International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

A Feature Selection Prediction Technique for Healthcare using Naive Bayes Algorithm

J.Betty Jane, E.N.Ganesh

Abstract: ***Nowadays, the data volume and its types and formats of data are very vast and complex in the field of health care. Bigdata is referred to as a huge volume of data that are complex to be handled with traditional database. The bigdata in healthcare plays an important role in enhancing treatments and facilities, hence the healthcare departments are in need to understand as much as they can, about the patient to prevent them from serious illness in the future. A feature selection technique is used for selecting subsets from large datasets and naive Bayes algorithm is used for classifying the datasets. The aim of the proposed work is to provide right information and accurate data to the organization, so that the data provided after predicting will enable the organization to ensure treating the patient's illness which may occur in the future by the help of the results found by Feature selection classification technique. Then the classification is done through naïve bayes algorithm (NB) that analyse data and gives the prediction accuracy of the future outcome healthcare datasets.***

Index Terms: Big data, Naïve bayes theorem, datasets, feature selection, wrapper method

I. INTRODUCTION

BigData is clearly centered on vast amount of data which enhance the accuracy and effects of the application for specified architectures [1]. The growth of datasets is rapid in part because of many of the internet of thingsmobile devices, and many other multimedia applications. Recent technological advancements in typical domains (e.g. Web, financial departments, health care, etc.) have directed to flood of data from these domains. Data collecting and pre-processing is difficult in recent times [2]. Most of the health care providers face some harder situation to handle data for them. To provide this information in planned and structured format, there is the essentiality of some algorithmic approach that can identify this structured format and accept the validated data. Bigdata in healthcare will have the data containing the records of the patients, reports, database of the patient etc. [3].Healthcare big data contains a large volume of datasets that are accessible to healthcare providers. As a result of healthcare data digitization and the emergence of valuable care, the healthcare department has taken make strategicdecisions using the advantage of bigdata analytics.

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
J.Bettyjane*, Department of computer science and Engineering,velsUiversity/, Chennai, India,
DR.E.N.Ganesh,Dean,School of Engineering,Vels University, Chennai, India,

The big data analytics application in healthcare gives lots of improvements and enhancements in life saving process of the patients Big data termed as the large amount of information that are analysed by the digitization of specific technologies.With respect to healthcare, it will use certain health care data of a population and prevent them from diseases, costs etc. Health care data is emerging with the patients records every day. Due to the emergence of electronic medical records (EMR), production of the data stored has risen fast to a point, so that we have a huge amount of data, called a flood of data. Doctors are in need to understand as much as about a patient as early as possible in their life, so that they can treat the patient's illness before they emerge as a serious illness, and if it is treated earlier the cost will also be reduced.

A.*Bigdata in Healthcare*

The healthcare dashboard gives you the requirements needed for the overview about the patient records to the healthcare manager. From a particular point all the data are gathered from the parameters, hence it will be helpful to find theprécised representation of your facility, will greatly help in smooth run .The mostly used parameters concerning various aspects: the number of patients utilizing the facility,the time span they stayed andwhere, treatment costs and the emergency rooms waiting time. Such a pattern view helps top-management identify potential narrowness, spot and patterns all over time, and general assessingof the situation. This gives the key in response to give better decisions,that will enhance the overall performance, with the goal of treating better for the patients and give right staff for assessing the patients. While the approach of bigdata in healthcare is still developing, it is well known that presenting of a healthcare platform for a better enhancement will not only support in the healthcare delivery but will also support in removing the boundary of space and time between the patients and medical departments. There is a feature selection technique which gets the relevant data from the irrelevant data of large datasets.[4]

B.*Healthcare Data sets*

Health care includes a huge set of public and private data systems, including health feedback, enrolment in the administrationand records billing, and healthcare data, used by various entities inclinics, CHCs, doctors, and health plan. Healthcare data include huge and large amount of healthcare data, various calculations, data for various classifications that are scattered and gathered from various healthcare data. Due to the variety of healthcare data sources data standardization is astrong pillar for perfect and relevant use of the information and combination of healthcare qualifiers, care providing departments, insurances, and government departments.

Table1:DATASETS FOR DIABETES

Sugar level	Bp	Age
200	103	85
170	139	73
117	139	73
110	118	57
112	103	34
170	100	71

II. RELATED WORK:

Feature selection, acquires for the most representative selections from the data obtained, which is critical for health analysis of the data. Feature selection gives the hidden data, meaning that the selected features provide feature extraction such as autoencoder based methods and PCA, feature selection is widely used in clinical healthcare databases rather than feature extraction.[5] The clinical databases consist of a large number of disease markers which may occur frequently. For these medical data analyses, some disease markers are not relevant and mostly gives some negative results. In order, to remove those unwanted features there is a need to apply feature selection. It also gives the Medical Decision support system enhancements by usefully reducingtime for learning the system. In this paper theycalculate three different feature selection methods, such as Principle Component Analysis,Analysis factor, and Ranking of the attribute method. Finally, using Naïve Bayesian (NB) classifier and K-nearest neighbour (KNN) classifier, the effective performance of PCA was calculated through a set of samples on a dataset. Data selection methods are classified in two ways: wrapper and filter [6].This paper study from the 2007 Behavioural Risk Factor Surveillance use twenty-three variables and 67,636 records. For the comparative analysis, feature extraction methods are of three types Chi-Square, Gain Ratio, and Info Gain, were used to select a set of useful features, which were then classified to the classification models, which helps to predict using AdaBoost, Random Forest, Radial Basis Function (RBF), Logistic Regression, and Naïve Bayes, for healthcare coverage. The most important factors were presented as a model are presented in this paper is researchers are extracting relevant datasets from large datasets.[7] In this paper, mutualinformation-based feature selection (MIBFS) method known as SURI, which increase features with high useful relevant information. A comparison is done between SURI to existing MIBFS methods using three classifiers with six healthcare data sets. In the final that is been produced, it is found in the interpretation that, SURI gives higher performance by selecting relevant features.[8] In this paper, they investigate the number of total laboratories in predicting patient deterioration in the Intensive Care Unit, investigate the effect of the average laboratory test value in which we consider laboratory tests as features. [9]

III. PROBLEM STATEMENT

From the clinical reports, clinical notes, and from body sensors, a huge volume of multi-dimensioned patient data is generated in the healthcare data. The healthcareanalysis of parameters and prediction of the subsequent health conditions for future are still in the stage of getting information. The useful way to analyze the structured and unstructured data generated from management of healthcare is the cloud cluster-enabled big data analytic platform. In this paper, adata collection probabilistic mechanism is given and the performance for collected data is done through correlation analysis. Finally, in order to predict the health condition for future a prediction model is designed on their health current status. Map reduce technique is used for bigdata text classification [10]. This study gives a new logical framework for Big representation of data, analytics of high throughput (variable selection and noise reduction), and free modelinference. Specifically, we find the core protocols of free distribution anddiagnostic model methods forinference based scientific Data sets. Compressive Big Data analytics (CBDA) iteratively produces random samples from a huge and complex dataset. Any diseases diagnosis using intelligent system produces high accuracy and treated as classification problem, also classification of health care data using machine learning techniques may be used as intelligent health care system (IHSM)SVM can also be used for classification of machine learning systems [11]. Diagnosis of diabetes is a major issue and commonly found in human being due to changing life style and also need to be identified. After applying feature selection technique (FST), a decision tree-based technique: CART is producing high accuracy with four features followed by other classification techniques. Restricted Boltzmann machines are used for segregating data from auxiliary data [12,13].The disease diagnosis at an early stage is critical. This paper deals with the prediction diagnosis of the chronic diseases using feature selection and classification techniques.Accuracy of classification systems plays a significant role in the satisfiable selection of features. In the machine learning algorithm, dimensionality reduction plays an important role. In the proposed work, the aim is to provide right information and accurate data to the organization, so that the data provided after predicting will enable the organization to ensure treating the patient's illness which may occur in the future by the help of the results found by Feature selection classification technique. With the given datasets we are applying the feature selection algorithm selecting the subsets. Then the classification is done through naïve bayes algorithm (NB) that analyse data and gives the prediction accuracy of the future outcomes' healthcare datasets

IV.METRICS USED

Metrics are computed which are:

Accuracy – This shows how often the classifier was correct. You can get this by adding the TP and the TN then dividing it with the total observations.

i)Acc=(TruePositive+TrueNegative)/(True Positive+TrueNegative+FalseNegative+FalsePositive)
ii)Precision/Positive Predicted Value (PPV) – This shows how often the model was correct when it predicted that a person is delinquent.
Prec= TruePositive/(TruePositive+FalsePositive)

iii)Recall/True Positive Rate/Sensitivity – This shows how often the model predicted that a person is delinquent when they are actually delinquent.
Rec=TruePositive/(TruePositive+FalseNegative)

iv)TNR/SPECIFICITY – This shows how often the model predicted that a person is not delinquent when they are actually not.

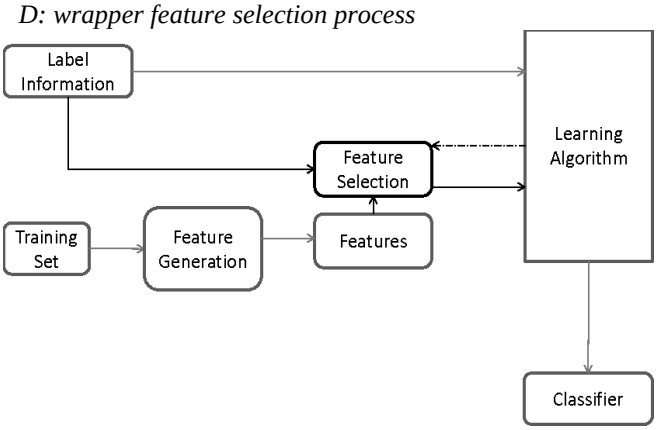
TNR = TN/(TN+FP)

V.SYSTEM OVERVIEW

C. feature selection:

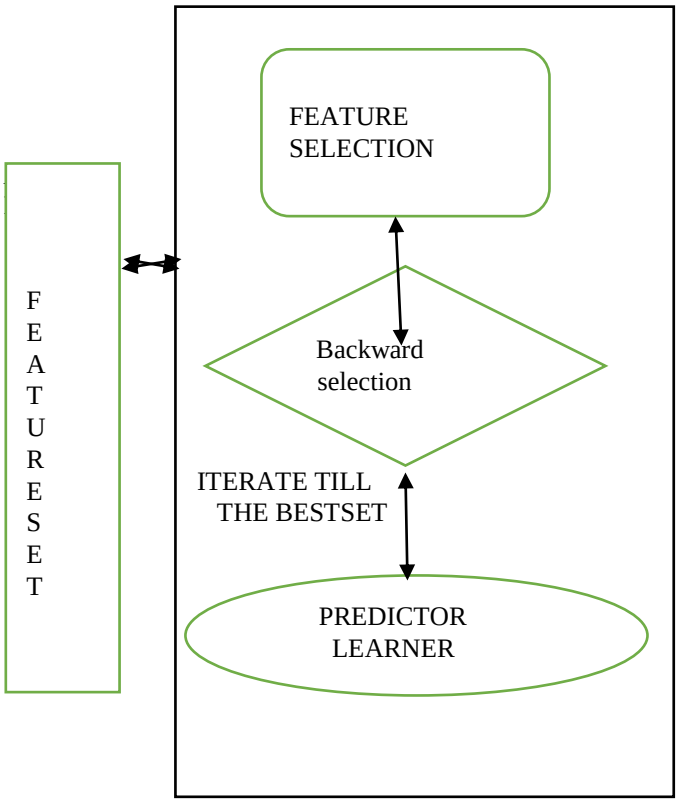
Feature selection is a process used to reduce the features before performing classification. unwanted features may produce unwanted effects while performing prediction tasks. The classification algorithm may suffer from the computational complexity ofthe adequate dimensionality often known asthe curse of dimensionality. Anover fitting is likely to occur, when a data set has too many unwanted variables and a few examples., the data are best characterized using as few variables as possible from an engineering point of view.

D: wrapper feature selection process



fig(a). FEATURE SELECTION PROCESS

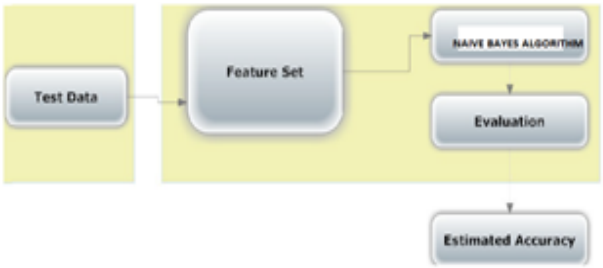
E. wrapper method feature selection search:
The Wrapper Methodology was made famous by researchers Ron Kohavi and George H. John in the year 1997. This method uses the interest of learning machine as a black box for apredictive power to score subsets of variables accordingly. The subsetfeature selection algorithm acts as a wrapper around the classification algorithm. One of the main drawbacks of this technique is the mass of computations required to obtain the feature subset.
1.select subset of features
2.give it to predictive learner a classifier
3.Performance of the classifier is influencing the feature selection process
4.Guidance for the feature selection process
After many iterations we can find the best one



Fig(b): ITERATING THE BEST SUBSET FOR FEATURE SELECTION

In the backward selection algorithm, it deletes the irrelevant data from the all feature sets.

F. classification

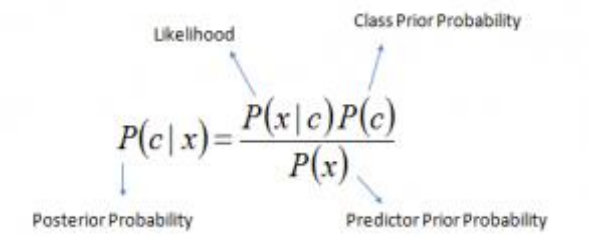


fig(c): NAÏVE BAYES CLASSIFICATION

A Naive Bayes Classifier is a supervised learning classification algorithm that assumes the Bayes' Theorem, and the features are independent computationally. Naive Bayes is a partof machine learning classification which uses the theorem of Bayes. It predicts membership probabilities for each class such as thegiven record probabilitiesor to a particular classdata point. The class with most likely class is called as the highest probability class.Naive Bayes model are used in huge data sets and are built. Naive Bayes is known to produce higher performancerather than otherhigher classification methods.

Bayes theorem are used for calculating P(c), P(x) and P(x|c) from posterior probability P(c|x). Look at the below equation:
Above,

- probability of the posterior of *class* $P(c|x)$ is the (c, *target*) given *predictor* (x, *attributes*).
- prior probability of *class is denoted* $P(c)$.
- probability likelihoodof*predictor* given class *is denoted as* $P(x|c)$.
  prior probability of *predictor is P(X)*



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

P(c)= testing dataset
P(x)= sample dataset
Naive Bayes algorithm:
Step 1. Create a frequency table from the featured dataset.
Step 2: Find the probabilities and probability of age, bp,sugar level to Create Likelihood table by finding.
Step 3: Theprobability of posterior class for each class use Naive Bayesian equation for calculation. The highest class of probability called posterior probability is the outcome of prediction.

G. feature evaluation:

True Positive Rate (TPR), False Positive Rate (FPR) are the measurements involved in calculating is performance, and only Accuracy.The proportion of the total number of predictions that are relevant is known as accuracy. TPR is denoted rightly classified instances proportion out of total classified. FPR is the shows the that were incorrectly classified as positive proportion of negative cases., first confusion matrix for the data set is then computed, to perform these metrics using these values.
Accuracy – This shows how often the classifier was correct. You can get this by adding the TP and the TN then dividing it with the total observations.
Acc=(TruePositive+TrueNegative)/(True Positive+TrueNegative+FalseNegative+FalsePositive)
Precision/Positive Predicted Value(PPV) – This shows how often the model was correct when it predicted that a person is delinquent.
Prec=True Positive/(TruePositive+FalsePositive)
Recall/True Positive Rate/Sensitivity – This shows how often the model predicted that a person is delinquent when they are actually delinquent.
Rec=TruePositive/(TruePositive+FalseNegative)
TNR/SPECIFICITY – This shows how often the model predicted that a person is not delinquent when they are actually not.
TNR = TN/(TN+FP)
G. use *a subset evaluator:*

1.From your feature vector it will create all possible subsets.

2. From the features in each subset, it will use a classification algorithm to induce classifiers.

3.It will take thefeaturesubset with which the classification algorithm gives the best performance.

VI.EVALUATION RESULTS:

The diabetes featured datasets are taken and from this datasets we are in need to predict the future criticality of the patients' health for the patients who are all between the age-68, bp-140.

Table 2: FEATURED DIABETES DATASET

Sugar level	bp	Age	Is diabetic/no
200	103	85	Yes
170	139	73	Yes
117	139	73	No
110	118	57	No
112	103	34	No
170	100	71	Yes
202	120	81	Yes
200	161	90	Yes
160	143	81	Yes
100	105	86	No
109	134	61	No
119	112	63	No
150	120	86	Yes
106	134	61	No
155	112	63	Yes
210	120	86	Yes
99	110	29	No

150	176	71	Yes
155	122	63	Yes
240	169	90	yes
117	127	47	no
160	133	65	yes

Frequency table			**Age**	
AGE	YES	NO	P(YES)	P(NO)
20-40	0	2	0/13	2/9
40-60	0	2	0/13	2/9
60-80	6	4	6/13	4/9
80-100	7	1	7/13	1/9

Frequency table			**Sugar Level**	
SUGAR LEVEL	YES	NO	P(YES)	P(NO)
50-100	0	1	0/13	1/9
100-150	2	8	02/13	8/9
150-200	8	0	08/13	0/9
200-250	3	0	03/13	0/9

Frequency table			**Bp Level**	
BP LEVEL	YES	NO	P(YES)	P(NO)
100-120	5	4	5/13	4/9
120-140	2	4	02/13	4/9
140-160	1	0	1/13	0/9
160-180	3	0	03/13	0/9

P(X|diabetic=yes) p(diabetes=yes)

P(sugarlevel=150-200||diabetes=yes)=>8/13
P(bp=140-160|diabetes=yes)=>1/13
P(age=68(60-80)|diabetes=yes)=>6/13
Total p(diabetes=yes)=>13/22

P(X|diabetic=no) p(diabetes=no)

P(sugarlevel=150-200||diabetes=no)=>0/9
P(bp=140-160|diabetes=no)=>0/9
P(age=68(60-80)|diabetes=no)=>4/9
Total p(diabetes=no)=>9/22
Let X={Age,bp,sugar level}
P(X|diabetic=yes) p(diabetes=yes)
8/13*1/13*6/13=0.0215
P(X|diabetic=no) p(diabetes=no)

0/9*0/9*4/9=0

Where 0.0215>0.so,the prediction of diabetes for the

Patient given age and bp will have diabetes in the range of 0.0215 as the future prediction.

VII. FEATUREEVALUATION&ESTIMATION ACCURACY

Considering the range fromsugarlevel is 170,bp-140, age-68.

It is been predicted that the patients who are all considered under the above conditions will have diabetes in the range of 0.0215 which is slightly greater than 0 [0.0215>0] and need to be given priority for future treatments.The estimated accuracy is calculated as 0.0215 which gives the prediction that the patients between the range of 60-80 age level

will have highest risk factor compared to that of other age group

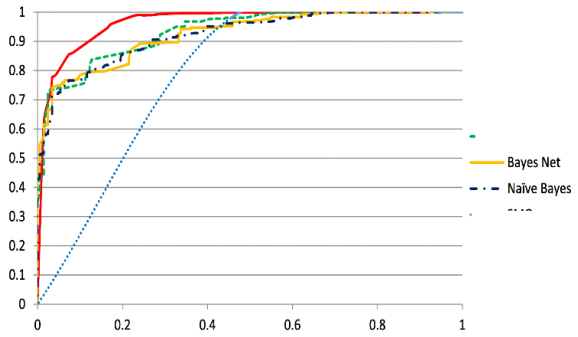


Fig:(d):NAÏVE BAYES ACCURACY

VIII.CONCLUSION

The bigdata in healthcare plays an important role in enhancing treatments and facilities, and since the datasets are large,the healthcare departments are in need to understand as much as they can, about the patient to prevent them from serious illness in the future. In this paper we are predictingthe complications of diabetes at the early stage through feature analysisby enhancing the classification techniques**.** The main aim is toprevent and cure diabetes patients and to raise the lives of the people affected by diabetes. Our proposed work does features analysis in the diabetes dataset and selects theuseful features based on the correlation values and performs classification. The naïve bayes classificationgives the good fit to the dataset notonly concerning the diabetic and non-diabetic patients but also would be easy and reliable to be enhanced with all healthcare diseases.

REFERENCES:

1. Laurent Thiry, Heng Zhao, Michel Hassenforder,IRIMASUniversit´e de Haute Alsace,"Categorical models for BigData", 2018 IEEE International Congress on Big Data
2. Fuad Rahman, PhD.MarvinSlepian, Ph.D,"A Novel Big-Data Processing Framwork for Healthcare Applications"2016 IEEE International Conference on Big Data (Big Data)978-1-4673-9005-7/16/$31.00 ©2016 IEEE 3548.
3. Weider D. Yu, Jaspal Singh Gill, MaulinDalal,Piyush Jha, Sajan Shah.
4. "BIG DATA APPROACH IN HEALTHCARE USED FOR INTELLIGENTDESIGN- Software As A Service"2016 IEEE International Conference on Big Data (Big Data)978-1-4673-9005-7/16/$31.00 ©2016 IEEE 3443.

5. S.Visalakshi; V.Radha.,"A literature review of feature selectiontechniques and applications: Review of feature selection in data mining",2014IEEE International Conference on Computational Intelligence and Computing Research.
6. Rahul Samant,SVKM'S NMIMS, Shirpur Campus, India;SrikanthaRao,TIMSCDR, Mumbai University, "A study on Feature Selection Methods in Medical Decision Support Systems". International Journal of Engineering Research & Technology Vol. 2 Issue 11, November - 2013 IJERT ISSN: 2278-0181.
7. Asha Gowda Karegowda1 , A. S. Manjunath2 & M.A.Jayaram3."COMPARATIVE STUDY OF ATTRIBUTE SELECTION USING GAIN RATIO AND CORRELATION BASED FEATURE SELECTION"International Journal of Information Technology and Knowledge Management July-December 2010, Volume 2, No. 2, pp. 271-277
8. Shruti Gupta ; Sandeep Kumar ; Akanksha Garg ; Dharmendra Singh ; N S Rajput ,"Class wise optimal feature selection for land cover classification using SAR" 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS).
9. Data Shiyu Liu and Mehul Motani, ,"Feature Selection Based on Unique Relevant Information for Health Data", Machine Learning for Health (ML4H) Workshop at NeurIPS 2018.
10. NouraAlNuaimi, Mohammad M Masud and Farhan Mohammed. "EXAMINING THE EFFECT OF FEATURE SELECTION ON IMPROVING PATIENT DETERIORATION PREDICTION"International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.6, November 2015 DOI : 10.5121/ijdkp.2015.5602 13.
11. Joan Santoso ;EkoMulyantoYuniarno ; MochamadHariadi,"Large Scale Text Classification Using Map Reduce and Naive Bayes Algorithm for Domain Specified Ontology Building",2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics
12. Manus Ross ; Corey A. Graves ; John W. Campbell ; Jung H. Kim "Using Support Vector Machines to Classify Student Attentiveness for the Development of Personalized Learning Systems",2013 12th International Conference on Machine Learning and ApplicationsYear: 2013 , Volume: 1
13. [12].Jian Zhang ,"Deep Transfer Learning via Restricted Boltzmann Machine for Document Classification",2011 10th International Conference on Machine Learning and Applications and WorkshopsYear: 2011 , Volume:1.