International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 06
May 2017

Medical Decision Data Mining using Naïve Bayes & K-Means Clustering

PATHALLAPALLI RAKESH VARMA[1], PUVVALA MANI KRISHNA[2] & UMMADI JANARDHAN REDDY[3]

[1]B-Tech Dept. Information Technology Vignan's University Vadlamudi-522213 Guntur Andhra Pradesh India Mail Id: - varma.283@gmail.com

[2]B-Tech Dept. Information Technology Vignan's University Vadlamudi-522213 Guntur Andhra Pradesh India Mail Id: - pmanibvm@gmail.com

[3]Assistant Professor Dept. Information Technology Vignan's University Vadlamudi-522213 Guntur Andhra Pradesh India Mail Id: - ummadi.janardhan@gmail.com

Abstract

In this paper, by utilizing data mining we can evaluate many patterns which will be used in future to make keenly intellective systems and decisions By data mining refers to sundry methods of identifying information or the adoption of solutions predicated on cognizance and data extraction of these data so that they can be utilized in sundry areas such as decision-making, the presage value for the presage and calculation. In our days the health industry has amassed astronomical amounts of patient data, which, infelicitously, is not "engendered" in order to give some obnubilated information, and thus to make efficacious decisions, which are connected with the base of the patient's data and are subject to data mining. This research work has developed a Decision Support in Heart Disease Presage System (HDPS) utilizing data mining modelling technique, namely, Naïve Bayes and Kmeans clustering algorithms that are one of the most popular clustering techniques; however, where the initial cull of the centroid vigorously influences the final result. Utilizing of medical data, such as age, sex, blood pressure and blood sugar levels, chest pain, electrocardiogram, analyzes of different study patient, etc. graphics can presage the likelihood of the patient. This paper shows the efficacy of unsupervised learning techniques, which is a k-betokens clustering to ameliorate edifying methods controlled, which is ingenuous Bayes. It explores the integration of K-designates clustering with verdant Bayes in the diagnosis of disease patients. It withal investigates different methods of initial centroid cull of the K-designates clustering such as range, inlier, outlier, arbitrary attribute values, and desultory row methods in the diagnosis of heart disease patients. The results designate that the integration of the K-betokens clustering with naïve Bayes with different initial centroid culling naïve Bayesian amend precision in diagnosis of the patient.

Keywords: Data Mining, Naïve Bayes, K-Means Clustering.

1. INTRODUCTION

Data mining this is revelation process in the raw data antecedently unknown, non-frivolous, virtually utilizable, the interpretation of the available erudition indispensable for decision-making in the sundry spheres of human activity. This search for relationship with subsisting astronomically immense associated data that are obnubilated among astronomically immense amounts of data and refers to the "mining" cognizance from sizably voluminous amounts of data. Subsisting systems are habituated to avail in decision-making, referred to as data mining. These systems represent an iterative sequence of pre-processing as cleaning, data integration, and data cull is veridical the pattern identification of data mining and erudition representation. Data mining is the search for relationships and ecumenical patterns that subsist in astronomically immense databases, but obnubilated among the plethoras of data. Computer diagnosis of diseases is the medico for the same instrument, the calculations for an engineer: design diagnostics does not supersede the medico, but it avails. The practice of examining immensely colossal preexisting data bases in order to engender incipient information. It coverts raw data into subsidiary information. It analyzes the data for relationships that have not antecedently been discovered. The steps of data mining are: Data cleaning, data integration, data cull, data transformation, data mining, pattern evaluation and cognizance representation. Medical data mining is a domain of lot of imprecision and skepticality. The clinical decisions are conventionally predicated on the medicos intuition. Consequently this may lead to disastrous consequences. Due to this there are many errors in the clinical decisions and it results in extortionate medical costs. Serialization is withal utilized in this system. It converts the data objects into streams of bytes and stores it into database.

2. RELATED WORK

Many hospital information systems are designed to fortify patient billing, inventory management and generation of simple statistics. Some hospitals use decision support systems, but they are largely constrained. They can answer simple queries like “What is the average age of patients who have heart disease?”, “How many surgeries had resulted in hospital stays longer than 10 days?”, “Identify the female patients who are single, above 30 years old, and who have been treated for cancer.” However, they cannot answer intricate

queries like "Identify the paramount Preoperative prognosticators that increase the length of hospital stay", "Given patient records on cancer, should treatment include chemotherapy alone, radiation alone, or both chemotherapy and radiation?", and "Given patient records, soothsay the probability of patients getting a heart disease." Clinical decisions are often made predicated on doctors‟ intuition and experience rather than on the erudition- opulent data obnubilated in the database. This practice leads to unwanted biases, errors and extortionate medical costs which affects the quality of accommodation provided to patients. Wu, et al proposed that integration of clinical decision support with computerbased patient records could reduce medical errors, enhance patient safety, decrement unwanted practice variation, and ameliorate patient outcome. This suggestion is promising as data modeling and analysis implements, e.g., data mining, have the potential to engender a cognizance-opulent environment which can avail to significantly amend the quality of clinical decisions

3. IMPLEMENTATION

Naïve Bayes Algorithm

Ingenuous Bayes classifier can be trained in supervised learning setting. It utilizes the method of maximum kindred attribute. It has been worked in involute authentic world situation. It requires iota of training data. It estimates parameters for relegation. Only the variance of variable need to be tenacious for each class not the entire matrix. Naïve bayes is mainly used when the inputs are high. It gives output in more sophisticated form. The probability of each input attribute is shown from the prognosticable state. Machine learning and data mining methods are predicated on naïve bayes relegation. Naïve bayes will rudimentally soothsay the output whether the patient will have chances of getting the heart disease or not. The model dataset which we get after applying K-Betokens algorithm will compared the values of dataset with a trained dataset. It will apply the bayes theorem and the probability will be obtained whether the patient will have heart disease or not

Algorithm Steps

Which one is the best predictor ?

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Fig 1 Sample Medical Data Set values

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 06
May 2017

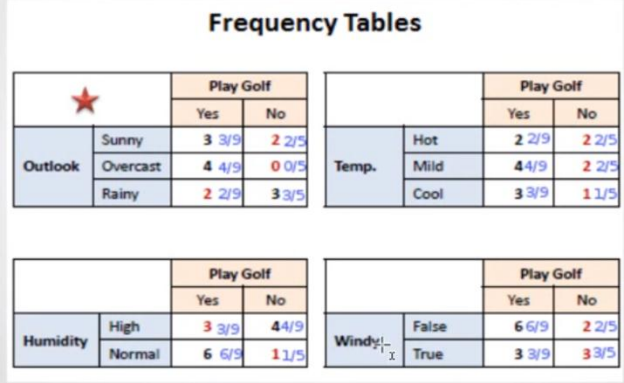

Fig 2 Frequency Tables from Data Set values

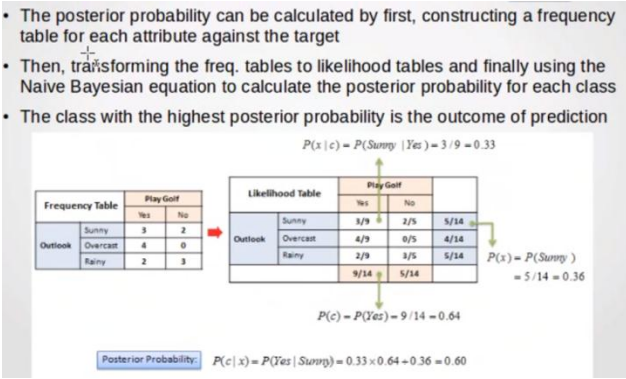

Fig 3 Posterior Probability Calculation

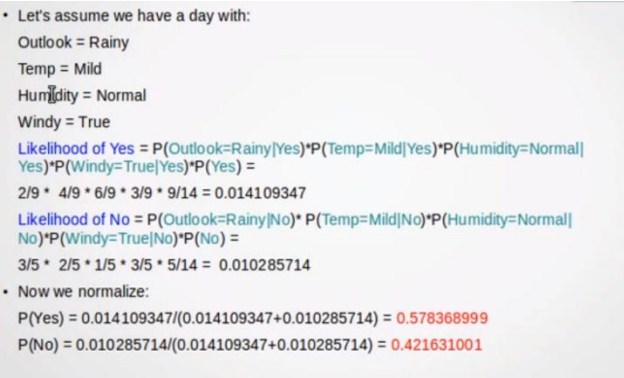

Fig 4 Result

4. K-MEANS CLUSTERING

K-denotes is simplest learning algorithm to solve the clustering quandaries. The process is simple and facile, it relegates given data set into certain number of clusters. It defines k centroids for each cluster. They must be placed as much as possible far away from each other. Then take each point belonging to given data set and relate into the most proximate centroid. If no point is pending then an group age is done. Then we re-calculate k incipient centroid for the cluster resulting from anterior steps. When we get the k centroid an incipient binding is to be done between lucid data points and most proximate centroid. A loop is been engendered because of this loop key centroid transmute the location step by step until no more changes are done.

Algorithm Steps

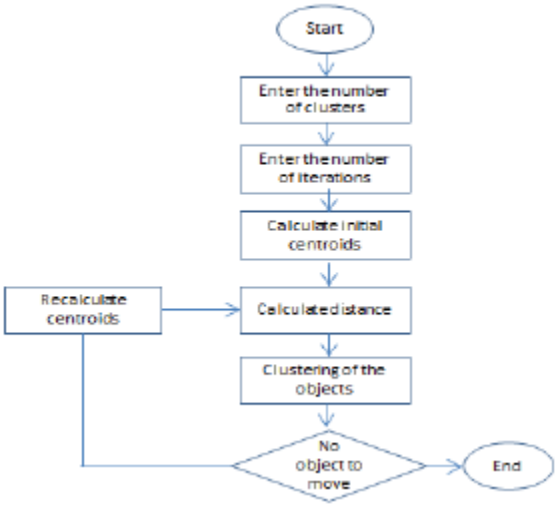

Fig 5 K-means clustering algorithm

5. EXPERIMENTAL RESULTS

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 06
May 2017

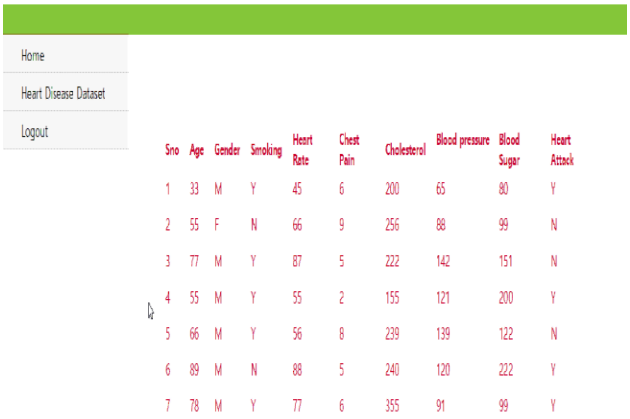

Fig 6 Medical Data Set

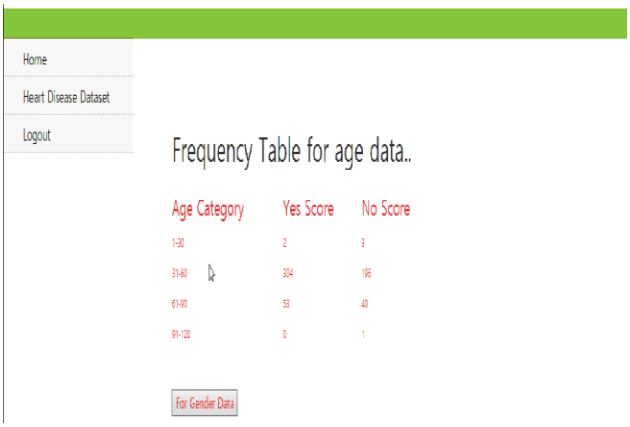

Fig 7 Frequency Table for Age Data

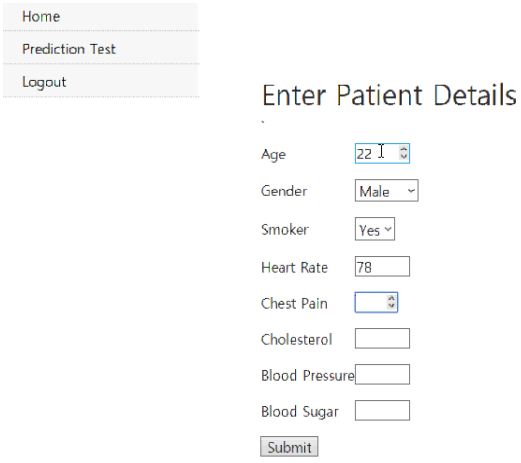

Fig 8 Input values

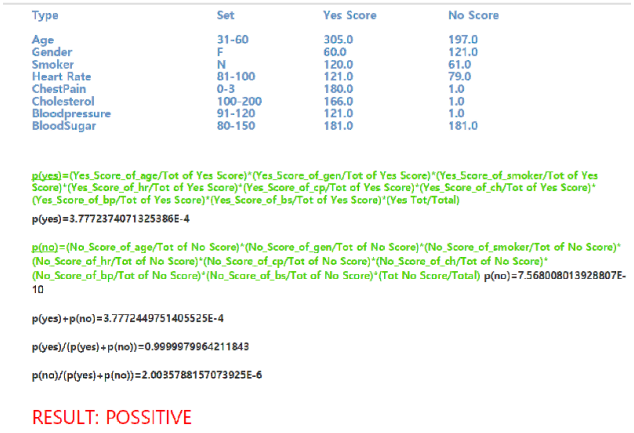

Fig 9 Result after Mining

6. CONCLUSION

In this paper we are proposing heart disease prognostication system utilizing naïve bayes and k-designates clustering. We are utilizing k-betokens clustering for incrementing the efficiency of the output. This is the most efficacious model to prognosticate patients with heart disease. This model could answer intricate queries, each with its own vigor with deference to facilitate of model interpretation, access to detailed information and precision

7. REFERENCES

[1] Sellappan Palaniappan, Rafiah Awang "Intelligent Heart Disease Prediction System Using Data Mining Techniques"Department of Information Technology Malaysia University of Science and Technology Block C, Kelana Square, Jalan SS7/26 Kelana Jaya, 47301 Petaling Jaya, Selangor, Malaysia .

[2] "CSV File Reading and Writing" (http:/ / docs. python. org/ library/ csv. html). . Retrieved July 24, 2011. "is no "CSV standard""

[3] Y. Shafranovich. "Common Format and MIME Type for Comma- Separated Values (CSV) Files" (http:/ / tools. ietf. org/ html/ rfc4180) Retrieved September 12, 2011.

[4] home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html "A tutorial on clustering algorithms".

[5] Shadab Adam Pattekari and Asma Parveen "Prediction System For Heart Disease Using Naïve Bayes" International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294.

[6] Mrs.G.Subbalakshmi (M.Tech), Mr. K. Ramesh M.Tech, Asst. Professor Mr. M. Chinna Rao M.Tech,(Ph.D.) Asst. Professor, "Decision Support in Heart Disease Prediction System using Naïve Bayes" G.Subbalakshmi et al. / Indian Journal of Computer Science and Engineering (IJCSE)2011.

[7] Jesmin Nahar, Tasadduq Imama, Kevin S. Tickle, Yi-Ping Phoebe Chen "Association rule mining to detect factors which contribute to heart disease in males and females" Expert Systems with Applications 40 (2013) 1086–1093.

[8] Oleg Yu. Atkov (MD, PhD), Svetlana G. Gorokhova (MD, PhD), Alexandr G. Sboev (PhD), Eduard V. Generozov (PhD), Elena V. Muraseyeva (MD, PhD), Svetlana Y. Moroshkina,Nadezhda N. Cherniy "Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters" Journal of Cardiology (2012) 59, 190—194.

[9] Shantakumar B.Patil Y.S.Kumaraswamy "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network" European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656.

[10] Sivagowry, Dr. Durairaj. M2 and Persia. "An Empirical Study on applying Data Mining Techniques for the Analysis and Prediction of Heart Disease" 2013.