ISSN (e): 2250 – 3005 || Volume, 08 || Issue, 6|| Jun – 2018 ||
International Journal of Computational Engineering Research (IJCER)

Development of a Data Mining Based Model for Classification of Child Immunization Data

Sourabh Shastri[1*], Paramjit Kour[2], Ankush Gupta[3], Shakshi Sambyal[4], Arun Singh Bhadwal[5], Amardeep Sharma[6], Professor Vibhakar Mansotra[7], Dr. Anand Sharma[8]

[1,2,3,4,5,6] *Department of Computer Science and IT, Kathua Campus, University of Jammu, Kathua, India*
[7] *Department of Computer Science and IT, University of Jammu, Jammu, India*
[8] *Department of IT & Research, UCCA, Guru Kashi University, Talwandi Sabo, Punjab*
**Corresponding Author: Sourabh Shastri*

ABSTRACT
In present times, healthcare domain has vast amount of data from prescriptions, treatment costs and outcomes, diagnostic tests, patient care records, insurance claims, immunization laws, available vaccines and many other fields because data gets collected on daily basis. The resultant healthcare data has certain characteristics that make its analysis very challenging and attractive. In this paper, we have developed a data mining based model for the classification of child immunization data. As immunization is an important part of healthcare, an attempt has been made to classify the child immunization data of Jammu and Kashmir state by applying Naïve Bayes classification approach of data mining. The data is classified into two class labels viz. Priority Districts and Non-Priority Districts. A tool using Java NetBeans has been developed to carry out the classification.
KEYWORDS: Healthcare, Databases, Data Mining, Classification, Immunization, Naïve Bayes, Net Beans.

Date of Submission: 06-06-2018 Date of Acceptance: 21-06-2018

I. INTRODUCTION

India is the second largest populated country in the world where 30% of its population is living below the poverty line and healthcare is a major issue of worry in India. People living in cities and towns acquire adequate health facilities but people of remote and rural areas are deprived of these health facilities and their living environment is not so healthy as compared to the people of cities, thereby encountering various health issues as diverse from malaria to uncontrolled diabetes, malnutrition, cancer and many other diseases. No doubt, rural areas have the facility of public healthcare centers but due to low quality of care, unavailability of skilled doctors and poor facilities majority of people transit towards private healthcare centers. On the other hand, private healthcare centers provide good quality and care to their patients but all families of rural areas cannot afford to visit private facilities. To overcome all these problems, Government of India has already taken number of initiatives but still lot of innovative ideas and decision making is required to solve these problems in the first instance and child immunization is one such initiative.

Immunization is a process for developing healthy environment, a necessary requirement in the lives of children to protect them against various infectious diseases and one of the most cardinal health interventions. It is a method of making individual's immune system healthy by protection against various infectious diseases by making use of various vaccines like tetanus, whooping cough, measles, polio etc. Much infectious, serious or life-threatening diseases can be cured within few weeks after the birth of the child by administering these vaccines. There are various reasons viz. inadequate delivery of health services, lack of awareness of people regarding vaccination, huge population, inaccessibility of remote areas and many more owing to which India lags behind in comparison to many developed countries in the world and to make India usher in the new era of global health, innovative thinking is required so that its future can be strengthened and many organizations have already started working in the field of healthcare in general and immunization in particular but still a lot of work is required to improve the performance globally.

Many missions and awareness camps in this field were started by the Government of India but the existing problems in the healthcare scenario pitches for the need to change the existing structure of the present health care services by applying data analytics i.e. Data Mining [1]**.** For making this possible, the data of remote or rural areas should be collected and on that basis, the decision pertaining to the area requiring immunization on prime basis should be taken by discovering knowledge which is useful for decision making. In this research paper, we have attempted to find out the districts of Jammu and Kashmir State where more attention is required in terms of child immunization by developing a data mining based model for child immunization data using Naïve Bayes classification. Besides model building, various model evaluation measures are used to assess the accuracy of the model.

II. LITERATURE REVIEW

By applying various data mining techniques, the healthcare system gets improved. We have explored various research papers on Naïve Bayes algorithm data mining technique and other related fields. Ankita Manwatkar and Prof. R. B. Mapari [2] applied this algorithm for analyzing the twitter data of engineering students and focus on their issues of learning experience. Priyanka Sao and Pro. Kare Prashanthi [3] pointed out that Naïve Bayes algorithm is mostly used algorithm in e-spam for spam classification. They have used LingSpam dataset for classification of spam and non-spam mails with various extraction techniques. The work of Preety and Sunny Dahiya [4] revealed that the effect and rapid growth of sentiments, evaluations, attitudes and emotions coincide with the field of social media on the Web. They worked on sentiment analysis system by using modified k-means and Naïve Bayes algorithms. Girija D. K, Dr. M. Giri and Dr. M. S. Shashidhara [5] identified the variations and the danger factors related to the female internal reproductive organ fibroids by using orange canvas tool and find significant hidden patterns that provide decision to the women's health. S. L. Ting, W. H. Ip and Albert H.C. Tsang [6] compared the performance of Naïve Bayes with other classifiers including decision tree, neural network and support vector machines in terms of accuracy and computational efficiency. Cfs Subset Evaluator and Rank Search with Gain ratio metric techniques were adopted for the feature selection. Priya. S and Dr. Antony Selvadoss Thanamani [7] examined the comparative study of Naïve Bayes classifier and K-Nearest Neighbor for handling the missing values. S. Vijayarani and M. Muthulakshmi [8] analyzed the best classification algorithm among Bayesian classifier including Bayes Net and Naïve Bayes and Lazy classifiers including IBL, IBK and Kstar by applying various performance factors. Sayali D. Jadhav and H. P. Channe [9] pointed out the advantages and disadvantages of various classification techniques such as K-Nearest Neighbor classifier, Naive Bayes and Decision Trees. Lina L. Dhande and Dr. Prof. Girish K. Patnaik [10] implemented the Naive Bayes and Neural Network classifier on sentiment classification for classifying the movie review into positive or negative polarities. They combined Naive Bayes classifier with Neural Network with the help of which the accuracy of sentiment analysis got increased upto 80.65%. Bekir Karlik and Emre Öztoprak [11] applied Naive Bayes classifier for analyzing the application of pharmacogenetics to personalized cancer treatment using data of TPMT polymorphisms. Sourabh Shastri, Anand Sharma and Vibhakar Mansotra [12] classified the child immunization data of J&K state using TAN structure of Bayesian Network into priority and non-priority districts.

III. DATA MINING

Data Mining is a powerful tool and has potential to influence public health in many ways from personalized, genetic medicine to studies of environmental health and many applications in between [13]. Besides, it provides various techniques and algorithms for analyzing the historical data to discover trends and patterns or extracting knowledge from such huge data. Data mining is a necessary step in the process of Knowledge Discovery from Data (KDD) and the process involved number of steps including Data Cleaning to remove noisy and inconsistent data, Data Integration to combine multiple heterogeneous or homogeneous data sources, Data Selection to consider only data relevant to the task, Data Transformation to transform data into forms appropriate for mining functions such as aggregation or summarization, Data Mining to apply the algorithms and techniques, Interpretation & Evaluation to identify interesting patterns based on interestingness measures and Knowledge Presentation where visualization or representation techniques are used [14]. The step Interpretation and Evaluation is important for analyzing the results in terms of calculating correctness and restore accuracy etc. [15].

On the basis of the type of data to be mined, two categories of functions are involved in data mining that are Descriptive and Predictive Mining. The descriptive data mining is used to discover pattern based on real data and identify the relationship between attributes. It is a type of unsupervised learning. Descriptive data mining includes Clustering, Association Rules, Sequential Analysis etc. where clustering is a process of partitioning the data set into sub-groups based on their similarities [16]. These groups are called clusters. Various types of clustering methods are available including Partitioning methods, Hierarchical methods, Density based methods etc. Association rule is the technique in the field of data mining which aims at extracting interesting correlation

and frequent patterns among item sets in the transaction database or other data repositories [17] and it can be used to improve decision making in wide variety of applications such as medical diagnosis, relational or distributed databases, banking, department stores etc. Sequential analysis is the method of finding interesting or sequential patterns among the large databases.

On the other hand, the predictive data mining uses historical data for prediction. The predictive model is the supervised learning method because the class label is given and on the basis of which model is constructed. Various techniques like Classification, Time Series Analysis, Regression etc. comes under predictive mining. Regression analysis is used to fit an equation to a dataset or to find the relationship between independent variables and dependent variables [18].

IV. CLASSIFICATION

Classification is the most commonly used technique for accurately predicting class label of target data [19] with implementation by Naïve Bayesian classifier, K-nearest neighbor classifier, C4.5, Support Vector Machine etc. Classification is a model finding process with two steps. The data is partitioned into two subsets: training data and testing data. The first step is learning process and in this step, the training data is analyzed on the basis of which model is constructed. The second step is classification process where the accuracy of the model is examined using test data. If the accuracy is acceptable, then the obtained model would be used to classify the new data. The partitioning of new data is according to the class label.

It is pertinent to mention that in a given set of objects, each object belongs to some class and thus the aim of classification technique is to construct a rule which will allow assigning future objects to a particular class. These kind of problems come under supervised classification and many methods for constructing such rules have been developed [20]. In this research work, we have developed a tool in Java NetBeans and used the Naïve Bayes classifier algorithm to develop a data mining based model on child immunization data.

V. NAÏVE BAYES CLASSIFIER

Naïve Bayes is one of the efficient learning algorithm in data mining and machine learning. Naïve Bayes is a simplest probabilistic classifier which makes use of Bayes theorem that works on conditional probability or is a way of calculating posterior probability. This classification is named after Reverend Thomas Bayes who introduced Bayes theorem and this was firstly applied to text classification by Mosteller and Wallace [21]. 'Naive' simply means that all the features that make up a document are independent of each other. Bayes theorem is a mathematical formula that performs calculation on the basis of conditional probability i.e. it uses knowledge from prior events to predict future events and combinations of values in the historical data. Conditional probability of an event is the probability of an event occurring given that another event has already occurred.

According to the definition of conditional probability:

P (B/A) = P (A and B)/P (A) where B represents dependent event and A represents the prior event.

It builds the model based on training data, lookup for new data to which class label it belongs to by calculating probability. Let D be a training set of tuples and each tuple is represented by an n-dimensional attribute vector, **X =** (x1, x2, . . . , xn) and their associated class labels are of m classes, C1, C2, . . . , Cm. Bayes rule can be stated as:

$$P(Ci/X) = \frac{P(X/Ci)\,P(Ci)}{P(X)}$$

$$\text{Posterior} = \frac{\text{Likelihood x Prior}}{\text{Evidence}}$$

Where P(Ci/X) is the posterior probability of class given predictor with which classification can be done in an optimal way for variety of effectiveness measures.

P(Ci) is the prior probability of class (Ci), the Naive Bayesian classifier predicts that tuple **X** belongs to the class Ci.

P(X/Ci) is the likelihood or posterior probability of X conditioned on Ci.

P(X) is the prior probability of X.

To reduce computational value, class independence assumption is introduced. The conditional independence assumption that the probabilities P(X/Ci) are independent given the class c and hence can be 'naively' multiplied as follows:

$$P(X1, X2, \ldots, Xn) = P(X1/Ci)P(X2/Ci) \ldots P(Xn/Ci).$$

This can be represented as:

$$P(X1 \ldots Xn|C) = \Pi_{i=1 \text{ to } n} P(Xi/C) \text{ [22]}$$

In case of zero probability, the laplacian correction is used for smoothing the data set called as expert parameter. There is a simple way to avoid zero probability by adding one value. We assume that our training set is too large

that if one value is added on to each count, in probability the negligible amount of difference is estimated. The advantages of Naïve Bayesian classifier is that it requires small training data set, easier for implementation, fast to classify and more efficient, thereby, working well on categorical data and can be used for both binary and multiclass classification problems etc.

VI. METHODOLOGY

Data Source: The secondary data used in this paper is mainly taken from National Health Mission portal. In this paper, the classification is carried out on the basis of child immunization data of the state of Jammu and Kashmir for the year 2014-2015 by using Naïve Bayes Classifier.

Data Set Description: Here, we are predicting the probability for various districts. The data set contain the profiles of 22 Districts and 10 indicators having numeric attributes listed in the table:

S. No.	Indicator	Description	Type
1	IS_H_A	Number of immunization sessions held during the month where ASHAs were present	Range
2	IS_H_M	Number of immunization sessions held during the month	Range
3	IS_P_M	Number of immunization sessions planned to be held during the month	Range
4	C16TT16	Number of children (more than 16 years old) given TT16	Range
5	C10TT10	Number of children (more than 10 years old) given TT10	Range
6	C5DT5	Number of children (more than 5 years old) given DT5	Range
7	M16OPV	Number of infants (more than 16 months old) received OPV Booster dose	Range
8	M16DPT	Number of infants (more than 16 months old) received DPT Booster dose	Range
9	TC911F1	Total number of children (9 to 11 months old) fully immunized	Range
10	TF911F1	Total number of female children (9 to 11 months old) fully immunized	Range

Table 1: Description of the Dataset

VII. EXPERIMENTAL RESULTS

In this paper, experiment is carried out by developing a data mining tool where we have used Naïve Bayes classifier algorithm for building a classification model. For developing this tool, Java NetBeans 6.1 was used as Integrated Development Environment (IDE). The objective of this dataset was to determine the testing samples which get classified into two different classes i.e. Priority and Non-Priority Districts. The execution of process is divided into the following steps [23]:

- Data loading to read the input dataset.
- Apply Naïve Bayes classification algorithm and classify the testing data on basis of training dataset.

Data loading to read the input dataset:

In the first screen of the tool, one has to browse file option. On clicking browse button, a file chooser dialog box shall appear from which one can choose any excel file to apply classification. After choosing the file, one has to click on the ADD button.

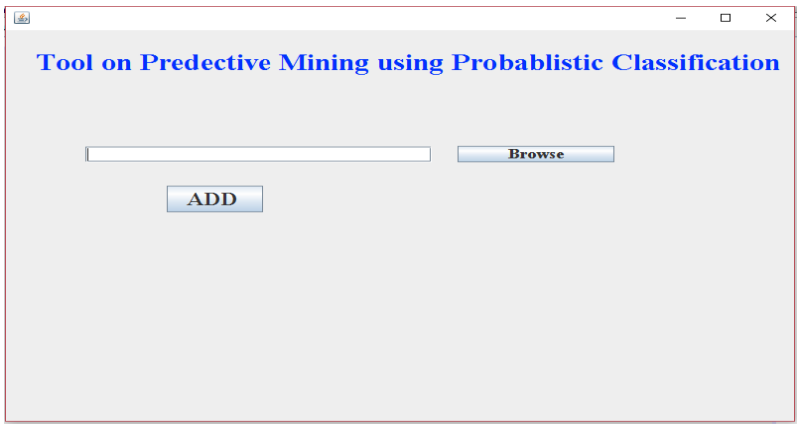

Figure 1. First Screen of the Tool

Apply Naïve Bayes classification algorithm and classify the testing data on basis of training dataset:
Input Screen for User: Labels and textboxes will be appear in this screen according to the data present in the excel file. The number of attributes present in dataset is shown and in textboxes user can enter the values of specified column. The CALCULATE PROBABILITY button is clicked after entering the data in textboxes for calculating probability of entered data and RESET VALUE button is used for clearing the entered values in textboxes.

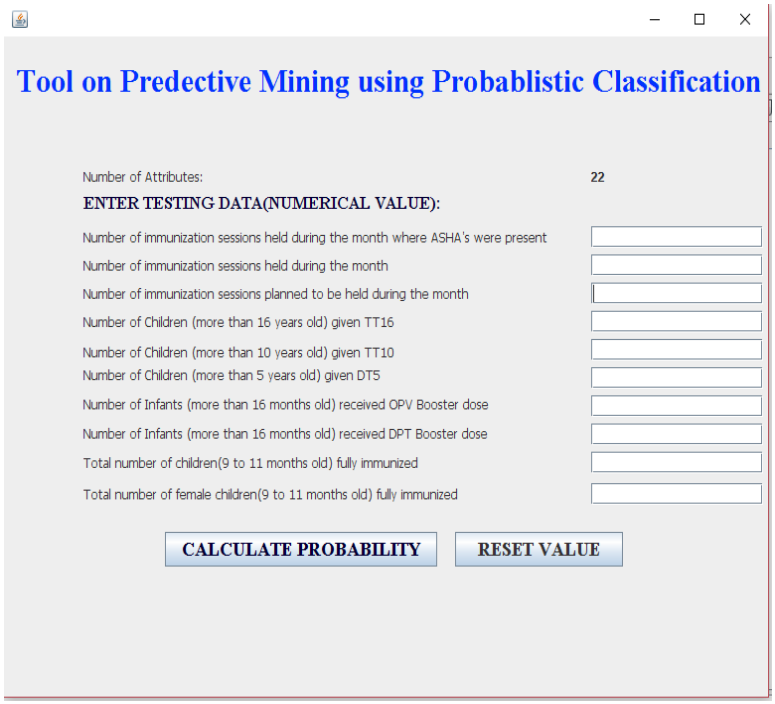

Figure 2. Input Screen for User

Output Screen: After entering the values in all textboxes, one has to click on the CALCULATE PROBABILITY button. It will calculate the probability of entered data and on the basis of which class label i.e. Priority Districts (PD) and Non-Priority Districts (NPD) is assigned to test data.

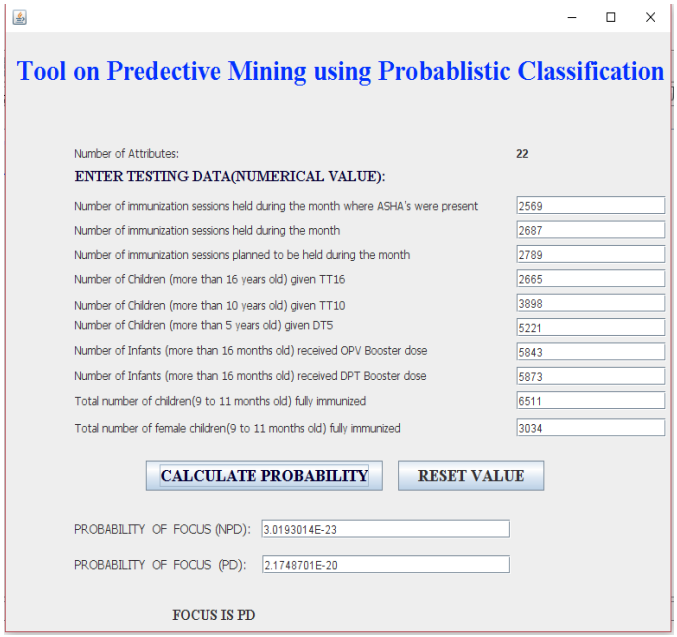

Figure 3. Output Screen of Tool

VIII. MEASURING PERFORMANCE

Classification accuracy is calculated by determining the percentage of tuples that are correctly classified. A classification model can be balanced by setting a threshold value which will operate at a desired value by using a curve. Using confusion matrix, performance parameters of a classifier can be calculated. For confusion matrix, we go through all the predictions made by the model i.e. information about actual and predicted classifications and accuracy of the solution to a classification problem. The following table 2 shows the confusion matrix for a two class classifier. The entries in the confusion matrix have the following meaning as discussed below [24]:

1. True Positives (TP): These refer to the positive tuples that were correctly labeled by the classifier i.e. number of correct predictions of positive instances.
2. True Negatives (TN): These are the negative tuples that were correctly labeled by the classifier i.e. the number of correct predictions of negative instance.
3. False positives(FP): These are the negative tuples that were incorrectly labeled as positive i.e. number of incorrect predictions of positive instances.
4. False negatives(FN): These are the incorrect predictions of positive tuples as negative instances [25].

Classes	Priority Districts	Non- Priority Districts	Total
Priority Districts	6(TP)	0(FN)	6 (P)
Non- Priority Districts	1(FP)	15(TN)	16 (N)
Total	7 (P)	15 (N)	22

Table 2: Confusion matrix

From the above confusion matrix, True Positives for 'Priority Districts' is 6 while False Negatives is 0 whereas for class 'Non-Priority Districts True Negatives is 15 and False Positives is 1. The diagonal values of matrix 6+15=21 represents the correct instances classified and other values 0+1=1 represents the incorrect instances. Using confusion matrix, performance parameters of a classifier can be calculated. The performance parameters includes: precision, recall, accuracy, and area under the curve (AUC) etc.

Accuracy: Accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier i.e. checks all the prediction is correct or not. The best accuracy is 1.0 whereas the worst is 0.0.

Thus, accuracy = (TP +TN)/(P+N)

= (6+15)/(6+16) = 0.9545

Error rate: Error rate or misclassification rate of a classifier is simply error rate =1-accuracy

=1-0.9545 = 0.0455

In this paper, the accuracy comes out to be 95% which lies near best accuracy value. As accuracy is inversely proportional to error i.e. high accuracy corresponds to low error.

Sensitivity: Sensitivity is referred to as the true positive (recognition) rate which gives the implication of the real positive that is correctly described. The best sensitivity is 1.0 whereas the worst is 0.0.

Sensitivity = TP/ (TP+FN)

=6/(6+0) = 1

Thus, Sensitivity obtained in this research work is the best sensitivity as because it comes out to be 1 which shows the proportion of positive tuples that are correctly identified.

Specificity: Specificity is known as true negative rate i.e. the proportion of negative tuples that are correctly identified. The best specificity is 1.0 whereas the worst is 0.0.

Specificity = TN/ (FP+TN)

= 15/ (1+15) = 0.9375

The specificity indicated from the model is 93% which is near the value of best specificity. This gives the implication how many true negative tuples that are correctly classified.

Precision: Precision is a measure of how many positive predictions were actually such i.e. measure of exactness. As higher value of precision means less false positives, while a lower precision value means more false positives. The best precision is 1.0 whereas the worst is 0.0.

Precision = TP/ (TP+FP)

=6/(6+1) =0.857

Thus, precision value is approximately 85% as there is less false positive values in dataset.

Recall: Recall checks all positive observation that is correct i.e. measure of completeness and is the fraction of relevant instances that are retrieved. Higher recall value means less false negatives while lower recall value means more false negatives and while if improving, recall value can often decrease precision.
Recall = TP/ (TP+FN)
= 6/(6+0)=1
Recall parameter obtained in this work is high which shows the proportion of false negative tuples that are correctly identified.

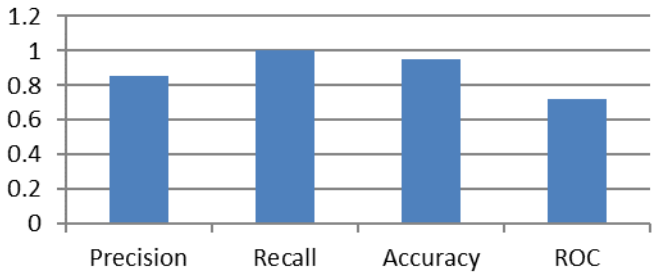

Figure 4. Graphical Representations of Performance Parameters

Curve using Naïve Bayes: AUC (Area under the Curve) of the Receiver Operating Characteristic (ROC) is a graph that applies to binary classifiers having some notion of a decision threshold. The ROC curve plots the rate of true positives versus false positives by evaluating the ranked predictions of the classifier. An ROC curve is actually two-dimensional graph in which True Positive Rate (TPR) is plotted on the Y-axis and False Positive Rate (FPR) is plotted on the X-axis [26]. The area under the curve is termed as AUC that gives the value of ROC. AUC is an evaluation of classifier as threshold varies over all possible values. The value of AUC is between 0 and 1. A rough guide for classifying the accuracy of a system is the traditional academic point system:

- 0.90 - 1 = excellent (A)
- 0.80 - 0.90 = good (B)
- 0.70 - 0.80 = fair (C)
- 0.60 - 0.70 = poor (D)
- 0.50 – 0.60 = fail (F) [27]

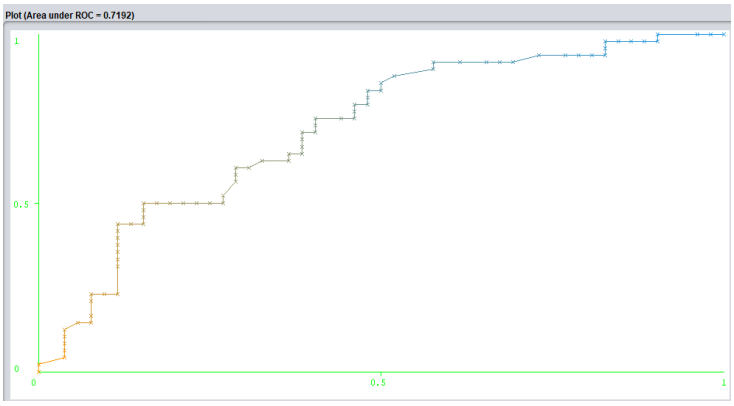

Figure 5. Area under ROC Curve

Here in this model, the Area under ROC comes 0.7192 which lies in 0.70 – 0.80 which shows result is fair.

IX. FINDINGS AND CONCLUSIONS

Classification is the very common and extensive task for information processing and it is a major data mining technique used in healthcare sectors for diagnosing and predicting diseases. In this paper, the tool has been developed to analyze the child immunization data of Jammu and Kashmir State by using the technique viz.

Naïve Bayes algorithm for classifying the dataset, as the districts are already classified into priority and non-priority districts by National Rural Health Mission, Government of India. In the light of the afore discussions and results, while using our own created model, we found certain differences in classifying the dataset. There are 22 districts in the dataset. The number of correctly classified districts are 21 and 1 district i.e. Baramula is incorrectly classified. The accuracy of the model comes out to be 95%. Now, with this model in hand, whenever a new district data shall enroll, that can easily be classified into priority and non-priority districts to achieve results.

X. FUTURE WORK

In the present study, we have developed a tool in Java NetBeans where Naïve Bayes classification approach of data mining is applied on child immunization data. In future, we shall add more algorithms and techniques in the tool to classify the child immunization data in particular and healthcare data in general.

REFERENCES

[1]. Muni Kumar N and Manjula R, "Role of Big Data Analytics in Rural Health Care -A Step Towards Svasth Bharath", International Journal of Computer Science and Information Technologies, vol. 5, no. 6, pp. 7172-7178, 2014.

[2]. Ankita Manwatkar and Prof. R. B. Mapari, "Twitter data mining using Naive Bayes Multi-label classifier", International Research Journal of Engineering and Technology, vol. 3, no. 6, pp. 2157-2160, June 2016.

[3]. Priyanka Sao and Pro. Kare Prashanthi, "E-mail Spam Classification Using Naïve Bayesian Classifier", International Journal of Advanced Research in Computer Engineering & Technology, vol. 4, no. 6, pp. 2792- 2796, June 2015.

[4]. Preety and Sunny Dahiya, "Sentiment Analysis using SVM and Naïve Bayes Algorithm", International Journal of Computer Science and Mobile Computing, vol. 4, no. 9, pp. 212-219, September 2015.

[5]. Girija D. K, Dr. M. Giri and Dr. M. S. Shashidhara, "Naive Bayesian algorithm employed in health care", International Journal of P2P Network Trends and Technology, vol. 3, no. 4, pp. 227-232, May 2013.

[6]. S. L. Ting, W. H. Ip and Albert H. C. Tsang, "Is Naïve Bayes a Good Classifier for Document Classification?", International Journal of Software Engineering and Its Applications, vol. 5, no. 3, pp. 37-46, July 2011.

[7]. Priya. S and Dr. Antony Selvadoss Thanamani, "Comparative Study of Naïve Bayes Classifier and K Nearest Neighbor in Imputation of Missing Values", in Proc. ASAT in CS'17, 2017 pp. 12-14.

[8]. S. Vijayarani and M. Muthulakshmi, "Comparative Analysis of Bayes and Lazy Classification Algorithms", International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, no. 8, pp. 3118-3124, August 2013.

[9]. Sayali D. Jadhav, H. P. Channe, "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques", International Journal of Science and Research, vol. 5, no. 1 pp. 1842-1845, January 2016.

[10]. Lina L. Dhande and Dr. Prof. Girish K. Patnaik, "Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier", International Journal of Emerging Trends & Technology in Computer Science, vol. 3, no. 4, pp. 313-320, July-August 2014.

[11]. Bekir Karlik and Emre Öztoprak, "Personalized Cancer Treatment by Using Naive Bayes Classifier", International Journal of Machine Learning and Computing, vol. 2, no. 3, pp. 339-344, June 2012.

[12]. Sourabh Shastri, Anand Sharma and Vibhakar Mansotra, "Classification of Child Immunization Data using Bayesian Network" in proc. 11th INDIACom, 2017, pp.1263-1268.

[13]. Stephanie J. Hickey, "Naive Bayes Classification of Public Health Data with Greedy Feature Selection", Communications of the IIMA, vol. 13, no. 2, pp. 87-98, 2013.

[14]. Souad Demigha, "Mining Knowledge of the Patient Record: The Bayesian Classification to Predict and Detect Anomalies in Breast Cancer", Electronic Journal of Knowledge Management, vol. 14, no. 3, pp. 128-139, 2016.

[15]. Mahesh Kini M, Saroja Devi H, Prashant G Desai and Niranjan Chiplunkar, "Text Mining Approach to Classify Technical Research Documents using Naïve Bayes", International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 7, pp. 386-391, July 2015.

[16]. Mythili S, Madhiya E, "An Analysis on Clustering Algorithms in Data Mining", International Journal of Computer Science and Mobile Computing, vol. 3, no. 1, pp. 334-340, January 2014.

[17]. K. Saravana Kumar and R. Manicka Chezian, "A Survey on Association Rule Mining using Apriori Algorithm", International Journal of Computer Applications, vol. 45, no. 5, pp. 47-50, May 2012.

[18]. Festim Halili and Avni Rustemi, "Predictive Modeling: Data Mining Regression Technique Applied in a Prototype", International Journal of Computer Science and Mobile Computing, vol. 5, no. 8, pp. 207-215, August 2016.

[19]. Heling Jiang, An Yang, Fengyun Yan and Hong Miao, "Research on Pattern Analysis and Data Classification Methodology for Data Mining and Knowledge Discovery", International Journal of Hybrid Information Technology, vol. 9, no. 3, pp. 179-188, 2016.

[20]. Raj Kumar and Dr. Rajesh Verma, "Classification Algorithms for Data Mining: A Survey", International Journal of Innovations in Engineering and Technology, vol. 1, no. 2, pp. 7-14, August 2012.

[21]. Daniel Jurafsky and James H. Martin, Speech and Language Processing, Pearson 2017.

[22]. Ishtiaq Ahmed, Donghai Guan and Tae Choong Chung, "SMS Classification Based on Naïve Bayes Classifier and Apriori Algorithm Frequent Itemset", International Journal of Machine Learning and Computing, vol. 4, no. 2, pp. 183-187, April 2014.

[23]. Maneesh Singhal, Ramashankar Sharma, "Optimization of Naïve Bayes Data Mining Classification Algorithm", International Journal For Research In Applied Science And Engineering Technology, vol. 2, no. 8, pp.145-154, August 2014.
[24]. Anshul Goyal and Rajni Mehta, "Performance Comparison of Naïve Bayes and J48 Classification Algorithms", International Journal of Applied Engineering Research, vol. 7, no. 11, 2012.
[25]. Tina R. Patil and S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", International Journal of Computer Science And Applications, vol. 6, no.2, pp. 256-261, April 2013.
[26]. Ahmad Ashari, Iman Paryudi and A Min Tjoa, "Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool", International Journal of Advanced Computer Science and Applications, vol. 4, no. 11, pp. 33-39, 2013.
[27]. J. Devi, N. Sehgal, "A Technique for Improving Software Quality using Support Vector Machine", International Journal of Computer Sciences and Engineering, vol. 5, no. 6, pp. 100-105, June 2017.

Sourabh Shastri. "Development of a Data Mining Based Model for Classification of Child Immunization Data" International Journal of Computational Engineering Research (IJCER), vol. 08, no. 06, 2018, pp. 41-49.