International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Predicting & Visulaizing the Clusters Assignments in Health Care Dataset for Disease Prediction

Neeraj Bhargava, Ritu Bhargava, Abhishek Kumar, Shikha Bhardwaj

Abstract— DM is the process which is used for the analyzing hidden patterns of data. This analyzing completed according to the several perspectives for categorization into usable information. Here, DM is referred as the Data Mining It is composed and assembled in same regions, like data warehouses, for effective analysis, DM algorithms. In paper we will use these records and will find the major attribute which plays an important role in disease prediction. To do so, first we implemented Naive bayes' algorithm where every pair of features being classified is independent of each other. Once we get the Naive Bayes' Result then we apply the Clustering technique on the same dataset. Simple K-Means Clustering is used to get the clusters of the data results. We can visualize the Cluster assignments for each attribute against the Resultant or prediction attribute. We can have the better understanding through these visualizations about the dependencies of attributes on the prediction variable. K-means algorithm is an iterative algorithm that tries to partition the dataset into K predefined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. And after final analysis of the result of both techniques we found two attributes which are having maximum weight as compare to others. These two attributes Glucose and Insulin must consider in the diabetes prediction.

Keywords—Weka ,data Mining,DM,Decision tree ,SVM.

I. INTRODUCTION

Over every one of the region, information is being gathered and collected at a clear pace. Computational hypotheses and tools to help people extract valuable data (information) from rapidly evolving versions of computerized information is an important requirement for another age. At the centre of the procedure is the use of specific information digging strategies for design extraction and disclosure [4]. Among the information mining systems created as of late, the information mining techniques are including speculation, portrayal, order, bunching, affiliation, development, design coordinating, information representation and meta-control guided mining. [2]. Healthcare is very important domain in which data mining is playing a very important role. Various efficient works has been performed in this domain with effective results to filter on further, many works has been analyzed along with the algorithms used and analysis is performed in this paper.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
Prof Neeraj Bhargava*, School of engineering and system Sciences, MDS University, Ajmer, India. Email: profneerajbhargava@gmail.com
Dr. Ritu Bhargava, Lecturer, Computer Science, Sofia Girls' college, Ajmer, India. Email: drritubhargava92@gmail.com
Abhishek Kumar, Asst. Prof, Computer Scienc Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India. Email: Abhishek.kumar@chitkara.edu.in
Shikha Bhardwaj, Research Scholar, Mjrp, Jaipur, India, Shikhabhardwaj09390@gmail.com

Result analysis is the most important part which our work has focused on .data mining basic algorithms has been applied and comparative results has been shown in the result section

Data Mining An Overview

Information measure are for the most part developing from everyday. The need to see expansive, complex, data enhanced informational collections has now expanded in all the changed fields of innovation, business and science. With these expansive measure of information, the capacity to extricate helpful learning covered up in these extensive measure of information and to follow up on the information is winding up progressively critical in the present focused world. The way toward applying PC based data framework (CBIS), including new systems, for finding learning from information is called information mining [3].

II. RELATED WORK

Daniel A. Keim focused on large volume of data that quit difficult to explore. To achieve high potential result the author used classification information visualization techniques and visual data mining. The extension of this work indicates some directions include integration of visualization techniques with machine learning, operation research, simulation and statistics. [5, 6]

Maria Cristina Ferreira de Oliveria et al. worked on information visualization techniques to explore the outcomes of mining algorithms. The authors also review the pioneering approaches by integrating into DM / KDD process. The conclusion leads on work which used to enhance user interaction. The extension of this work that denotes more than the traditional application of visualization techniques to support non-analytic staged of a KDD process but analytic DM algorithms in which visualization plays a major role.[7,8]

Tang et al. worked on classification methods to classify the medical data on the basis of Correction Rate, leaf number and tree depth. The work targeted to improve the correction rate of coronary heart disease data by comparing decision tree algorithm and system reconstruction analysis. The evaluation process carried on the performance of ID3, C4.5, CART, CHAID and exhausted CHAID. The analysis leads to the conclusion that system reconstruction method gets higher level of correction rate but contained little effect on leaf number and tree depth of decision tree.[9, 10]

Weiguo Han proposed Intelligent Transportation System (ITS) monitoring device to aid traffic congestion to increase traffic situation. The author applied several visualization techniques to explore the traffic volume data that implemented on data analysis tools.

The methods were applied to gain more effective result or patterns on massive datasets. The outcomes of data analysis help manager, engineer & planner to take efficient & effective decision making of traffic operations. For the future extension, the author wanted to work on transportation scientifically

Palaniappan S. et al] developed a prototype IHDPS on .NET platform. Here IHDPS is referred as the Intelligent Heart Disease Prediction System. The analysis conducted on different DM techniques – NB, Neural Network and Decision Tree . The outcomes of the efforts illustrates that each techniques achieved unique strength[11].

Hailiang Jin et al. conducted visual data exploration with directly involved containing data mining process. The analysis leads to the conclusion that the combination of dedicated knowledge by applying the data mining algorithms. The visual data mining techniques were implemented using national advance data mining tools based on the task and objectives.[12]

A.Rajkumar et al.worked on the use of supervised machine learning algorithms to classify data of datasets which supports diagnosis of heart disease with larger level of accuracy. The datasets contained data of heart patients and the algorithms were implemented using tanagra s/w tool to process and then compare the performance of Data Mining algorithms. The analysis of the results targeted to performance study of algorithms based on the measuring parameters Recall and precision. The analysis lead to the conclusion specify Naïve Bayesian algorithm is far superior in compact time compare to the other consider algorithm and the accuracy achieved of the order of 53% satisfy the attributes of importance in the process of classification.[13]

K.Srinivas et al. explored the applicability of the Neural Network for Cardiovascular disease (CAD). Since the Medical diagnosis is an crucial act to be performed efficiently and accurately , the authors have taken automated system for medical diagnosis to increase medical care at much reduced costs. There are some DM methods available which is the Neural Network, Naïve Bayes, and Decision Trees are used for prediction about the heart disease employing Neural Network technique. The results for Neural Network technique justified higher level accuracy and sensitivity considering performance evaluation parameters.[14]

- Xiu-yu zhong, tried to enhance the website optimize website structure, design and build intelligence website. To accomplish the targeted purposes, the author utilized ML methods and for the apriori algorithm, customer identification method, and session identification algorithm implemented on the pre-processed dataset. Here ML is referred as the Machine Learning. It is worked on WEKA open source DM s/w instrument. The resulted outcomes reached on the origins of apriori method. The author offers the issues that must implement on another DM method and the comparative analysis of result gives more optimal result.[15]

P. Santhi et al. explored the comparative performance of the clustering and classification methods through heart disease dataset. The work targeted to higher level of prediction accuracy through comparing both the methods. The evaluation carried on the performance of classifiers of, functions (SMO), Bayes (Naïve Bayes, Naïve Bayes updateable), Meta Multi BoostAB, Multiclass Classifier) Lazy (IB1, IBK) , trees(NB Tree) , Rule(Decision Table), and the clustering methods of EM, Make Density Based Clusters , Cobweb, Simple K-Means , Farthest First methods . The analysis lead to the conclusion which states that the NB Tree having higher prediction Accuracy compared to the clustering method.[60]

Mai Shouman et al. focused on the work statistical and the DM instrument to diagnosis disease. To accomplish the more perfect results of DM methods they implemented hybridization on the selected techniques. First to diagnosis the heart disease patient the signal Data Mining techniques were used which illustrate the acceptable levels of accuracy than for enhancing the accuracy of disease the hybridization DM method . Hybrid Data mining techniques produces more effective result in diagnosis of heart disease. Different hybrid techniques like fuzzy artificial immune recognition system and k-nearest neighbor are applied together which produced accuracy of 87%. NN gives accuracy of 89.01% which is better. One case of genetic algorithm and neural network is also discussed which created well result in determining heart disease.[17]

Anuj Sharma et al. explores the applicability of feature selection methods with different classification techniques namely- Naïve Bayes, SVM, Decision Tree, KNN, Maximum Entropy , Adaboost , Winnow. The proposed work conducted to improve the performance of classification techniques. The experimental result lead to the conclusion that Gain Ration gives the best performance for feature selection method and SVM performance achieved higher level performance of classification techniques.

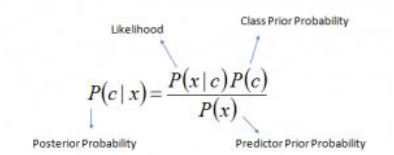
Chaitrali S. Dangare et al. performed comparative study of classification techniques. The dataset contain the data of heart disease and the algorithm were implemented in the WEKA[15] tool to increase the level of accuracy by including two more attributes in the dataset. The analysis leads to the conclusion which state that the Neural Network having higher level of accuracy (with 100%) as compare to Naïve Bayes (with 99.62%) and Decision Tree (with 90.74%). For further expansion to achieve the more accurate result apply more number of attributes with other Data mining techniques for prediction.[18]

Mai Shouman et al. worked on several DM techniques to diagnosis a heart disease. The analysis of the results targeted to performance of DM techniques based on the measuring parameters – sensitivity, specificity and accuracy. The exploration leads to the conclusion that specify KNN achieved higher level of accuracy of 97.4 % than Neural Network.[18]

III. METHODOLOGY

In this research paper, we applied the Naïve Bayes algorithm on the given dataset. The naïve bayes algorithm is basically a classification technique.

This technique stats that a special feature of any section is not related with other features of that group. Because this algorithm is well suited for the large dataset with simplicity thus we used it on our dataset.



$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c) \quad (1)$$

Where:
P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
P(c) is the prior probability of class.
P(x|c) is the likelihood which is the probability of predictor given class.
P(x) is the prior probability of predictor.

3.1 Algotithm

Step 1: Input dataset and convert it into frequency table.
Step 2: Find probability and create a likelihood table.
Step 3: Now calculate the the Posterior probability for each class from eq. (1). The class with highest Posterior probability would be the final prediction.

IV. SIMULATION AND RESULTS

4.1 Dataset

This dataset is taken from web containing 9 attributes named as Pregnancies having nominal values (P for Positive and N for Negative), Glucose having numeric values, Skin Thickness having numerical values, Insulin having numerical values, BMI having numerical values, Diabetes Pedigree Function having numerical values, Age having numerical values, and our Prediction Attribute having nominal values (TRUE and FALSE). The dataset contains 768 instances or records.[16]

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
P	85	72	35	0	33.6	0.627	50	TRUE
P	100	66	29	0	26.6	0.351	31	FALSE
N	85	64	0	0	23.3	0.672	32	TRUE
P	89	66	23	94	28.1	0.167	21	FALSE
N	137	40	35	168	43.1	2.288	33	TRUE
P	116	74	0	0	25.6	0.201	30	FALSE
N	78	50	32	88	31	0.248	26	TRUE
P	115	0	0	0	35.3	0.134	29	FALSE
N	197	70	45	543	30.5	0.158	53	TRUE
N	125	96	0	0	0	0.232	54	TRUE
N	110	92	0	0	37.6	0.191	30	FALSE
P	168	74	0	0	38	0.537	34	TRUE
P	139	80	0	0	27.1	1.441	57	FALSE
P	189	60	23	846	30.1	0.398	59	TRUE
P	166	72	19	175	25.8	0.587	51	TRUE
P	100	0	0	0	30	0.484	32	TRUE
N	118	84	47	230	45.8	0.551	31	TRUE
P	107	74	0	0	29.6	0.254	31	TRUE
P	103	30	38	83	43.3	0.183	33	FALSE
P	115	70	30	96	34.6	0.529	32	TRUE
N	126	88	41	235	39.3	0.704	27	FALSE
N	99	84	0	0	35.4	0.388	50	FALSE
P	196	90	0	0	39.8	0.451	41	TRUE
N	119	80	35	0	29	0.263	29	TRUE

Fig 1: Dataset for Diabetes

4.2 Attribute Visualization

1. Pregnancies Attribute

This attribute contains the status of Pregnancies in the particular case of a patient. If the patient is Pregnant then the attribute value is P else N.

The below figure shows the count of each Value (P or N).

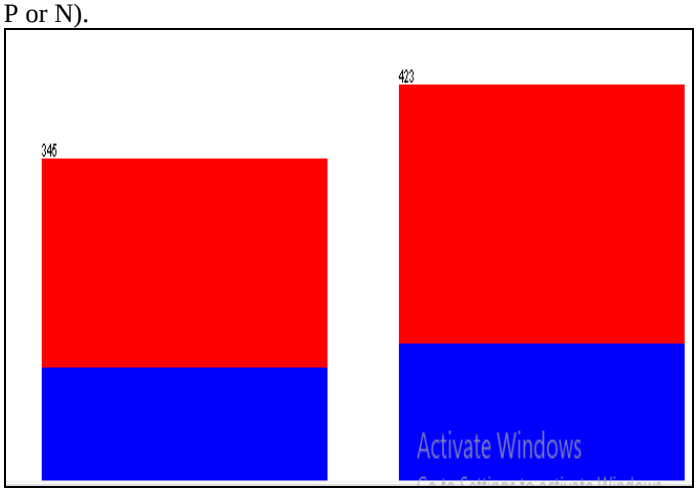


Fig 2: Pregnancies Attribute Visualization

2. Glucose Attribute: This attribute contains the range of Glucose measured in the particular case of a patient. The below figure shows three types of the values which is the minimum, maximum, mean & Standard Deviation values of the attribute value range.

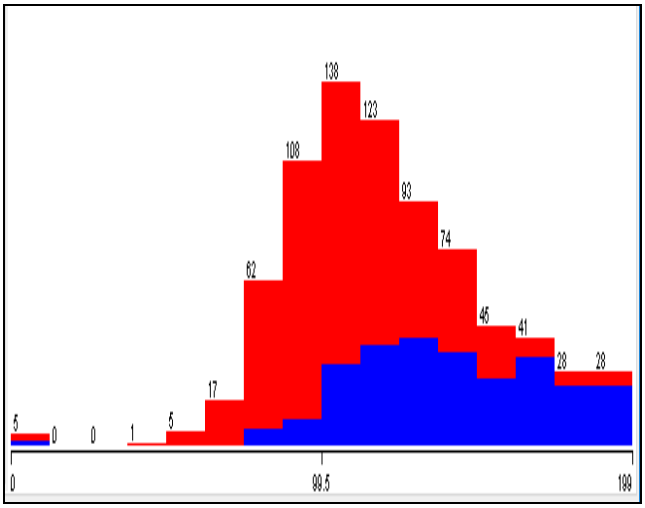


Fig 3: Glucose Attribute Visualization

3. Blood Pressure Attribute: This attribute contains the range of Blood Pressure measured in the particular case of a patient. The below figure shows three types of the values which is the minimum, maximum, mean & Standard Deviation values of the attribute value range.

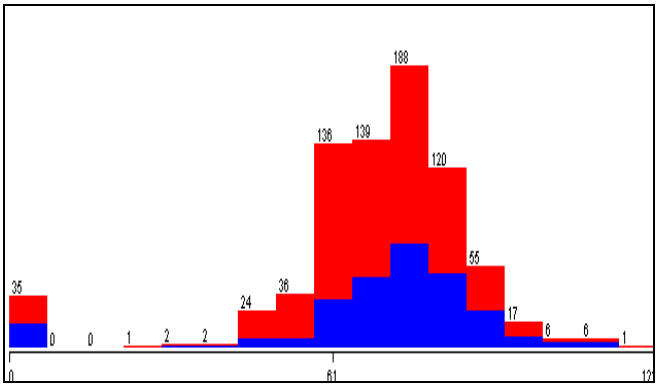


Fig 4: Blood Pressure Attribute Visualization

4. **Skin Thickness attribute:** This attribute contains the range of Skin Thickness measured in the particular case of a patient. The below figure shows three types of the values which is the minimum, maximum, mean & Standard Deviation values of the attribute value range.

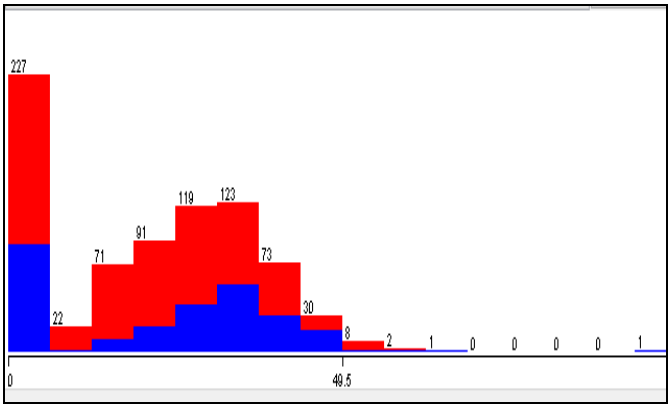


Fig. 5: Skin Thickness

5. **Insulin Attribute:** This attribute contains the range of Insulin measured in the particular case of a patient. The below figure shows three types of the values which is the minimum, maximum, mean & Standard Deviation values of the attribute value range.

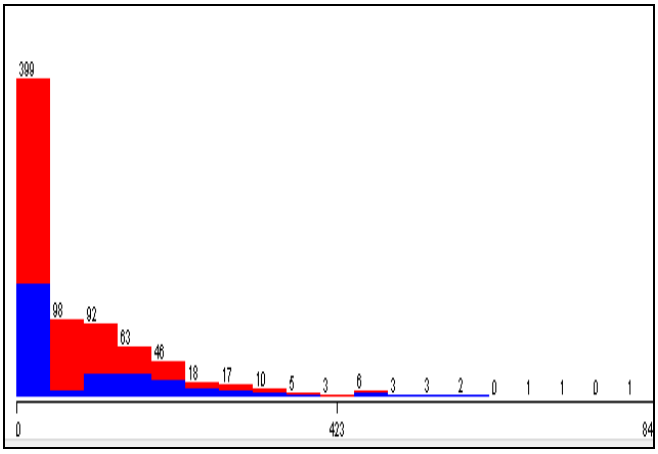


Fig. 6: Insulin Attribute

6. **BMI Attribute:** This attribute contains the range of BMI measured in the particular case of a patient. The below figure shows three types of the values which is the minimum, maximum, mean & Standard Deviation values of the attribute value range.

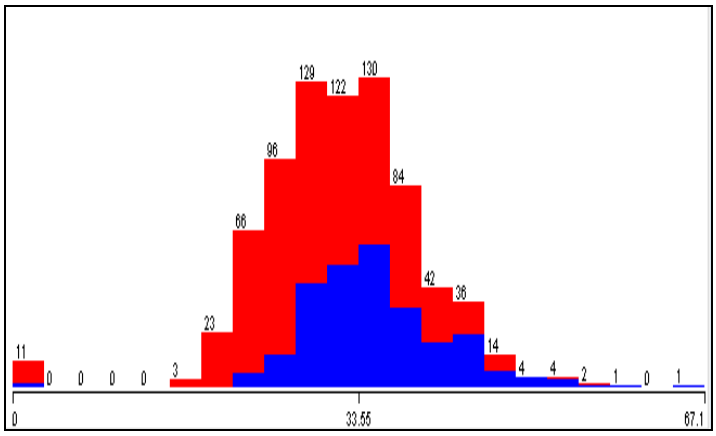


Fig.7: BMI Attribute

7. **Diabetes Pedigree Function Attribute:** This attribute contains the range of Diabetes Pedigree Function measured in the particular case of a patient. The below figure shows three types of the values which is the minimum, maximum, mean & Standard Deviation values of the attribute value range.

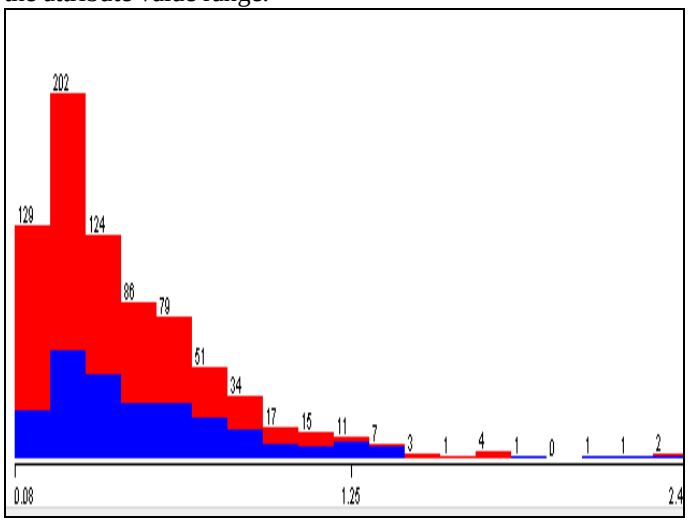


Fig. 8: Skin Thickness attribute

8. **Age Attribute:** This attribute contains the range of Diabetes age measured in the particular case of a patient. The below figure shows three types of the values which is the minimum, maximum, mean & Standard Deviation values of the attribute value range.

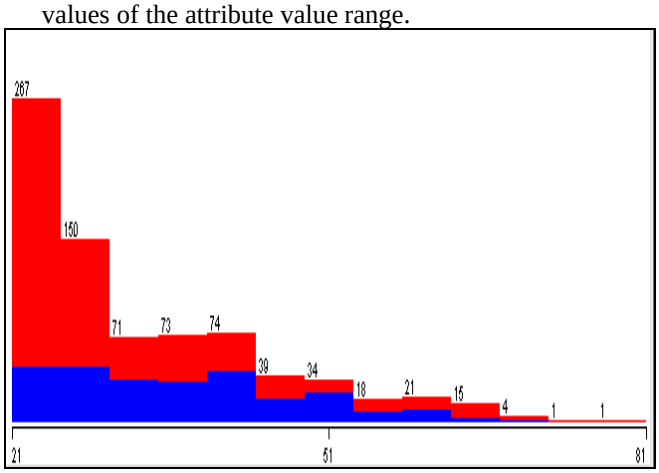


Fig. 9: Age attribute

9. **Outcome Attribute:** This attribute contains the status of Result in the particular case of a patient. If the Result is TRUE then the attribute value is TRUE else FALSE.

The below figure shows the count of each Value (TRUE or FALSE).

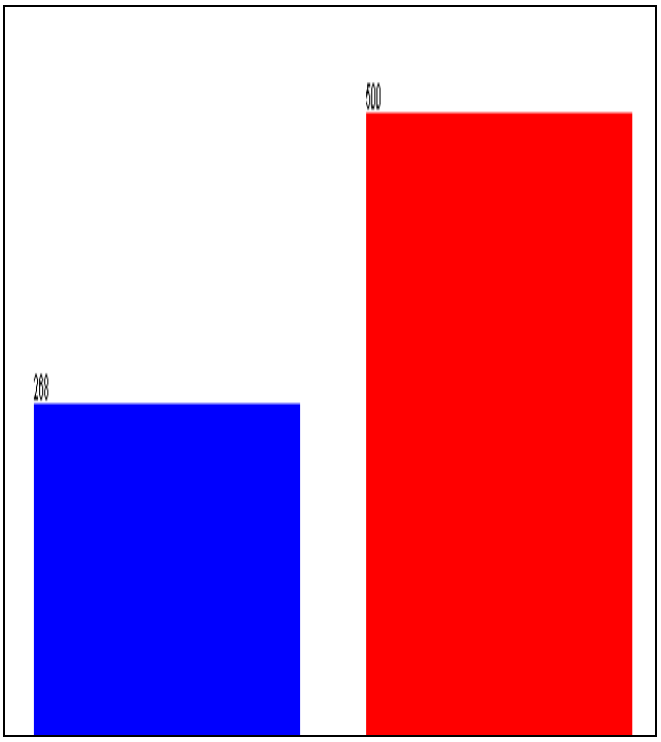
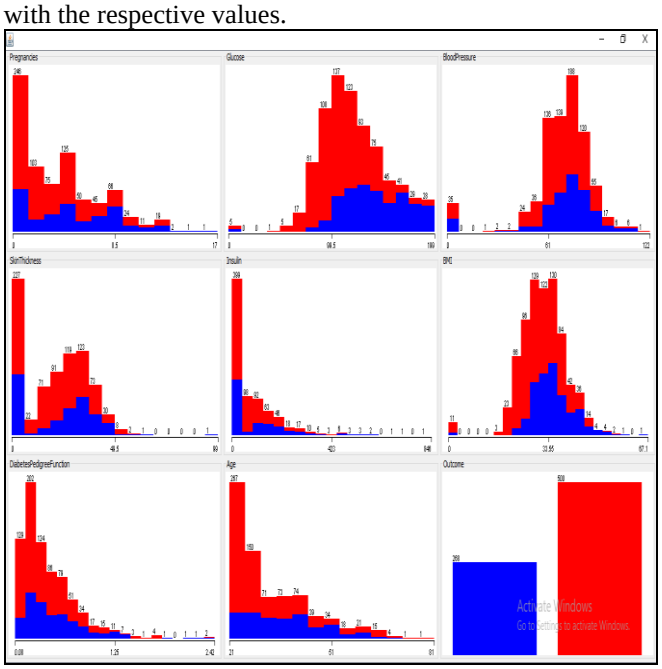


Fig.10: Outcome Attribute

Here we have a visualization of all attribute together with the respective values.



These all visualizations show the minimum, maximum, mean and std. Deviation value of the each attribute in graphical form. We can also get the information about the data type of the attribute in this section

V. CONCLUSION

In basic terms, a Naive Bayes classifier accept that the nearness of a specific feature in a class is inconsequential to the nearness of some other feature. When we get the Naive Bayes' Result then we apply the Clustering system on the equivalent dataset. Basic K-Means Clustering is utilized to get the clusters of the data results. We can picture the Cluster assignments for each property against the Resultant or expectation characteristic. We can have the better understanding through these perceptions about the conditions of characteristics on the expectation variable. K-implies algorithm is an iterative algorithm that attempts to parcel the dataset into Kpre-characterized unmistakable non-covering subgroups (clusters) where every datum point has a place with just one gathering. It attempts to make the between cluster data focuses as comparable as would be prudent while additionally keeping the clusters as various (far) as could be allowed. It assigns data focuses to a cluster to such an extent that the aggregate of the squared separation between the data focuses and the cluster's centroid (number juggling mean of the considerable number of data indicates that have a place that cluster) is at the base. The less variety we have inside clusters, the more homogeneous (comparative) the data focuses are inside a similar cluster. What's more, after conclusive examination of the consequence of the two procedures we discovered two properties which are having greatest load as contrast with others. These two traits Glucose and Insulin must consider in the diabetes forecast.

REFERENCES

1. Developing a CIHR Framework to Measure The Imact of Health Research. http://www.cihr-irsc.gc.ca/e/documents/meeting_synthesis_e.pdf
2. National Consensus Conference on Population Health Indicatores – Final Report. Canadian Institue for Health Information, Ottawa, 1999.
3. Heathy Canadians: A Federal Report on Comparable Health Indicatores, 2004. Health Canada, Ottawa.
4. Buxton M, S Hanney, T Jones 2004. Estimation the economic value to societies of the impact of health research: a critical review. Bulletin of the World Health Organiztion. 82(10):733-739.
5. Sharpe A, Smith J. (2005). Measuring the Impact of Research on Well-being: A Survey of Indicators of Well-being. Centre for the Study of Living Standards Report 2005-02
6. N. Bhargava, S. Dayma, A. Kumar and P. Singh, "An approach for classification using simple CART algorithm in WEKA," *2017 11th International Conference on Intelligent Systems and Control (ISCO)*, Coimbatore, 2017, pp. 212-216. doi: 10.1109/ISCO. 2017. 7855983
7. R.L. SimpsonBig data and nursing knowledge Nurs Adm Q, 39 (1) (2015), pp. 87-89
8. B. Buxton, V. Hayward, I. Pearson, L. Kärkkäinen, H. Greiner, E. Dys on, et al.Big data: the next Google Interview by Duncan Graham-Rowe Nature, 455 (7209) (2008), pp. 8-9
9. C.D. StrobelAmerican recovery and reinvestment act of 2009
10. J Corp Account Financ, 20 (5) (2009), pp. 83-85
11. J.T. Overpeck, G.A. Meehl, S. Bony, D.R. EasterlingClimate data challenges in the 21st century Science, 331 (6618) (2011), pp. 700-702
12. N. Bhargava, R. Purohit, S. Sharma and A. Kumar, "Prediction of arthritis using classification and regression tree algorithm," *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, 2017, pp. 606-610. doi: 10.1109/CESYS.2017.8321150

13. K. Jee, G.H. KimPotentiality of big data in the medical sector: focus on how to reshape the healthcare system Healthc Inform Res, 19 (2) (2013), pp. 79-85
14. Y.Y. PanConstruction of nursing consultation information system in the age of big data Medical Information, 27 (8) (2014), p. 10
15. C. Auffray, R. Balling, I. Barroso, L. Bencze, M. Benson, J. Bergeron, et al.Making sense of big data in health research: towards an EU action plan Genome Med, 8 (1) (2016), pp. 1-13
16. NIH.NINRBig data in symptoms research methodologies boot camp.[2017-02-18]
17. G.H. Zhou, Y. Xin, Y.J. ZhangStudy on big data's applications in medical and health field Chinese Journal of Health Information Management, 10 (4) (2013), pp. 296-300304.
18. B. SchwerdtleBig data in nurse education Nurse Educ Today, 51 (2016), pp. 114-116

AUTHORS PROFILE

Prof. Neeraj Bhargava, working as Professor in M.D.S University, Ajmer. He is Head of the Department of Computer Science and school of engineering and System Science, MDS University, Ajmer.He has more than 26 years of teaching experience and guided many research projetcs through out .He has been prominnat in teaching and research and his papers are having great impact among young researchers in India and Abroad.

Dr.Abhishek Kumar, is working as Assistant Professor in Chitkara University he is senior member IEE .He is having 8 years of teaching epxerience.He has authored and edited more than 13 books with reputated publishers like Wiley, Springer,CRC USA etc. Having

Dr. Ritu Bhargava, is working is Lecturer in Sophia girls' College, Ajmer.She has been senior acedmecian and prominent faculty of computer Science .She has benn teaching in many government and private firm as visiting faculty.

Shikha Bharadwaj, is research scholar in mjrp university.she is currently doing her research work in domain of data mining