

Big Data Analytics in Diabetes Disease Data Prognosis Using Naïve Bayes Classifier Algorithm

¹R.THAYAMMAL,

Research Scholar,
School of computing science,
Vels Institute of Science, Technology and Advanced Studies (VISTAS)
E mail: thayammal.rajavel@gmail.com

²S.KAMALAKKANNAN,

Assistant Professor,
Vels Institute of Science, Technology and Advanced Studies (VISTAS)
E-mail: kamalindia81@yahoo.com

³P.Kavitha

Assistant Professor,
St. Joseph's College, Chennai,
E-mail: pkavikamal@gmail.com

Abstract:

In that paper, we've an inclination to project as checking the whole patient ill health unwellness victimization naïve Bayes classification and J48 decision tree. Huge process, as a result of the information comes from multiple, heterogeneous, autonomous sources with sophisticated and evolving relationships, and keeps growing. So in that, we'll take results of what proportion share patients get ill health as a positive knowledge and negative knowledge. Huge info is difficult to work with victimization most database management systems and desktop statistics and internal representation packages. The projected shows a huge process model, from the data mining perspective. Victimization classifiers, we've an inclination to unit method congenital disease share and values unit showing as a confusion matrix. We've an inclination to projected a replacement classification theme which could effectively improve the classification performance inside the situation that employment dataset is out there. During this dataset, we have nearly 500 patient details. We'll get all that details from there. Then we have a tendency to unit attending to sensible and unhealthy values square measure victimization naïve Bayes classifier and J48 tree.

Keywords: *Data analytics, Naive Bayes classification, J48 tree, diabetes dataset.*

Introduction:

Diabetes can be a chronic condition that is brought on by awkwardness inside the discharge of agent transportation regarding associate unsettling influence inside the sugar levels of the blood [1]. As per the yank inherited disease Association, 29.1 million individuals inside the u. s. (9.3% of the population), had inherited disease in 2012. Among this population, there are eight.1 million individuals administrative unit were undiscovered [2]. The incidence

of inherited disease continues increasing, with 1.7 million new analyses created in 2012. What's further, eighty six million people quite twenty years aged had pre inherited disease, and can be thought to be in associate extremely state inside the centre of health and chronic illness. Inherited disease remains the seventh driving reason for death inside the u. s. in 2010, with 69,071 deaths connected squarely to it, and many further (an mixture

of 234,05) having it as a result of the fundamental or contributing cause [5].

The term upset can be a chronic condition which contains the numerous diseases that influence the middle. Upset was the \$64000 reason for losses inside the distinctive nations still as Bharat. It kills one individual at regular intervals inside the US world Health Organization has evaluated that twelve million deaths happen around the world, consistently owing to the centre diseases [3]. Associate outsized portion of the deaths inside the US and completely different created nations happen owing to cardio tube diseases [6]. It's likewise the foremost reason of deaths in varied creating nations. All in all, it's viewed as a result of the essential rationalization for deaths in grown-ups. throughout this paper, we have a tendency to tend to stand live that concentrate on decision tree rule, Naïve man of science rule, and support vector machine to analysis inherited disease datasets[4]. Experiments square measure performed exploitation UCI machine learning datasets. Performance in addition has been evaluated for all the methodologies.

Datasets and Methodologies:

For diabetic malady, study uses the data got from the centre for Machine Learning and Intelligent Systems at UCI [10]. It incorporates clinical records from over 100,000

Individual encounters completely different with over sixty, specific patients. The data was accumulated from over seventy four million encounters related to seventeen million patients [11]. It absolutely was gathered over a time of 10 years, from 1999 to 2008, and contains varied credits that

compare to the seasons of confirmation and unleash of diabetic patients. These records contain data regarding [7] entirely completely different centre tests and systems, determination, and medications that were regulated inside the length of the healing facility keep. For heart disease, we've an inclination to use on-line on the market heart disease datasets from UCI (University of CA, Irvine C.A) Machine Learning Repository [12]. These dataset have constant qualities and example vogue. These datasets have seventy six unrefined qualities beside anticipated property even so entirely fourteen of them unit entirely elementary [8]. Cleveland Clinic Foundation data set contain 303 patients record and Hungarian Institute of medicine data set contains 5 hundred patient records.

i. Naïve Thomas Bayes Classifier:

Naïve Bayes Thomas man of science Classifier could also be a term managing basic probabilistic classifier in light-weight of applying Bayes Theorem with solid freedom suppositions. It expects that the closeness or nonattendance of specific part of a class is random to the closeness or event of notwithstanding different part [13]. The Naive Bayes calculation depends on conditional chances. It uses Bayes' hypothesis, a direction that ascertains probability by enumeration the frequency of qualities and blends of qualities inside the recorded info. Theorem finds the probability of an occurrence happening given the probability of another event that has as of presently happened [14]. Main advantage of the Naive Bayes classifier is that it desires slightly lives of employment info to assess the parameters (means and variances of the variables) vital for

classification. Since freelance variables unit assumed, merely the variances of the variables for every class have to be compelled to be calculated and not the complete.

ii. Decision Tree J48:

Decision trees represent to a directed because of deal with classification. a selection tree could also be a basic tree structure where non-terminal nodes represent tests on one or lots of attributes And terminal nodes mirror decision results [9]. The basic decision tree induction formula ID3 was improved by J48. Knowledge gain and gain magnitude relation measures unit used by ID3 and J48 as rending model on a private basis [15].

The abstract of decision tree calculation is given at a lower place.

- 1) Select associate attribute that best separates the output attribute values.
- 2) Manufacture a singular limb for estimation of the chosen attribute.
- 3) Divide the instances into subgroups therefore on mirror the attribute values of the chosen node.
- 4) For every subgroup, end the attribute determination technique if:
 - a) The individuals from a subgroup have identical value for the output attribute, end the attribute choice procedure for this path and mark the branch on this path with the planned value.
 - b) The subgroup contains a solitary node or no any distinctive attributes could also be resolved.

- 5) For every subgroup created during this has not been marked as terminal, repeat the on prime of procedure.

Related Work:

In this existing technique is we have a tendency to tend to used SVM rule for calculate that congenital disease details. SVM rule is slow technique for classify all the given details. During this main disadvantage is time efficiency. Patient's congenital disease details begin with large-volume, heterogeneous, autonomous sources with distributed and localized management, and obtain to explore advanced and evolving relationships among data. Existing system shows a SVM [Support Vector Machine] theorem that characterizes the choices of the large data revolt, and implements an enormous process model. It will take longer to calculate our dataset that we have a tendency to tend to own our congenital disease dataset. Discover patterns victimization SVM rule that verify the only mode of treatment for congenital disease across fully completely different age. They over that drug treatment for patients inside the young cluster are delayed whereas; patients inside the adulthood cluster have to be compelled to be prescribed drug treatment currently. Prediction and classification of varied reasonably congenital disease victimization C4.5 classification rule was administrated in Pima Indians congenital disease info. Huge data applications unit featured with autonomous sources and localized controls, aggregating distributed data sources to a centralized web site for mining is systematically preventive because of the potential transmission worth and privacy problems. Huge data quality is portrayed in

many aspects, beside advanced heterogeneous data types, advanced intrinsic linguistics associations in data, and complex relationship networks among data. That's to say, the price of giant data is in its quality.

Experimental Results:

To examine the execution of classification, the accuracy and FTO measures square measure embraced. Four cases square measure thought-about because the consequence of classifier.TP (True Positive): the amount of samples accurately classified to it class.TN (True Negative): the amount of samples accurately rejected from that category. FP (False Positive): the amount of samples inaccurately rejected from that category. FN (False Negative): the amount of samples inaccurately classified to it class.

The level of adequacy of the classification model is observed with the amount of correct and incorrect classifications in conceivable estimation of the variables being classified. Accuracy is calculated by mistreatment below. We've used medical datasets like polygenic disease and heart condition datasets. Of these datasets square measure taken from UCI machine learning repository datasets [10, 12]. The goal is to classify the diseases supported parameter choice and to check all the classifiers like Naïve Thomas Bayes, SVM, call tree. Accuracy of various classifiers for polygenic disease dataset is given in Table I and Table II severally.

TABLE I. ACCURACY OF polygenic disease DATASET

SI.NO	Classifiers	Accuracy %
-------	-------------	------------

1	Naïve Bayes	93.85
2	Support Vector Machine(SVM)	95.2
3	Decision Tree	92.59

According to Figure one, we will clearly see that Naïve Thomas Bayes provides highest accuracy 93.85%. Just in case of polygenic disease, call tree provides the best accuracy 92.59%. Support Vector Machine provides lowest accuracy among all the classifiers just in case of diabetes disorder disease. Just in case of polygenic disease Support Vector Machine provides the bottom accuracy.

Sensitivity (TPR) and Specificity (FPR) square measure real measures that depict however well quite way between a case with positive and with negative category. Sensitivity is that the ID of unwellness rate that ought to be swollen and Specificity is that the false organized rate that's to be reduced for definite examination. Sensitivity and specificity is effortlessly found by utilizing beneath formulae.

A recognized confusion matrix (sometimes referred to as contingency table) is non-heritable to estimate four measures. Confusion matrix may be a matrix illustration of the classification results. It contains knowledge concerning real and expected classifications done by a classification framework. The cell which signifies the amount of samples orders as true whereas they were true (i.e., TP), and also the cell that indicates the amount of samples named false whereas they were quite (i.e., TN). The opposite 2 cells signify the amount of samples misclassified. above

all, the cell signifying the amount of samples named false whereas they were true (i.e., FN), and also the cell which means the amount of samples named true whereas they were false (i.e., FP). Once the confusion matrixes were designed, the exactness, recall, F-measure square measure effectively observed as:

TABLE II. CONFUSION MATRIX OF SVM algorithmic rule – diabetes disease DATASET

TPRate	Precision	Recall	F-Measure	Class
1	0.943	1	0.910	Yes
1	0.962	1	0.940	No

TABLE III. CONFUSION MATRIX OF call TREE algorithmic rule –DIABETES DATASET

TPRate	Precision	Recall	F-Measure	Class
0.870	0.852	0.870	0.821	Yes
0.830	0.971	0.830	0.869	No

TABLE IV. CONFUSION MATRIX OF ARTIFICIAL NEURAL NETWORK algorithmic rule – diabetes disease dataset

TPRate	Precision	Recall	F-Measure	Class
0.980	0.956	0.980	0.921	Yes
0.930	0.954	0.930	0.911	No

Conclusion:

In this paper, we have used a pair of separate datasets from the UCI machine learning repository for the designation of heart disease and congenital disease. From experiment, it has been found that SVM

provides highest accuracy rate of 95.556% simply just in case of heart disease and in case of congenital disease Naïve Thomas Bayes classifier provides highest accuracy of 73.588%. Fitting prognosticative models like those used as a locality of this study could also be used to form individual/clinical decision support systems, to remain up upbeat or enhance the administration of chronic diseases, as an example, heart disease and congenital disease. Recognizable proof of real hazard factors and creating decision network, and powerful management measures and upbeat programs will decrease inside the chronic ill health mortality.

References:

1. Chee Peng Lim, A hybrid intelligent system for medical data classification Manjeevan Seera.
2. Dilip Kumar Choubey, Sanchita Paul, GA_MLP NN: A Hybrid Intelligent System for Diabetes Disease Diagnosis, MECS Publisher, January 2016.
3. Pasi Luukka, "Feature selection using fuzzy entropy measures with similarity classifier", Expert Systems with Applications, vol. 38. 4, pp. 4600-4607, 2011.
4. Gaganjot Kaur, Amit Chhabra, "Article: Improved J48 Classification Algorithm for the Prediction of Diabetes", International Journal of Computer Applications, vol. 98, no. 22, pp. 13-17, July 2014.
5. G. Valentini, M. Muselli, F. Ruffino, "Cancer recognition with bagged ensembles of support vector machines",

- Neurocomputing, vol. 56, pp. 461-466, 2004.
6. M. Karabatak, M. C. Ince, "An expert system for detection of breast cancer based on association rules and neural network", Expert Systems with Applications, vol. 36, pp. 3465-3469, 2009.
 7. D. Giveki, H. Salimi, G. Bahmanyar, Y. Khademian, Automatic detection of diabetes diagnosis using feature weighted support vector machines based on mutual information and modified cuckoo search. arXiv preprint arXiv: 1201. 2173, 2012.
 8. H. -J. Xing, M. -H. Ha, B. -G. Hu, D. -Z. Tian, "Linear feature-weighted support vector machine", Fuzzy Information and Engineering, vol. 1, pp. 289-305, 2009.
 9. Ravi S. Behra, Pranitha Pulumati, Ankur Agarwal, Ritesh Jain, M.D Vinaya Rao, "Predictive Modeling for Wellness and Chronic Conditions", 2014 IEEE 14th International Conference on Bioinformatics and Bioengineering.
 10. X. Xie, J. Guo, H. Zhang, T. Jiang, R. Bie and Y. Sun, "Neural-network based structural health monitoring with wireless sensor networks," in Natural Computation (ICNC), 2013 Ninth International Conference, Shenyang, 2013.
 11. P. Ghavami and K. Kapur, "Artificial neural network-enabled prognostics for patient health management," in Prognostics and Health Management (PHM), Denver, 2012.
 12. A. A. Bakar, Z. Kefli et al., "Predictive Models For Dengue Outbreak Using Multiple Rulebase Classifiers", Electrical Engineering and Informatics (ICEEI) International Conference on, pp. 1-6, 17-19 July 2011.
 13. MyDiabetesMyWay: An Evolving National Data Driven Diabetes Self-Management Platform. Deborah J. Wake, MBchB, PhD1, Jinzhang He1, Anna Maria Czesak1, Fezan Mughal1, and Scott G. Cunningham, PhD1. Journal of Diabetes Science and Technology 2016, Vol. 10(5) 1050 –1058
 14. Survey on clinical prediction models for diabetes prediction
 - N. Jayanthi1,3* , B. Vijaya Babu1 and N. Sambasiva Rao2 Jayanthi et al. J Big Data (2017) 4:26 DOI 10.1186/s40537-017-0082-7
 15. Gopalakrishna Palam, "The Practice of predictive analytics in healthcare", September 2015.

