

MEDICAL DATA SETS USING K-MEAN CLUSTERING AND NAIVE BAYES

K. VENKATA SUBBA REDDY

Department of Computer Science and Engineering

Muffakham Jah College of Engineering and Technology, Hyderabad.

E Mail: kvsreddy2012@gmail.com

Abstract

By utilizing data referral this paper will judge several patterns which can be used in future to form keenly intellectual systems and selections. By data referral refers to likewise ways of characteristic data or the adoption of solutions predicated on cognizance and information extraction of that information, so they will be used in likewise areas like decision-making, the presale price for the presage and calculation. In our days the health trade has concentrated astronomical amounts of patient information, which, infelicitously, is not engendered to administer some ventilated data, and therefore to form efficacious selections, that area unit connected with the bottom of the patient's information and area unit subject to data referral. This analysis work has developed a call Support in cardiograph Presage System, utilizing data reformulating technique, namely, Naive Thomas Bases and K-means clod algorithms that area unit one amongst the foremost in style clod techniques; but, wherever the initial cull of the center of mass smartly influences the ultimate result. Utilizing of medical information, like age, sex, pressure and glucose levels, chest pain, graphical record, categories of various study patient, etc. Graphics will presage the chance of the patient. This paper shows the effectualness of unsupervised learning techniques, which could be a k-betokens clod to ameliorate edifying histrionically that is obvious Thomas Bases. It explores the mixing of K-designates clod with abundant Thomas Bases within the diagnosing of wellness patients. It investigates different ways of an initial center of the mass cull of the K-designates clod like thievery inline, an outlier, capricious attribute values, and purposeless row ways within the diagnosing of cardiograph patients. The results designate that the mixing of the K-betokens clod with naive Thomas Bases

with a totally different initial center of mass culling naive Bayesian amend exactitude in diagnosing of the patient.

Keywords: K-Means Clustering, Naive Bayes, Medical Data Set.

1. INTRODUCTION

Data mining will be revelation technique inside the knowledge antecedently unknown, non-frivolous, with reference to serviceable, the interpretation of the accessible ruddiness indispensable for decision-making inside the sundry spheres of act. This hunt for a relationship with subsisting astronomically massive associated data that unit of measurement ventilated among astronomically massive amounts of knowledge and refers to the "mining" cognizance from sizable voluminous amounts of knowledge. Subsisting systems unit of measurement habituated to avail in decision-making, noted as processing. These systems represent Associate in Nursing iterative sequence of pre-processing as cleansing, data integration, and data cull is real the pattern identification of knowledge mining and ruddiness illustration. Processing is that the hunt for relationships and ecumenical patterns that survives in astronomically massive databases, but ventilated among the plethora of knowledge. Laptop identification of diseases is that the medico for an identical

instrument, the calculations for Associate in Nursing engineer: vogue medical science does not follow the medico, but it avails. The observe of examining immensely large pre-existing databases thus on engender early information. It coverts info into subsidiary information. It analyzes the information for relationships that haven't antecedently been discovered. The steps {information of the data} mining are knowledge cleansing, data integration, data cull, data transformation, processing, pattern analysis and cognizance illustration. Medical processing may well be a site of a ton of inexactness and skeptically. The clinical choices unit of measurement conventionally predicated on the medicos intuition. Consequently, this might end in unfortunate consequences. As a result of this there unit of measurement many errors inside the clinical choices and it ends up in steep medical costs. Publication continues to be used throughout this method. It converts objects into streams of bytes and stores it into info.

2. RELATED WORK

Many hospital information systems unit designed to fortify patient asking, inventory management and generation of simple statistics. Some hospitals use decision support systems, but they are for the foremost half forced. They are going to answer simple queries like “ What is that the common age of patients, United Nations agency have heart disease, “ How many surgeries had resulted in hospital stays longer than 10 days, “ Identify the female patients United Nations agency unit single, above thirty years previous, and United Nations agency area unit treated for cancer. However, they can't answer knotty queries like Identify the predominant surgical prognosticators that increase the length of hospital stay, Given patient records on cancer, have to be compelled to treatment embody medical aid alone, radiation alone, or every medical aid and radiation, and “ Given patient records, soothsay the probability of patients getting a upset.” Clinical choices unit usually created predicated on doctors intuition and ability rather than on the erudition- deluxe info ventilated inside the information. This follows winds up in unwanted biases, errors and steep medical costs that affect the

quality of accommodation provided to patients. Integration of clinical decision support with computerized patient records might reduce medical errors, enhance patient safety, decrement unwanted follow variation, and ameliorate patient outcome. This suggestion is promising as info modeling and analysis implements, e.g., processing, have the potential to engender a cognizance-opulent atmosphere which can avail to significantly amend the quality of clinical choices.

3. IMPLEMENTATION

Naive Bayes Algorithm

Ingenuous mathematician classifier may be trained in a supervised learning setting. It utilizes the tactic of the most kindred attribute. It's been worked in involutes authentic world state of affairs. It needs an iota of coaching knowledge. It estimates parameters for relegation. Solely the variance of a variable has to be compelled to be tenacious for every category, not the whole matrix. A naive mathematician is principally used once the inputs area unit high. It offers output in additional subtle kind. The chance of every input attribute is shown from the prognostic able state. Machine learning and data processing

strategies area unit predicated on naive mathematician relegation. A naive mathematician can radiotelephones say the output whether or not the patient can have probabilities of obtaining the center unlawfulness or not. The model dataset that we tend to get when applying K-Betokens formula can be compared the values of a dataset with a trained dataset. Apply the mathematician theorem and also the chance is going to be obtained whether or not the patient will have cardiograph or not.

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

Table 1 Sample Medical Data Set values

		Play Golf	
		Yes	No
Outlook	Sunny	3 (3/9)	2 (2/5)
	Overcast	4 (4/9)	0 (0/5)
	Rainy	2 (2/9)	3 (3/5)

Table 2 Frequency Table of Outlook

		Play Golf	
		Yes	No
Temp	Hot	2 (2/9)	2 (2/5)
	Mild	4 (4/9)	2 (2/5)
	Cool	3 (3/9)	1 (1/5)

Table 3 Frequency Table of Temp

		Play Golf	
		Yes	No
Humidity	High	3 (3/9)	4 (4/5)
	Normal	6 (6/9)	1 (1/5)

Table 4 Frequency Table of Humidity

		Play Golf	
		Yes	No
Windy	False	6 (6/9)	2 (2/5)
	True	3 (3/9)	3 (3/5)

Table 5 Frequency Table of Windy

		Play Golf	
		Yes	No
Outlook	Sunny	3 (3/9)	2 (2/5)
	Overcast	4 (4/9)	0 (0/5)
	Rainy	2 (2/9)	3 (3/5)

$$P(x/c) = P(\text{Sunny/Yes}) = 3/9 = 0.33$$

		Play Golf		
		Yes	No	
Outlook	Sunny	3 (3/9)	2 (2/5)	5/14
	Overcast	4 (4/9)	0 (0/5)	4/14
	Rainy	2 (2/9)	3 (3/5)	5/14
		9/14	5/14	

$$P(c) = P(\text{yes}) = 9/14 = 0.64$$

$$P(x) = P(\text{Sunny}) = 5/14 = 0.36$$

Posterior Probability P

$$(c/x) = P(\text{Yes/Sunny}) = 0.33 * 0.64 / 0.36 = 0.60$$

Table 6 Posterior Probability Calculation

The experimental results shown by
 calculating the Posterior probability

Let's assume we have a day with:

Outlook = Rainy

Temp= Mild

Humidity = Normal

Windy = True

Likelihood of Yes=P

$$(P(\text{Outlook}=\text{Rainy}|\text{Yes}) * P(\text{Temp}=\text{Mild}|\text{Yes}) * P(\text{Humidity}=\text{Normal}|\text{Yes}) * P(\text{Windy}=\text{True}|\text{Yes})) * P(\text{Yes}) = \frac{2}{9} * \frac{4}{9} * \frac{6}{9} * \frac{3}{9} * \frac{9}{14} = 0.014109347$$

Likelihood of No =

$$(P(\text{Outlook}=\text{Rainy}|\text{No}) * P(\text{Temp}=\text{Mild}|\text{No}) * P(\text{Humidity}=\text{Normal}|\text{No}) * P(\text{Windy}=\text{True}|\text{No})) * P(\text{No}) = \frac{3}{9} * \frac{2}{5} * \frac{1}{5} * \frac{3}{5} * \frac{5}{14} = 0.010285714$$

Now we normalize:

$$P(\text{Yes}) = \frac{0.014109347}{(0.014109347 + 0.010285714)} = 0.578368999$$

$$P(\text{No}) = \frac{0.010285714}{(0.014109347 + 0.010285714)} = 0.421631001$$

4. K-MEANS CLUSTERING

K-denotes is simplest learning formula to unravel the agglomeration quandaries. The method is straightforward and facile, it

relegates given knowledge set into a bound variety of clusters. It defines k centered for every cluster. They have to be placed the maximum amount as potential secluded from one another. Then take every purpose happiness to given knowledge set and relate into the foremost proximate center of mass. If no purpose is unfinished then associate in nursing cluster age is completed. Then we tend to re-calculate k inchoate center of mass for the cluster ensuing from anterior steps. Once we get the k center of mass Associate in nursing inchoate binding is to be done between lucid knowledge points and most proximate center of mass. A loop is been engendered as a result of this loop key center of mass transmute the situation step by step till no additional changes area unit done.

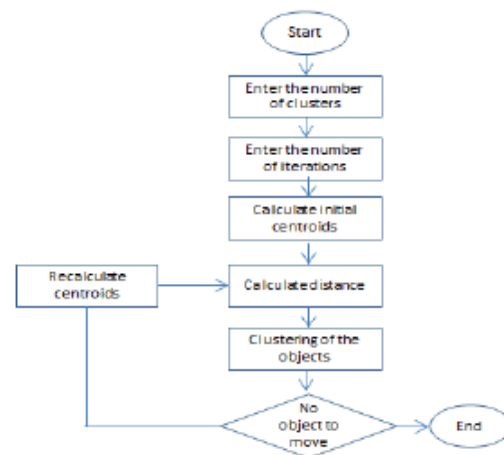
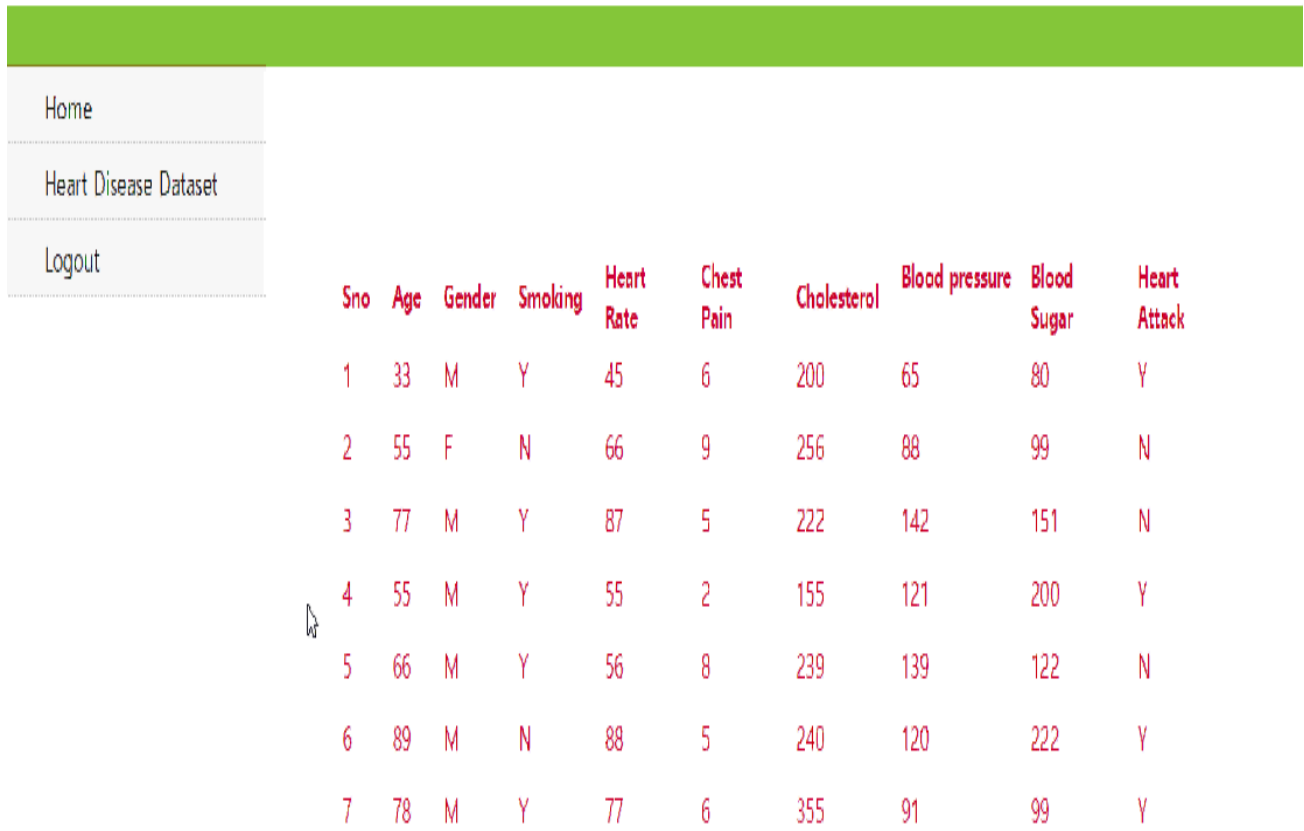


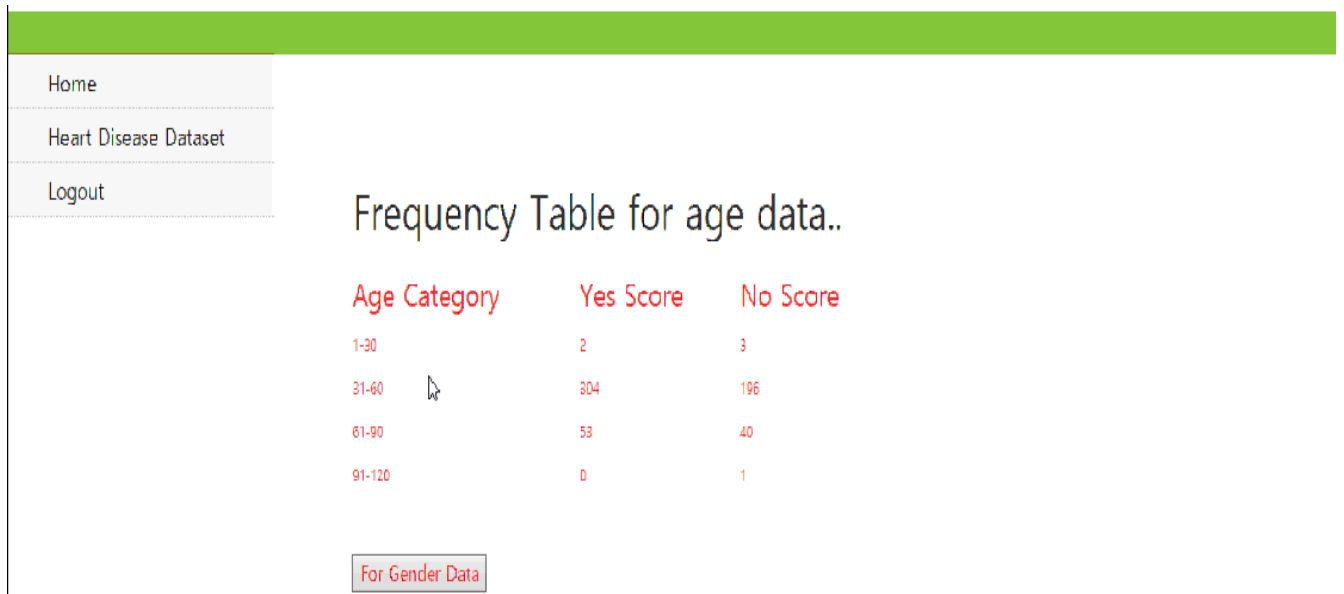
Fig 1 K-means clustering algorithm

5. EXPERIMENTAL RESULTS



Sno	Age	Gender	Smoking	Heart Rate	Chest Pain	Cholesterol	Blood pressure	Blood Sugar	Heart Attack
1	33	M	Y	45	6	200	65	80	Y
2	55	F	N	66	9	256	88	99	N
3	77	M	Y	87	5	222	142	151	N
4	55	M	Y	55	2	155	121	200	Y
5	66	M	Y	56	8	239	139	122	N
6	89	M	N	88	5	240	120	222	Y
7	78	M	Y	77	6	355	91	99	Y

Fig 2 Medical Data Set



Frequency Table for age data..

Age Category	Yes Score	No Score
1-30	2	3
31-60	304	196
61-90	53	40
91-120	0	1

For Gender Data

Fig 3 Frequency Table for Age Data

Type	Set	Yes Score	No Score
Age	31-60	305.0	197.0
Gender	F	60.0	121.0
Smoker	N	120.0	61.0
Heart Rate	81-100	121.0	79.0
ChestPain	0-3	180.0	1.0
Cholesterol	100-200	166.0	1.0
Bloodpressure	91-120	121.0	1.0
BloodSugar	80-150	181.0	181.0

$p(\text{yes}) = (\text{Yes_Score_of_age} / \text{Tot of Yes Score}) * (\text{Yes_Score_of_gen} / \text{Tot of Yes Score}) * (\text{Yes_Score_of_smoker} / \text{Tot of Yes Score}) * (\text{Yes_Score_of_hr} / \text{Tot of Yes Score}) * (\text{Yes_Score_of_cp} / \text{Tot of Yes Score}) * (\text{Yes_Score_of_ch} / \text{Tot of Yes Score}) * (\text{Yes_Score_of_bp} / \text{Tot of Yes Score}) * (\text{Yes_Score_of_bs} / \text{Tot of Yes Score}) * (\text{Yes Tot} / \text{Total})$

$p(\text{yes}) = 3.7772374071325386E-4$

$p(\text{no}) = (\text{No_Score_of_age} / \text{Tot of No Score}) * (\text{No_Score_of_gen} / \text{Tot of No Score}) * (\text{No_Score_of_smoker} / \text{Tot of No Score}) * (\text{No_Score_of_hr} / \text{Tot of No Score}) * (\text{No_Score_of_cp} / \text{Tot of No Score}) * (\text{No_Score_of_ch} / \text{Tot of No Score}) * (\text{No_Score_of_bp} / \text{Tot of No Score}) * (\text{No_Score_of_bs} / \text{Tot of No Score}) * (\text{Tot No Score} / \text{Total})$

$p(\text{no}) = 7.568008013928807E-10$

$p(\text{yes}) + p(\text{no}) = 3.7772449751405525E-4$

$p(\text{yes}) / (p(\text{yes}) + p(\text{no})) = 0.9999979964211843$

$p(\text{no}) / (p(\text{yes}) + p(\text{no})) = 2.0035788157073925E-6$

RESULT: POSSITIVE

Fig 4 Experimental Results after Mining

6. CONCLUSION

In this paper propose heart condition prognostication system utilizing naive Bases and k-designates agglomeration. This paper tends to utilize k-betokens agglomeration for incrementing the efficiency of the output. This is often the

foremost efficacious model to prognosticate patients with a heart condition. This model might answer tortuous queries, each with its own vigor with deference to facilitate of model interpretation, access to detailed knowledge and preciseness.

REFERENCES

[1] SellappanPalaniappan, RafiahAwang “Intelligent Heart Disease Prediction System Using Data Mining Techniques, Malaysia University of Science and Technology, Malaysia.

[2] Y. Shafranovich. "Common Format and MIME Type for Comma- Separated Values (CSV) Files”, September 12, 2011.

[3] Shadab Adam Pattekari and AsmaParveen “Prediction System For Heart Disease Using Naive Bayes” *International Journal of Advanced Computer and Mathematical Sciences* ISSN 2230-9624.Vol 3, Issue 3, 2012, pp 290-294.

[4] Mrs.G.Subbalakshmi, Mr. K. Ramesh, Mr. M. ChinnaRao, “Decision Support in Heart Disease Prediction System using Naive Bayes”, *Indian Journal of Computer Science and Engineering (IJCSE)*2011.

[5] JesminNahar, TasadduqImama, 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.

[6] Oleg Yu. Atkov, Svetlana G. Gorokhova , Alexandr G. Sboev, Eduard V. Generozov , Elena V. Muraseyeva, 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.

[7] ShantakumarB.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.

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