International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

A Model for Accurate Prediction of Child Immunization Data for Knowledge Discovery using Bayesian TAN and Naive Bayes Classifiers

Sourabh, Vibhakar Mansotra

Abstract: ***Knowledge Discovery in Databases (KDD) is a splendid methodology of discovering knowledge from gigantic databases by using its various stages viz. Data Selection, Data Preprocessing, Data Transformation, Data Mining and Interpretation/Evaluation. Data Mining is a vital sub-process of KDD methodology that is particularly used to apply the various mining algorithms on the data. In the present research paper, the authors have made an attempt to discover new knowledge by classifying the child immunization data of Jammu and Kashmir State of India. The data for the present work was collected from a web portal named as Health Management Information System (HMIS) facilitated by Ministry of Health and Family Welfare (MoHFW), Government of India. The data consists of diverse health parameters pertaining to the immunization of children and for the present study, the child immunization data of all districts of Jammu and Kashmir State was considered. Two classifiers viz. Bayesian TAN and Naïve Bayes were employed for classifying the districts of Jammu and Kashmir State into High IMR and Low IMR districts based on the available past data from 2014 to 2018. Additionally, various measurement methods have been used to evaluate the performance of the models developed by Bayesian TAN and Naïve Bayes.***

Index Terms: ***KDD, Data Mining, Classification, Bayesian TAN, Naïve Bayes, Child Immunization.***

I. INTRODUCTION

A prominent public healthcare challenge worldwide is the maintenance of health of children of a country. It is a mirror that reflects the entire spectrum of social development and reflects as a vital indicator of the well- being in a country or state. In India also, poor health among children has remained a question of worry for a long time. The necessary healthcare facilities provided to children in order to ensure their good health is central to qualitative development and a high number of newborn deaths indicate the lack of accessibility to proper medical facilities and a wide gap between the rich and poor.

Manuscript published on November 30, 2019.
* Correspondence Author
Sourabh*, Department of Computer Science & IT University of Jammu
Professor Vibhakar Mansotra, Department of Computer Science & ITUniversity of Jammu

Better provisioning of primary health services to the children can be measured by the level of health outcome such as Infant Mortality Rate (IMR).

Infant Mortality Rate (IMR) is defined as the number of deaths of infants less than one year of age per 1,000 live births [1].

It is a health outcome directly related to health indicators such as immunization of infants which forms a part of the primary health services. The information as regards child immunization is required not only to understand the status of health of the population but also to know the requisite need of the population residing in a defined geographical area. Globally about one-third or 500,000 deaths occur annually in India while most of these are vaccine preventable deaths [2].

The Universal Immunization programme was launched in the year 1985 [3] and was one of the key interventions on protection of children from preventable life-threatening conditions viz. Diphtheria, Pertussis, Tetanus, Polio, Measles, Hepatitis-B, Tuberculosis, Pneumonia and Meningitis. It became a part of Child Survival and Safe Motherhood Programme in 1992 and is currently one of the key areas under National Rural Health Mission (NRHM) since 2005 [4]. Some advances have been achieved germane to the coverage of immunization in India under this programme, but it has faced important management challenges and thus, has fallen short of the "for all children" coverage it had intended. Every year immunization saves an estimated 2–3 million lives [5]. In 2015, by focusing on 216 high focus districts across 27 states/UTs, Mission Indradhanush was envisaged to achieve 90% of full immunization coverage by 2020 [6]. Similarly, Navjat Shishu Suraksha Karyakaram (NSSK) [7], a simple and scalable training module on basic newborn care and resuscitation was developed and in order to implement this scheme, about 1.27 lakhs health care providers were trained in the field of essential newborn care and resuscitation and the same were placed at delivery points. Janani Suraksha Yojana (JSY) is the largest cash transfer program in the world for safe motherhood intervention under the National Rural Health Mission (NRHM) which was funded by the Government of India (GOI) with the objective of reducing maternal and neo-natal mortality by promoting institutional deliveries among the poor pregnant women [8].

The target of Mission Indradhanush (MI) was to fully immunize 90% of India's [9] 26 million children born each year till the age of five and to immunize all children under the age of two years besides all pregnant women against seven preventable diseases including Diphtheria, Whooping Cough, Tetanus, Polio, Tuberculosis, Measles and Hepatitis-B by 2020.

Mission Indradhanush after four phases has resulted in 6.7% increase in full immunization coverage of children [10].

It is worthwhile to mention here that the Infant Mortality Rate (IMR) of India is highest among all the BRICS countries and even the top five countries (population wise) [11]. With the major growth of data mining and applications, it is possible to collect, analyze and compare specific child immunization data for identifying previously unknown and hidden patterns, relationships and knowledge for decision making from large datasets that was not possible with traditional techniques. In this research paper, an attempt has been made to discover the knowledge from the child immunization data of Jammu and Kashmir State for the years 2014-18 with the help of the data mining based models using Bayesian TAN and Naïve Bayes classification. The two aforementioned algorithms are studied on the basis of the performance that has been evaluated by using the various measures of classification.

II. REVIEW OF LITERATURE

Some of worth mentioning works which helped in the current study are mentioned in this section. S. Shastri and V. Mansotra [12] designed a conceptual framework model KDD-MHCI based on KDD methodology having two stages viz. KDD and Decision Making. The KDD-MHCI model was proposed for maternal health and child immunization public healthcare data and they illustrated how data mining can assist decision making at different levels of health industry in India. A. K. Singha et al. [13] identified major maternal risk factors including chronic hypertension, pre-pregnancy diabetes, eclampsia and number of previous C-section deliveries that cause neonatal infant mortality. They emphasized that the performance of LR model was the best with high precision score followed by NB and LSVM. In the study carried out by Y. Fang et al. [14], CHAID was discerned to reflect the better accuracy of 88.8% than LR and identified maternal anemia during pregnancy, exclusive breastfeeding in the first 6 months, floating population and maternal educational level as risk factors for infant anemia in Beijing, China. In the research carried out by A. P. Idowu et al. [15], the ANN model displayed better results as compared to DT and Naïve Bayes algorithms for predicting the occurrence of immunizeable diseases including Yellow Fever, Measles, Polio, Hepatitis-B, Tuberculosis and Whopping Cough that affect children between age 0-5 years in the selected locations. Z. Markos et al. [16] discerned the predictive model developed using PART to be the best performer than J48 and Naïve Bayes having 92.6% of accuracy and 97.8% AUC for assessing the nutritional status of under-five children in Ethiopia. S. Shastri et al. [17] developed a tool in Java NetBeans for classifying child immunization data of Jammu and Kashmir State by applying Naïve Bayes algorithm.

D. S. Kumar et al. [18], in their study found that the classification tree performed best as compared to other algorithms including LR, NB, RF, SV, NN in the studies pertaining to the infants having low birth weight. The risk factors that were identified for low birth weight of infants were mother's last weight, mother's age, no. of premature labors, hypertension, uterine irritability and smoke. A. Alemu et al. [19] identified poor breastfeeding practices to be the reason for 24-27% of infant deaths in Ethiopia. The findings indicated that the delivery place, maternal educational status, resident place, child weight and watching television were determinate factors of child breastfeeding practice. The J48 decision tree provided high accuracy of 96.41% as compared to other algorithms including ANN, Bagging and NB. A. Singh et al. [20] discovered knowledge from the child immunization data of India by building a predictive model using Artificial Neural Network (ANN).

In the research work of S. G. Ghahfarokhi et al. [21], the RF indicated better results in terms of accuracy to predict low birth weight than LR, NB, DT and J48. The results showed that the most important factor in predicting low birth weight was gestational age less than 36 weeks. The number of fetuses, preeclampsia and premature rupture of membrane, placenta previa, the number of pregnancies and the degree of mother education were also predictors of low birth weight. S. Shastri et al. [22] classified the child immunization data of Jammu and Kashmir State for the year 2014-15 into priority and non-priority districts. All the priority districts match with the criteria of priority districts but in case of non priority districts, two out of sixteen match with the criteria of priority districts. R. Gawande et al. [23] worked on the trends in child mortality from 2000 to 2016 in India and revealed that though the child mortality has decreased over the years, still the present toll of child mortality is very high.

III. METHODOLOGY

A. Conceptual Framework

The KDD-MHCI conceptual framework proposed by [12] was used in this research paper as shown in Figure 1. The KDD-MHCI framework was divided into two phases viz. KDD and Decision Making. KDD stages are performed in multiple iterations with the emphasis on all

stages from the selection of child immunization data from public healthcare database to the knowledge discovery. The decision-making process involves collaboration of available child immunization data, knowledge base and knowledge discovery. It provides a wide spectrum of decisions for healthcare planners depending upon their roles in management levels. For this research work, child immunization data of Jammu and Kashmir State (declared now as a Union Territory by the Government of India) has been selected and all the stages of KDD and Decision Making were performed.

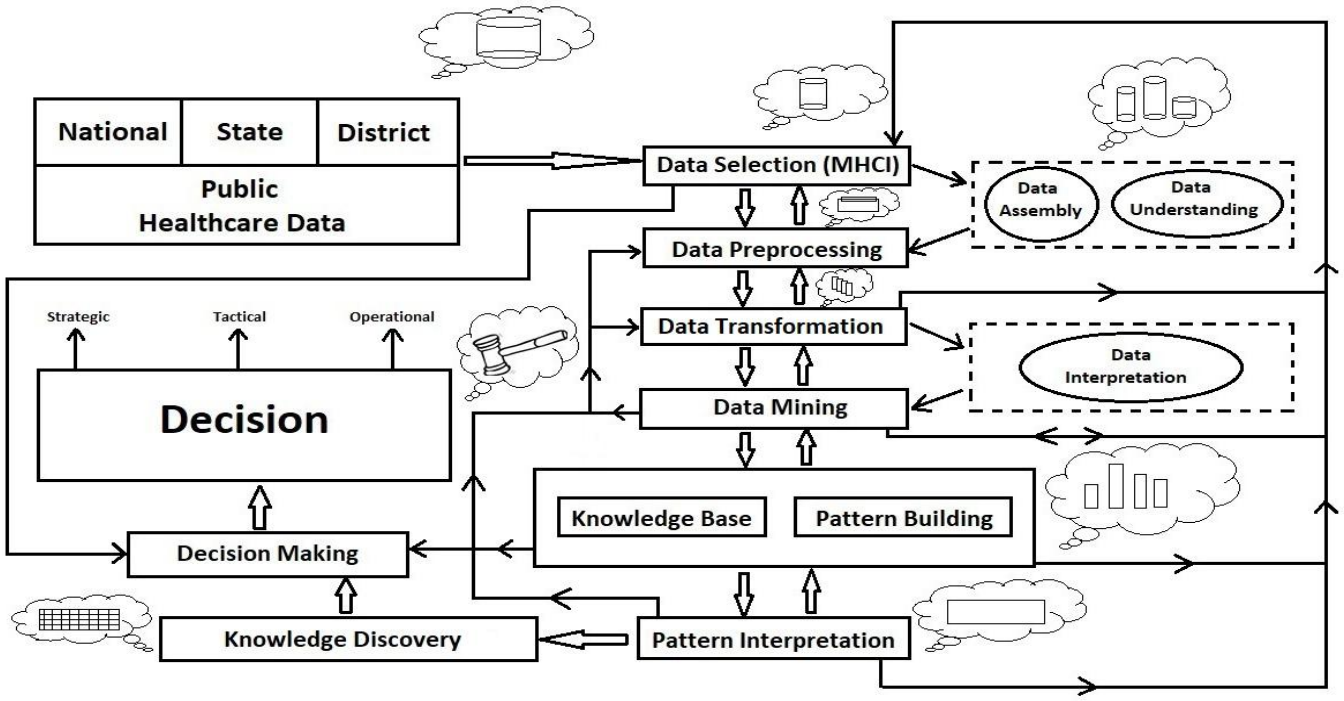

Fig. 1. KDD-MHCI Conceptual Framework [12].

B. *Proposed Algorithm*

This research introduces a new approach by studying two classifiers to discover the knowledge from child immunization databases. In this section, we first discuss the formal definition of the research problem. Let $D_S = \{F_1, F_2, \ldots\ldots F_{n-1}, F_n\}$ be a dataset with n features. The objective of the algorithm as shown in Algorithm 1 is to find out the best classifier between the two classifiers Bayesian TAN and Naïve Bayes. The Group Mean substitution method was used for the imputation of missing values and outliers have been replaced by the nearest non-outlier group values. To find the best features from the State-CI dataset, three feature selection methods were applied viz. One R Feature Evaluation, Correlation Feature Evaluation and Relief F Feature Evaluation. The top twenty aggregated ranked features F_s were taken into consideration initially for modeling. The Bayesian TAN and Naïve Bayes approaches of Bayesian Networks were applied on the selected features from the dataset and their performances were evaluated on the basis of Accuracy and AUC. The knowledge was discovered from the best classifier out of the two supra mentioned approaches viz. Bayesian TAN and Naïve Bayes.

ALGORITHM 1

INPUT: *Dataset,* $D_s = \{F_1, F_2, \ldots\ldots, F_{n-1}, F_n\}$
OUTPUT: *Best_Classifier*

// Data Preprocessing

1. Apply Imputation by Group Mean Substitution.
2. Apply Nearest Non-outlier Group Substitution.

// Feature Selection

Initially, $F_s = \{\}$

3. Apply Feature Selection Methods.
4. One R Feature Evaluation: { D_s}
5. Correlation Feature Evaluation: { D_s}
6. Relief F Feature Evaluation: { D_s}
7. Find aggregated top 20 ranked features among 3 FS methods.
8. $F_s = 1/n\sum(1R, COR, ReliefF)$
9. Output: *F_s is new Dataset for Data Partition.*

// Modeling & Prediction

A) *Apply Bayesian TAN:*

10. O_1 = Bayesian TAN (F_s)
11. $PE_{BT} = O_1$ (Performance_Evaluation)

B) Apply Naïve Bayes:

12. O_2 = Naïve Bayes (F_s)
13. $PE_{NB} = O_2$ (Performance_Evaluation)

// Comparison for Best_Classifier

14. Find the Best_Classifier (Accuracy & AUC)
15. Compare (PE_{BT} , PE_{NB})

Output: *Best_Classifier*

Algorithm 1: State-CI Algorithm

C. *State-CI Flow Diagram*

The working of the State-CI algorithm is shown in the flow diagram in Figure 2. The dataset for the present research work is the State-CI dataset of Jammu and Kashmir State during the years 2014-18.

Two simple classification learning algorithms viz. Bayesian TAN and Naïve Bayes were applied and their performance was evaluated using Accuracy and AUC to find out the best classifier for knowledge discovery.

The preliminary screening of the features was carried out using the aggregation of three feature selection methods to reduce the training and evaluation time of the models.

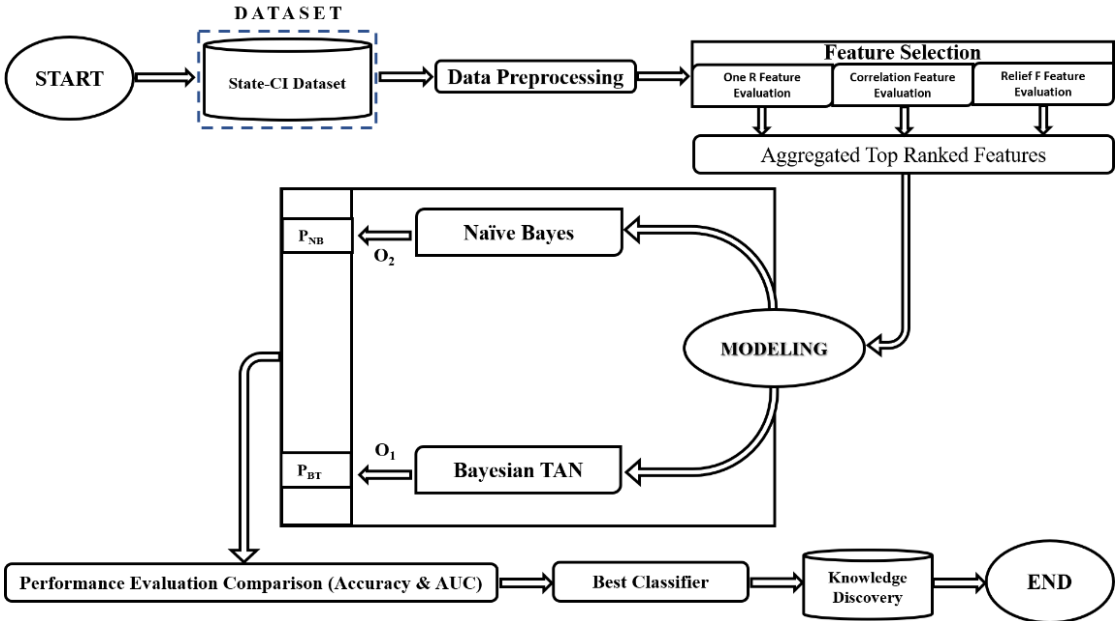

Fig. 2. State-CI Flow Diagram.

D. *Classifiers Used*

The classifiers used for the present research work included Bayesian TAN and Naïve Bayes. Both of the classifiers used in this research work belong to Bayesian networks. Bayesian networks are statistical classifiers that can predict a class membership probability such as probability that a given tuple belongs to a particular class [24]. A brief introduction about these classifiers is presented in the following subsections.

Naïve Bayes

Naïve Bayes classifier is based on Bayes theorem and can be used for the purpose of classification of the data among the class labels [24, 25]. The structure of Naïve Bayes is shown in Figure 3 where *c* is the class label and f_1, f_2, f_3, …, f_n are the various features.

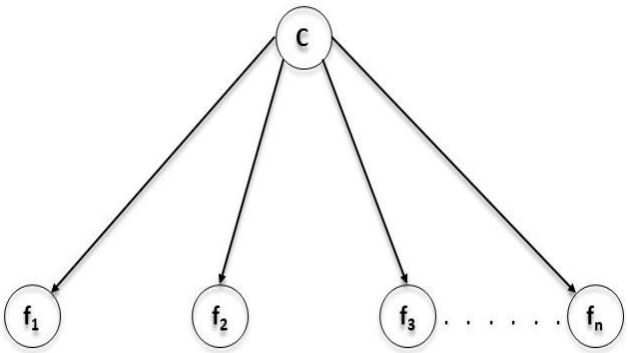

Fig. 3. A Simple Naïve Bayes Structure.

Naïve Bayes assumes that the effect of a feature value on a given class is independent of the values of the other features. This assumption is called as conditional independence and that's why it is called as Naïve. The working of the Naïve Bayes classifier is shown in Algorithm 2.

ALGORITHM 2

INPUT: testing tuple
OUTPUT: O_f = final_prediction

D_t = training set of tuples and their associated class labels (C_1, C_2)
T_t = testing tuple; r = number of rows in D_t ; c = number of columns in D_t

1. for i=1 to r do
2. count C_1; C_2
3. end for

//Calculate prior probability for each class

4. for i=1 to 2 do
5. P (C_i)= count C_i / total tuples
6. end for

// calculate condition probabilities

7. for k=1 to 2 do
8. for i=1 to c-1 do
9. for j=1 to r do
10. $C_{pk} = P(T_t/C_k)$
11. end for
12. end for
13. end for
14. if($p(T_t/ C_k) = 0$)
15. calculate posterior probability with Laplace smoothing
16. for i=1 to 2 do
17. for j=1 to c-1

18. c[j] = (C_{pk}+1) / (count C_k+((c-2)*r)*1)
19. $r_1 = r_1$*c[j]
20. end for
21. end for
22. else
23. for j=1 to 2
24. for i=1 to c-1
25. c[i] = C_{pj} / count C_j
26. end for
27. end for
28. if($c_1 > c_2$)
29. $O_f = c_1$
30. else
31. $O_f = c_2$

Algorithm 2: Naïve Bayes Algorithm

Bayesian TAN

The Tree Augmented Naïve Bayes (TAN) is from the family of Bayesian Network models that is less restrictive and thus an improvement over the standard Naïve Bayes model [26]. In TAN model, every feature is dependent on its class and one other feature from the feature set unlike the Naïve Bayes. In this way, it adds one more level of interaction among the features of the network and thus considered as more realistic than the Naïve Bayes. The TAN model gives better performance results if there are correlations among the features but the performance is almost same as that of Naïve Bayes model if there are not enough correlations among the features of the model. The structure of Bayesian TAN is shown in Figure 4 where *c* is the class label, *x* is the head feature and f_1, f_2, f_3, f_4,..., f_n are the dependent features. Here f_1, f_2, f_3, f_4,..., f_n depends on the feature x as well as on the class label c.

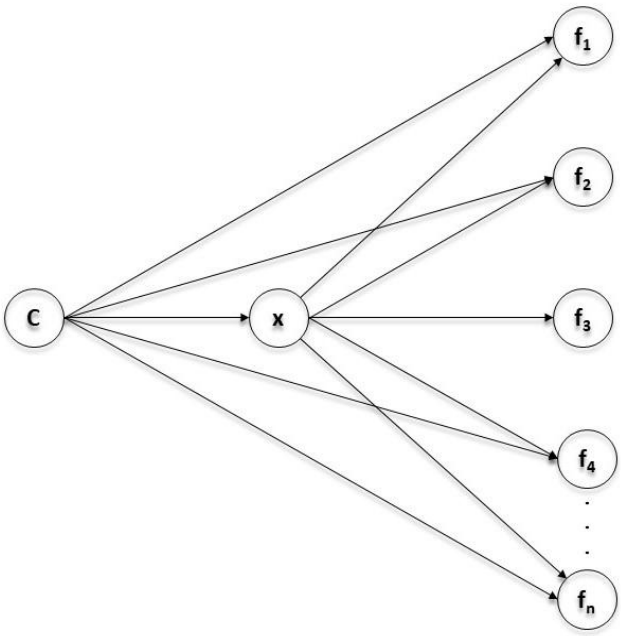

Fig. 4. A Simple TAN Structure.

IV. EXPERIMENTS

A. *Dataset*

The child immunization dataset was used in these experiments. This dataset has been collected from the Health Management Information System (HMIS) portal of Ministry of Health and Family Welfare, Government of India [27]. It initially included 39 features of all the districts of Jammu and Kashmir State and after the preliminary screening of the input features using feature selection methods, only twenty top aggregated features were used for the modeling by both of the classifiers.

B. *Feature Selection*

Feature Selection enables the learning algorithms to train faster and also reduces the complexity of a model. It increases the performance of a model if the right subset of features is selected [28]. To find the best features from the State-CI dataset, three ranking feature selection methods were applied in this research work viz. One R Feature Evaluation, Correlation Feature Evaluation and Relief F Feature Evaluation.

One R algorithm was proposed by Holte that constructs one rule for each feature in the training data and then chooses the rule with the minimum error [29]. Correlation Feature Evaluation method evaluates the worth of a feature by measuring the correlation between feature and class. It uses Pearson's Correlation Method to evaluate the correlation among each attributes and target class attribute [30]**.** Relief F feature evaluation estimates the significance of a feature by repeatedly sampling an instance and taking into consideration the value of the given feature for the nearest instance of the same and different class [29, 31]. The top twenty aggregated ranked features were taken into consideration initially for modeling as shown below in Figure 5.

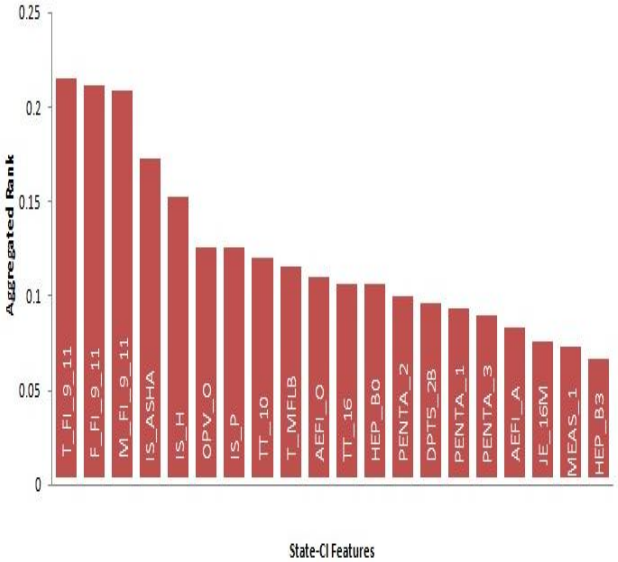

Fig. 5. Top 20 aggregated Features.

V. RESULTS AND DISCUSSION

A. *Results with Bayesian TAN*

The TAN structure for the present work is shown in Figure 6 where the features that were selected for the modeling out of 20 supra mentioned features are shown and depict a relationship with the class IMR and one other head feature i.e. T_MFLB.

There are casual relationships among input features between T_MFLB : M_FI_9_11, T_MFLB : IS_P, T_MFLB : F_FI_9_11, T_MFLB : IS_ASHA, T_MFLB : IS_H and T_MFLB : T_FI_9_11.

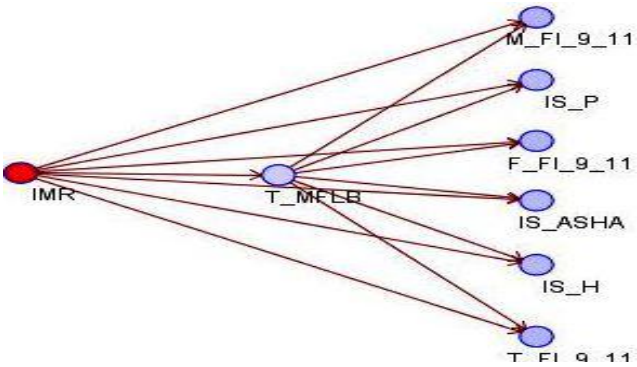

Fig. 6. Bayesian Network.

Table 1 explains the conditional probability of total number of male and female live births (T_MFLB) feature. In low IMR areas, 62% of the districts have T_MFLB value less than 8,787 whereas in high IMR areas, 67% of the districts have T_MFLB value less than 8,787 and the rest of the values of T_MFLB with regard to low and high IMR districts are shown in Table 1.

Table 1: Conditional Probabilities of T_MFLB.

T_MFLB

Parents	Probability			
IMR	<= 8,787	8,787 ~ 20,025	20,025 ~ 31,263	> 31,263
LOW	0.62	0.25	0.06	0.06
HIGH	0.67	0.33	0.00	0.00

Tables 2, 3, 4 and 5 explains the conditional probabilities of the features M_FI_9_11, IS_H, F_FI_9_11 and IS_ASHA given feature T_MFLB. These are discussed as below.

Table 2: Conditional Probabilities of M_FI_9_11 given T_MFLB.

M_FI_9_11

Parents		Probability				
IMR	T_MFLB	<= 3,002.8	3,002.8 ~ 5,380.6	5,380.6 ~ 7,758.4	7,758.4 ~ 10,136.2	> 10,136.2
LOW	<= 8,787	0.50	0.30	0.20	0.00	0.00
LOW	8,787 ~ 20,025	0.00	0.00	0.25	0.75	0.00
LOW	20,025 ~ 31,263	0.00	0.00	0.00	0.00	1.00
LOW	> 31,263	0.00	0.00	0.00	1.00	0.00
HIGH	<= 8,787	0.75	0.25	0.00	0.00	0.00
HIGH	8,787 ~ 20,025	0.00	0.50	0.50	0.00	0.00

Table 3: Conditional Probabilities of IS_H given T_MFLB.

IS_H

Parents		Probability				
IMR	T_MFLB	<= 3,398.8	3,398.8 ~ 5,484.6	5,484.6 ~ 7,570.4	7,570.4 ~ 9,656.2	> 9,656.2
LOW	<= 8,787	0.30	0.50	0.10	0.10	0.00
LOW	8,787 ~ 20,025	0.00	0.25	0.75	0.00	0.00
LOW	20,025 ~ 31,263	0.00	0.00	0.00	0.00	1.00
LOW	> 31,263	0.00	0.00	1.00	0.00	0.00
HIGH	<= 8,787	1.00	0.00	0.00	0.00	0.00
HIGH	8,787 ~ 20,025	0.50	0.50	0.00	0.00	0.00

Table 4: Conditional Probabilities of F_FI_9_11 given T_MFLB.

F_FI_9_11

Parents		Probability				
IMR	T_MFLB	<= 2,577	2,577 ~ 4,660	4,660 ~ 6,743	6,743 ~ 8,826	> 8,826
LOW	<= 8,787	0.50	0.20	0.30	0.00	0.00
LOW	8,787 ~ 20,025	0.00	0.00	0.25	0.75	0.00
LOW	20,025 ~ 31,263	0.00	0.00	0.00	0.00	1.00
LOW	> 31,263	0.00	0.00	0.00	1.00	0.00
HIGH	<= 8,787	0.75	0.25	0.00	0.00	0.00
HIGH	8,787 ~ 20,025	0.00	0.50	0.50	0.00	0.00

Table 5: Conditional Probabilities of IS_ASHA given T_MFLB.

IS_ASHA

Parents		Probability				
IMR	T_MFLB	<= 2,889.8	2,889.8 ~ 4,524.6	4,524.6 ~ 6,159.4	6,159.4 ~ 7,794.2	> 7,794.2
LOW	<= 8,787	0.20	0.30	0.40	0.00	0.10
LOW	8,787 ~ 20,025	0.00	0.00	0.75	0.25	0.00
LOW	20,025 ~ 31,263	0.00	0.00	0.00	0.00	1.00
LOW	> 31,263	0.00	1.00	0.00	0.00	0.00
HIGH	<= 8,787	1.00	0.00	0.00	0.00	0.00
HIGH	8,787 ~ 20,025	0.50	0.00	0.50	0.00	0.00

The confusion matrix for the TAN model is shown in Table 6 indicating that out of six districts named under high IMR, all of them match with the criteria of high IMR but out of sixteen low IMR districts, two districts match with the criteria of high IMR.

Table 6: Bayesian TAN Confusion Matrix.

Partition	HIGH	LOW
HIGH	06 (TP)	0 (FN)
LOW	02 (FP)	14 (TN)

The ROC curve for the model is shown in Figure 7 and the AUC and Gini values are shown in Table 7 with other measures.

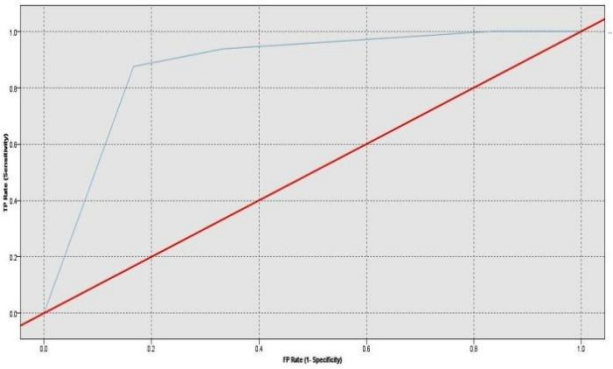

Fig. 7. Bayesian TAN ROC Curve.

Table 7: Bayesian TAN Evaluation Measures.

State-CI Evaluation	
Accuracy	90.91%
Sensitivity (TPR)	1
Specificity (TNR)	0.875
Precision	0.750
Recall	1
F-measure	0.857
AUC	0.948
Gini	0.896

B. Results with Naïve Bayes

The target field used for building the model was IMR and the input fields used were T_FI_9_11, F_FI_9_11, M_FI_9_11, IS_ ASHA, IS_H and IS_P. The testing phase of Naïve Bayes is shown in Figure 8.

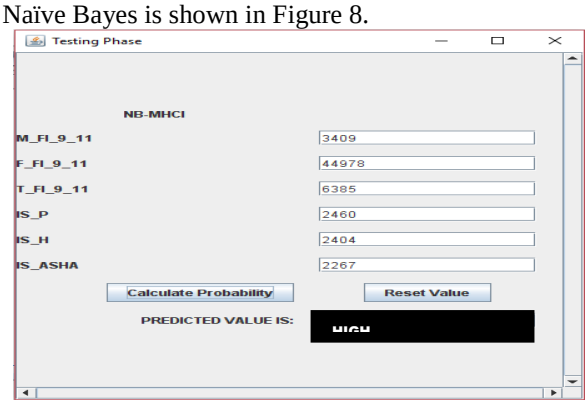

Fig. 8. Naïve Bayes Testing Phase.

The confusion matrix of Naïve Bayes is shown in Table 8 showing that out of six districts named under high IMR, one district match with the criteria of low IMR and out of sixteen low IMR districts, two districts match with the criteria of high IMR.

Table 8: Naïve Bayes Confusion Matrix.

Partition	HIGH	LOW
HIGH	05 (TP)	1 (FN)
LOW	02 (FP)	14 (TN)

The ROC curve for the model is shown in Figure 9 and the AUC and Gini values are shown in Table 9 with other measures.

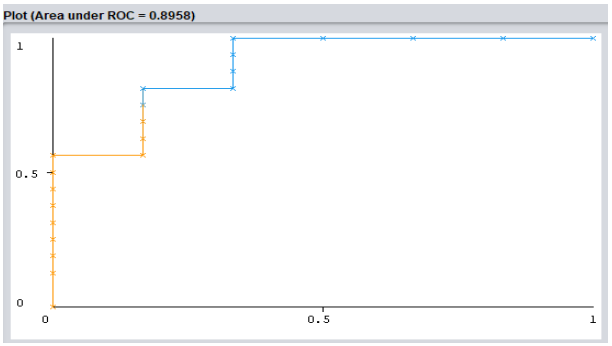

Fig. 9. Naïve Bayes ROC Curve

Table 9: Naïve Bayes Evaluation Measures.

State-CI Evaluation	
Accuracy	86.36%
Sensitivity (TPR)	0.833
Specificity (TNR)	0.875
Precision	0.714
Recall	0.883
F-measure	0.769
AUC	0.895
Gini	0.790

C. Evaluation of the results of Bayesian TAN and Naïve Bayes

The Bayesian TAN has achieved an accuracy of 90.91% that is higher than the Naïve Bayes accuracy of 86.36%. Moreover, the AUC of Bayesian TAN is also better than the AUC value of Naive Bayes. The other evaluation measures including TPR, TNR, Precision, Recall, F-measure and Gini of Bayesian TAN are also better than the Naïve Bayes as shown in Table 10. It concludes that the performance of Bayesian TAN is better than the performance of Naïve Bayes on this dataset of child immunization of Jammu and Kashmir State. Therefore, the Bayesian TAN classifier has been selected as the best classifier between Naïve Bayes and Bayesian TAN and thus considered for the knowledge discovery.

Table 10: Evaluation of the State-CI Classifiers.

Evaluation of the State-CI Classifiers		
State CI Model Measures	**Bayesian TAN**	**Naïve Bayes**
Accuracy	90.91%	86.36%
Sensitivity (TPR)	1	0.833
Specificity (TNR)	0.875	0.875
Precision	0.750	0.714
Recall	1	0.883
F-measure	0.857	0.769
AUC	0.948	0.895
Gini	0.896	0.790

D. Knowledge Discovery from State-CI Data using Bayesian TAN

- In high IMR areas, when total live births were less than 8,788 then the immunization sessions held in these districts were less than 3,399 each. In contrast, 30% low IMR districts witnessed less than 3,399 immunization sessions and 70% districts with more than 3,399 sessions.
- When total live births were less than 8,788, then the maximum immunization sessions where ASHAs were present were noted to be less than 2,890 in all districts of high IMR areas whereas in low IMR areas, 20% districts were noticed with less than 2,890 immunization sessions where ASHAs were present and 80% districts were observed to be having more than 2,890 immunization sessions.
- Further, when total live births were less than 8,788, the maximum number of male children who got fully immunized between 9-11 months were discerned to be less than 5,380 in all districts of high IMR areas whereas in case of low IMR areas, 80% districts have less than 5,380 male children who got full immunization between 9-11 months and 20% districts ranged between 5,380 to 7,758.
- In high IMR areas, when total live births were less than 8,788, the maximum number of female children fully immunized between 9-11 months were less than 4,660 in all districts whereas 70% districts have less than 4,660 female children fully immunized between 9-11 months and 30% districts were between 4,660 to 6,743 in low IMR areas.

VI. CONCLUSION

A community's future depends on the health of its citizens, particularly its children. Besides, being beneficial on an individual level, spreading awareness and improving access to immunization is crucial to economic and societal progress. In this paper, two approaches of Bayesian Networks viz. Bayesian TAN and Naïve Bayes were applied on the child immunization data of Jammu and Kashmir State of India (declared now as a Union Territory by Government of India) from 2014-18 to discover the knowledge from the data. The approaches viz. Bayesian TAN and Naïve Bayes were applied on State-CI dataset and the final results of State-CI indicated the Bayesian TAN as a better classifier than Naïve Bayes and thus the knowledge was discovered by using the TAN approach of Bayesian networks. The present study reveals that the situation in low IMR areas is better than high IMR areas and there is a dire need of conducting more awareness programmes regarding the full immunization at high IMR areas. More programmes like World Immunization Week (April 24-30) need to be consummated periodically. Various suitable and effective policies need to be put into place by the State Government to ensure full participation in immunization programmes like it should be made certain by the Governments that if parents refuse the mandatory vaccines, the main consequence should be that their children would not be accepted in schools, nurseries, etc., thereby encouraging them to understand the vitality of child immunization. In the future, we will extend this work at national level with various ensemble techniques and classification algorithms.

REFERENCES

1. Alok K. Deb et al., "A case control study investigating factors associated with high infant death in Saiha district of Mizoram, India bordering Myanmar," BMC Pediatrics, vol. 17, no. 23, 2017.
2. Mark Rohit Francis et al., "Factors associated with routine childhood vaccine uptake and reasons for non-vaccination in India: 1998–2008," Vaccine, 2017. http://dx.doi.org/10.1016/j.vaccine.2017.08.026.
3. Joe Varghese et al., "Advancing the application of systems thinking in health: understanding the growing complexity governing immunization services in Kerala, India," Health Research Policy and Systems, vol. 12, no. 47, pp. 105-116, 2014.
4. Ministry of Health & Family Welfare. [Online]. Available: https://mohfw.gov.in [Accessed 29 Jul. 2019].
5. Duncan N. Shikuku et al., "Door–to–door immunization strategy for improving access and utilization of immunization Services in Hard-to-Reach Areas: a case of Migori County, Kenya," BMC Public Health, vol. 19, 2019. https://doi.org/10.1186/s12889-019-7415-8.
6. Ajeet Singh Bhadoria et al., "National Immunization Programme – Mission Indradhanush Programme: Newer Approaches and Interventions," The Indian Journal of Pediatrics, vol. 86, no. 7, pp. 633-638, 2019. https://doi.org/10.1007/s12098-019-02880-0.
7. Shivaprasad S Goudar et al., "Institutional deliveries and perinatal and neonatal mortality in Southern and Central India," Reproductive Health, vol. 12, no. 2, 2015.
8. Kristi Sidney et al., "Out-of-pocket expenditures for childbirth in the context of the Janani Suraksha Yojana (JSY) cash transfer program to promote facility births: who pays and how much? Studies from Madhya Pradesh, India," International Journal for Equity in Health, vol. 15, no. 71, 2016. 10.1186/s12939-016-0362-4.
9. A. K. Dutta and Anju Aggarwal, "Newer Development in Immunization Practices," Indian J Pediatr, vol. 85, no. 1, pp. 44-46, 2018. https://doi.org/10.1007/s12098-017-2530-y.
10. National Health Mission. [Online]. Available: https://nhm.gov.in [Accessed 07 Aug. 2019].
11. Central Intelligence Agency. [Online]. Available: https://cia.gov [Accessed 09 Aug. 2019].
12. S. Shastri and V. Mansotra, "KDD-Based Decision Making: A Conceptual Framework Model for Maternal Health and Child Immunization Databases" in Proceedings of International Conference on Advances in Computer, Communication and Computational Sciences, 2018, pp. 243-253. https://doi.org/10.1007/978-981-13-6861-5_21.
13. Aayush Kumar Singha et al., "Application of Machine Learning in Analysis of Infant Mortality and its factors,". 10.13140/RG.2.1.3857.3687.
14. Fang Ye et al., "Chi-squared Automatic Interaction Detection Tree Analysis of Risk factors for Infant Anemia in Beijing, China," Chin Med J, vol. 129, no. 10, pp. 1193-1199, 2016.
15. A. P. Idowu et al., "Data Mining Techniques for Predicting Immunize-able Diseases: Nigeria as a Case Study," International Journal of Applied Information Systems, vol. 5, no. 7, pp. 5-15, 2013.
16. Z. Markos et al., "Predicting Under Nutrition Status of Under-Five Children Using Data Mining Techniques: the Case of 2011 Ethiopian Demographic and Health Survey," Journal of Health & Medical Informatics, vol. 5, no. 2, 2014. 10.4172/2157-7420.1000152.
17. S Shastri et al., "Development of a Data Mining Based Model for Classification of Child Immunization Data," International Journal of Computational Engineering Research, vol. 8, no. 6, pp. 41-49, 2018.
18. Senthilkumar D and Paulraj S, "Prediction of Low Birth Weight Infants and Its Risk Factors Using Data Mining Techniques" in Proceedings of the International Conference on Industrial Engineering and Operations Management Dubai, United Arab Emirates (UAE), 2015, pp. 186-194.

19. Abede Alemu et al., "Assessment of Breastfeeding practices in Ethiopia using different data mining techniques," Indian Journal of Computer Science and Engineering, vol. 7, no. 1, 2016.
20. Arun Singh et al., "Performance Evaluation of ANN Classifier for Knowledge Discovery in Child immunization Databases," International Journal of Computational Engineering Research, vol. 9, no. 3, pp. 70-76, 2019.
21. S. G. Ghahfarokhi et al., "A model to predict low birth weight infants and affecting using data mining techniques," J Bas Res Med Sci, vol. 5, no. 3, 2018.
22. S. Shastri, A. Sharma and V. Mansotra, "Classification of Child Immunization Data using Bayesian Network" in Proceedings of 4th IEEE International Conference on Computing for Sustainable Global Development (11th INDIACom), 2017, pp.1263-1268.
23. R Gawande et al., "Analysis and Prediction of Child Mortality in India," International Research Journal of Engineering and Technology, vol. 6, no. 3, pp. 5071-5074, 2019.
24. J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques. Waltham, MA: Morgan Kaufmann Publishers, 2012.
25. M. Abdar et al., "A new nested ensemble technique for automated diagnosis of breast cancer," Pattern Recognition Journal. https://doi.org/10.1016/j.patrec.2018.11.004.
26. IBM Knowledge Center. [Online]. Available: https://ibm.com [Accessed 09 Aug. 2019].
27. Health Management Information System. [Online]. Available: https://nrhm-mis.nic.in [Accessed 21 Aug. 2019].
28. Analytics Vidhya [Online]. Available: https://www.analyticsvidhya.com [Accessed 22 Aug. 2019].
29. Jasmina Novakovic, Perica Strbac and Dusan Bulatovic, "Toward Optimal Feature Selection using Ranking Methods and Classification Algorithms," Yugoslav Journal of Operations Research, vol. 21, no.1, pp. 119-135, 2011.
30. The Learntech [Online]. Available: http://learntech.uwe.ac.uk [Accessed 29 Aug. 2019].
31. R. P. L. Durgabai, "Feature Selection using ReliefF Algorithm," International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 10, pp. 8215-8218, 2014.

AUTHORS PROFILE

Sourabh, Is Pursuing Ph.D. From Department Of Computer Science & IT, University Of Jammu Under The Guidance Of Professor Vibhakar Mansotra And Currently Working As Assistant Professor In Department Of Computer Science And IT, Kathua Campus, University Of Jammu. He Is A Life Member Of Computer Society Of India And Has Published 20 Research Papers In International Journals. He Has More Than 10 Years Of Teaching Experience And His Main Research Work Focuses On Analysis And Design Of Algorithms, Data Mining And Machine Learning.

Prof. Vibhakar Mansotra, Is The Senior Most Professor In Department Of Computer Science And IT, University Of Jammu. He Has Completed His M.Tech. From IIT Delhi And Ph.D. From University Of Jammu. He Is Also Dean Mathematical Sciences And Director IT, University Of Jammu. Additionally, He Is Holding The Charge Of Coordinator IGNOU (S.C-1201), University Of Jammu. He Has More Than 27 Years Of Teaching Experience At University Of Jammu. He Has Been Conferred With The Best Teacher Award In Information Technology By The Amar Ujala B-School Excellence Awards. He Is A Life Member Of Computer Society Of India (CSI) And Presently Holding The Charge Of Chairperson Division-IV. He Has Published Several Research Papers In National And International Journals And His Research Interests Are Data Mining, Information Retrieval, Computer Graphics And Machine Learning.