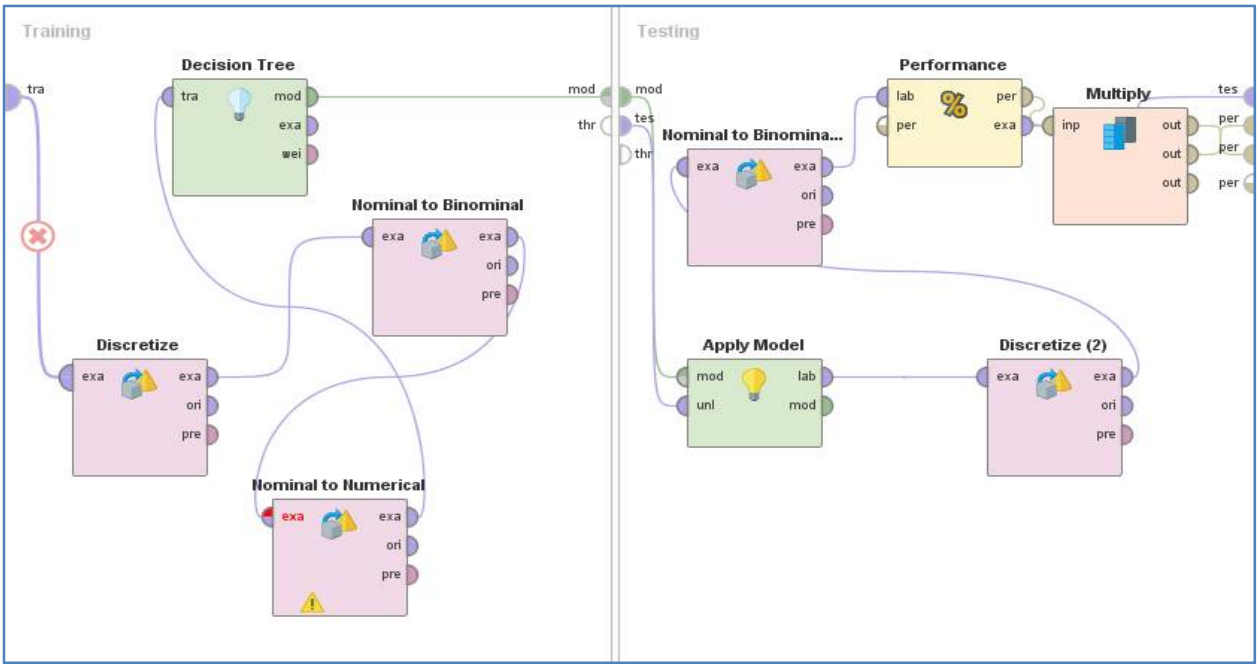

Fig 13: The complete process of performance evaluation

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

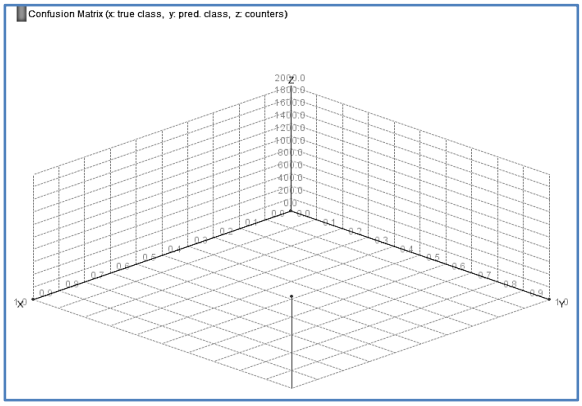

Fig 14: The confusion matrix for Accuracy

In above figure there is a confusion Matrix for evaluating the accuracy measure of the decision tree process. It is a 3-D visualization having the regular intervals of the accuracy values.

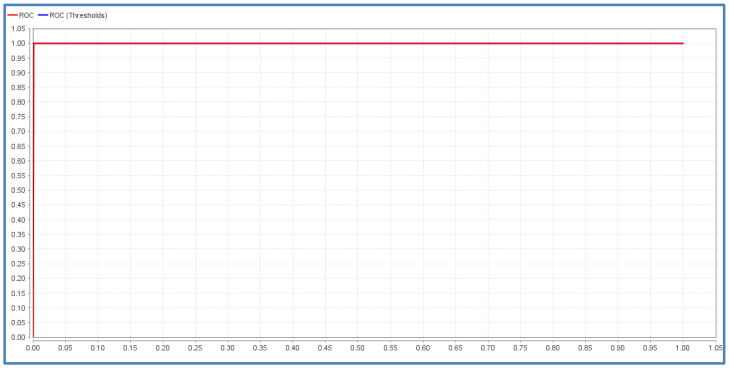

Fig 15: AUC curve measure of Decision Tree

In above figure the graph show the AUC curve of the decision tree process for performance on the given dataset. According to the graph, the result is satisfactory. Because it reaches its maximum level of accuracy.

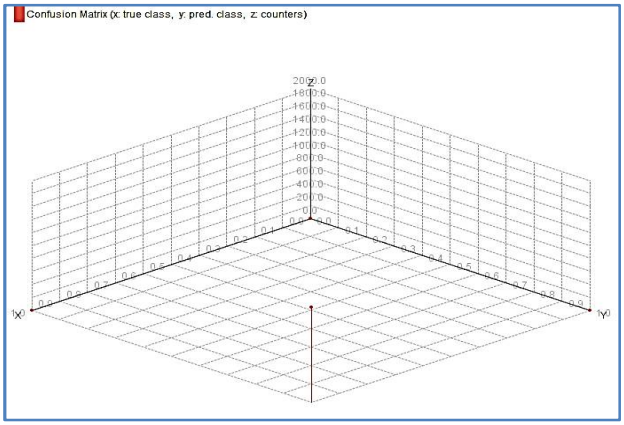

Fig 16: Confusion Matrix for Precision

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

In above figure there is a confusion Matrix for evaluating the accuracy measure precision of the decision tree process. It is a 3-D visualization having the regular intervals of the accuracy values.

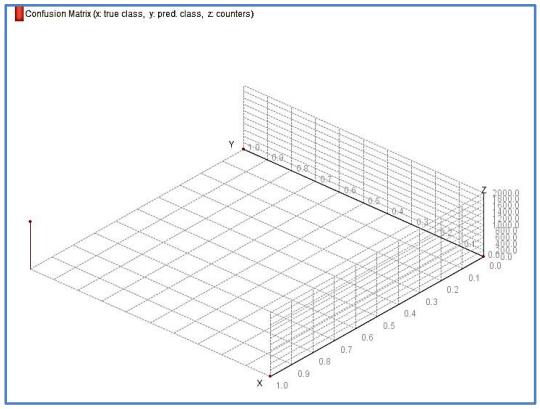

Fig 17: Confusion Matrix for TP rate

In above figure there is a confusion Matrix for evaluating the accuracy measure of the decision tree process i.e. True positive rate. It is a 3-D visualization having the regular intervals of the TP rate accuracy values.

ii. **W-Logistic Performance Evaluation:** We have some performance evaluation measures in W- Logistic data analytics to evaluate the performance of the technique. These are Accuracy confusion matrix, AUC curves for Threshold, and precision Matrix. We'll test each of them step by step. This is will done by using the performance operator in the tool. The screen capture of complete process is given in the figure placed below:

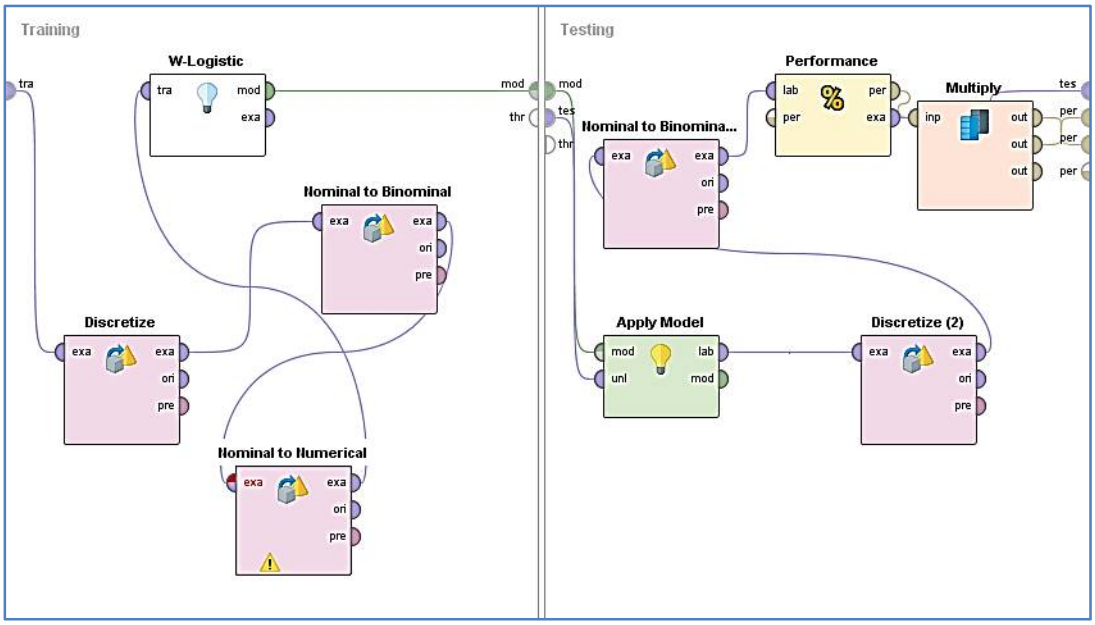

Fig 18: The complete process of performance evaluation of W-Logistic

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

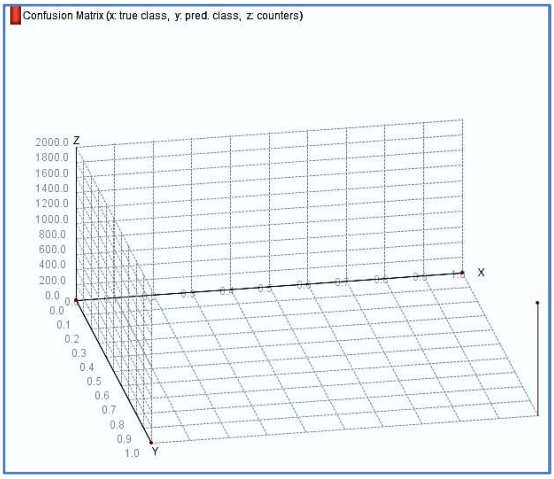

Fig 19: The confusion matrix for Accuracy

In above figure there is a confusion Matrix for evaluating the accuracy measure of the decision tree process. It is a 3-D visualization having the regular intervals of the accuracy values.

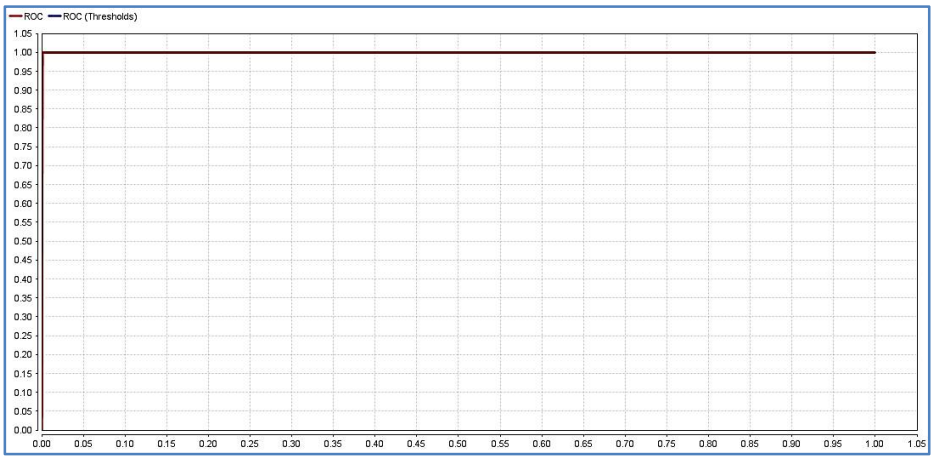

Fig 20: AUC curve measure of W-Logistic

In above figure the graph show the Threshold AUC curve of the W-Logistic process for performance on the given dataset. According to the graph, the result is satisfactory. Because it reaches its maximum level of accuracy.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

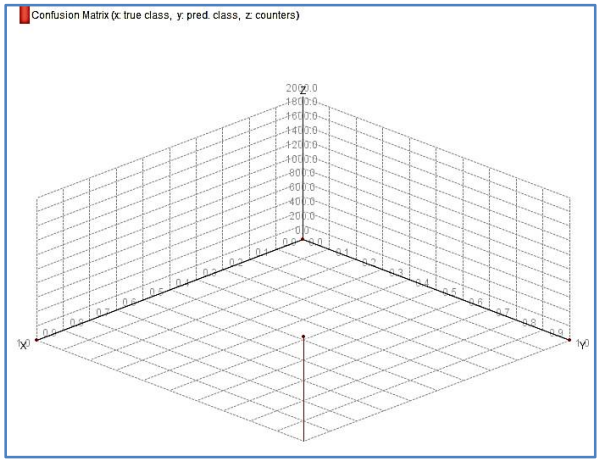

Fig 21: Confusion Matrix for precision

In above figure there is a confusion Matrix for evaluating the accuracy measure of W-Logistic process i.e. True positive rate. It is a 3-D visualization having the regular intervals of the precision values.

IV. CONCLUSION

Machine learning is the most prominent platform for the implementation of the different future scope domains like health care, vision technologies etc. Decision tress are the most prominent model which is being used for different real time problems like prediction models, and some of the most important thing is identifying the correlation between the features of the models. These are the things which are required for getting highest accuracy rate for the model. This is the most important thing for the machine learning engineer to identify the best features and the correlation among the features.

REFERENCES

[1] "An improved approach for prediction of Parkinson's Disease using Machine Learning Techniques", Kamal Narayan Reddy Challa* ,2016, IEEE
[2] Adoption of Electronic Health Record System: Multiple Theoretical Perspective", Qiwei Gan, Qing Cao – 2014 IEEE
[3] "A Scalable mHealth System for Noncommunicable Disease Management", G D Clifford* - 2014 IEEE
[4] "Predictive Medication and use of BigData", Avijit Goswami – 2017 IEEE
[5] "Variation in Outcome in Tethered Cord Syndrome", Norulain Iqbal*, 2016, Asian Spine Journal
[6] "Resource Frequency Prediction in Healthcare: Machine Learning Approach" Daniel Vieira, 2016 IEEE
[7] National Patient Safety Association, "Safer care for acutely ill patients:Learning from serious accidents," Tech. Rep., 2007.
[8] National Institute for Clinical Excellence, "Recognition of and response to acute illness in adults in hospital," Tech. Rep., 2007.
[9] H. Gao, A. McDonnell, D. Harrison, S. Adam, K. Daly, L. Esmonde, D. Goldhill, G. Parry, A. Rashidian, C. Subbe, and S. Harvey, "Systematic review and evaluation of physiological track and trigger warning systems for identifying at-risk patients on the ward," *Intensive Care Med.*, vol. 33, no. 4, pp. 667–679, 2007.
[10] L. Tarassenko, D. Clifton, M. Pinsky, M. Hravnak, J. Woods, and P.Watkinson, "Centile-based early warning scores derived from statistical distributions of vital signs," *Resuscitation*, vol. 82, no. 8, pp. 1013–1018, 2011.
[11] D. Prytherch, G. Smith, P. Schmidt, P. Featherstone,K. Stewart,D.Knight,and B. Higgins, "Calculating early warning scores—A classroom comparisonof pen and paper and hand-held computer methods," *Resuscitation*,vol. 70, pp. 173–178, 2006.
[12] A. Hann, "Multi-parameter monitoring for early warning of patient deterioration,"Ph.D. dissertation, Univ. Oxford, Oxford, U.K., 2008.
[13] D. Wong, I. Strachan, and L. Tarassenko,"Visualisation of highdimensional data for very large data sets," presented at the Workshop Mach. Learn. Healthcare Appl., Helsinki, Finland, 2008.
[14] B. Sch¨olkopf, J. Platt, J. Shawe-Taylor,A. J. Smola, and R. C.Williamson,"Estimating the support of a high-dimensional distribution," *Neural Comput.*,vol. 13, no. 7, pp. 1443–1471, 2001.
[15] S. Hugueny, D. Clifton, and L. Tarassenko, "Probabilistic patient monitoring with multivariate, multimodal extreme value theory," *Commun. Comput. Sci.*, vol. 127, pp. 199–211, 2011.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

[16] R. Kavitha, E. Kannan, S. Kotteswaran,"Implementation of Cloud based Electronic Health Record (EHR) for Indian Healthcare Needs",Indian Journal of Science and Technology,2016 Jan, 9(3), Doi no:10.17485/ijst/2016/v9i3/86391
[17] Meenakshi Sharma, Himanshu Aggarwal,"EHR Adoption in India: Potential and the Challenges",Indian Journal of Science and Technology,2016 Sep, 9(34), Doi no:10.17485/ijst/2016/v9i34/100211
[18] ResScan Software : ResScan version 4.2 Clinical Guide from "ResMed Ltd 1 Elizabeth Macarthur Drive Bella Vista NSW 2153 Australia"
[19] R. Lozano, and C. J. L. Murray, "Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Disease Study 2010," *The Lancet*, vol. 380, no. 9859, pp. 2095–2128, 2012
[20] A. T. Campbell, S. B. Eisenman, N. D. Lane, E. Miluzzo, and R. A. Peterson, ``People-centric urban sensing," in *Proc. 2nd Annu. Int. Workshop Wireless Internet (WICON)*, New York, NY, USA, 2006, Art. no. 18. [Online]. Available: http://doi.acm.org/10.1145/ 1234161.1234179.
[21] R. K. Ganti, F. Ye, and H. Lei, ``Mobile crowdsensing: Current state and future challenges," *IEEE Commun. Mag.*, vol. 49, no. 11, pp. 32_39, Nov. 2011.
[22] M. Pouryazdan, B. Kantarci, T. Soyata, and H. Song, ``Anchor-assisted and vote-based trustworthiness assurance in smart city crowdsensing," *IEEE Access*, vol. PP, no. 99, pp. 1_1, doi: 10.1109/ACCESS.2016.2519820.