IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

Classification of Maternal Healthcare Data using Naïve Bayes

[1]P. Kour, [2]S. Shastri, [3]A.S. Bhadwal, [4]S. Kumar, [5]K. Singh, [6]M. Kumari, [7]A. Sharma, [8]V. Mansotra

[1,2,3,4,5,6,8]University of Jammu, Jammu and Kashmir, India
[7]Guru Kashi University, Talwandi Sabo, Punjab, India

DOI: https://doi.org/10.26438/ijcse/v7i3.388394 | Available online at: www.ijcseonline.org

Accepted: 19/Mar/2019, Published: 31/Mar/2019

***Abstract*—** Data Mining and Machine Learning are the emerging research fields that are gaining popularity in many areas including healthcare, education, spam filtering, manufacturing, CRM, fraud detection, intrusion detection, financial banking, customer segmentation, research analysis and many others due to their infinite applications and methodologies to discover the trends and knowledge from voluminous databases in the novel manner. Healthcare industry produces gigantic amount of data related to child immunization, maternal health, family planning, clinical data, health surveys, diagnosis etc. As the process of data collection in health sector increases, the usage of data mining and machine learning techniques for analyzing and decision making also increases. There is one major health issue in health sector i.e. maternal health that needs to be worried about. In this research paper, the maternal health data of the state of Jammu and Kashmir, India has been collected from HMIS portal and Naive Bayes classification algorithm of data mining has been used for the analysis. Various performance measures including Accuracy, Precision, Recall, Kappa, F-measure, AUC and Gini have also been used for calculating the performance.

***Keywords*—**Data Mining, Machine Learning, Maternal Health, Naïve Bayes

I. INTRODUCTION

Women are the pillars of any family and society but maternal deaths during pregnancy and childbirth is a great loss to baby, family, society and country. Most maternal deaths occur due to preventable causes including bleeding after birth, unsafe abortions, obstructed labour, high blood pressure, hemorrhage, hypertensive disorders, ruptured uterus, hepatitis, anemia etc. Maternal mortality refers to women's death per 100,000 live births due to childbearing or within 42 days of termination of pregnancy and childbirth. It is one of the major health issues that need to be worried about globally. World widely 6.4 million women getting pregnant every year and on daily basis, approximately 800 women die from causes related to pregnancy and childbirth [1]. The developing countries like India having highest maternal mortality rate with 25% of world's maternal deaths get accounted in 2015 only [2]. According to the latest report of World Health Organization (WHO), India's maternal mortality rate (MMR) in 2016 was 130 deaths per 100,000 live births which comes after the mortality rate of Bhutan with 148 deaths but the mortality rate of India got declined from the figure of 215 deaths that was reported in year 2010 [3].

In India, the higher number of maternal deaths occur in rural areas as most of the population lives in rural areas. No doubt the people of remote and rural areas have the facility of public healthcare centers but low quality of care, unavailability of skilled doctors or health workers and poor facilities are the major problems. Some other factors that affect the mortality rate are tribal Gujjars and Bakerwals women of Jammu and Kashmir State who are not aware of their rights, drawbacks of early marriage, proper care and balanced diet during pregnancy. The nomadic way of life, superstitions belief, distance from health centres, lack of information or illiteracy and cultural practices are also the issues. To overcome all these problems and to meet the challenge for saving lives of mothers and new born babies, Government of India has taken many initiatives and thus, started many schemes. Indira Gandhi Matritva Sahyog Yojana is a conditional cash transfer scheme for pregnant women of age 19 years and above for the first two live births. Integrated Child Development Services is for fighting malnutrition, preschool education and healthcare to children below 6 years of age and their mothers. Janani Suraksha Yojana aims to decrease maternal deaths happening in the country by promoting institutional delivery of babies or home birth through skilled assistance and onetime cash incentive to pregnant women.

Inspite of all these schemes, the problem of maternal deaths exists in India. Nowadays, medical science is also using the recent technologies for improvements and advancements in healthcare. These recent technologies are working on the enormous amount of past and present data for various innovative outcomes. The present study is an initiative for analyzing the maternal health data of Jammu and Kashmir

State, India for discovering knowledge by applying the Naïve Bayes classifier that may assist the healthcare policy makers for decision making.

Rest of the paper is organized as follows; Section II contains related work. Various classification techniques for healthcare data have been discussed in section III. In section IV, Naïve Bayes algorithm has been explained in detail. Section V contains the details regarding dataset used for study. The experimental results and discussions are presented in section VI. Section VII contains the measuring performance whereas section VIII contains the conclusion part.

II. LITERATURE REVIEW

This section summarizes various technical articles on Maternal Health, Data Mining, Machine Learning, Classification and Naïve Bayes Classifier. W. L. Moreira et al. [4] in their study compared two Bayesian classifiers viz. Naive Bayes and AODE to better classify the hypertensive disorders severity. The Naïve Bayes classifier had an excellent performance as compared to the other experimented classifier. In another study, A. K. Singha et al. [5] presented an idea for analyzing data of mothers during their pregnancy to identify major maternal risk factors which can cause neonatal infant mortality. The analysis was based on different attributes and they build a model based on machine learning using the techniques including Naive Bayes, Linear Support Vector Machine and Logistic Regression. This model could be used to predict the key factors and high-risk mothers so that they can be given proper medical attention for reducing the infant mortality rate. In another study conducted by S. Vijayarani and S. Deepa [6] using Naïve Bayes classifier, there was prediction of hemoglobin protein based diseases. Naive Bayes has been proved to be an effective method for the disease sequence classification method and its performance can be analyzed by the factors including classification accuracy and execution time.

Another research study by S. J. Hickey [7] explored the combination of a Naive Bayes classifier with greedy feature selection that identified the most-predictive attributes from given target attribute and produced modest increases in classification accuracy. Further, C. Sundar [8] evaluated the performance of Naïve Bayes probabilistic model based classification method and find out that cardiotocography patterns are different from the suspicious and pathologic patterns with respect to precision, recall and F-score measures but not capable of distinguishing suspicious and pathologic patterns. S. Saiyed et al. [9] pointed out that Naive Bayes is one of the simple data mining techniques for predicting heart diseases based on risk factors such as age, family history, diabetes, hypertension, high cholesterol, tobacco smoking, alcohol intake, obesity or physical inactivity etc. Such risk factors based systems would not only help to reduce the mortality rate of heart disease patients in the rural areas but also give patients a warning and suggestion about the presence of heart disease even before visiting hospital or costly medical checkups. A. Kamat et al. [10] made use of two classification algorithms viz. Naive Bayes and ID3 for determining the mode of delivery based on several parameters present in obstetric ultra-sonography reports, urine test reports of pregnant women and blood. In another study, A. R. Borkar and P. R. Deshmukh [11] applied Naïve Bayes classifier technique and used it for predicting swine flu diseases.

III. CLASSIFICATION TECHNIQUES IN HEALTHCARE

Data Mining and Machine learning have been applied in various fields for improving decision making and the applications of classification technique can also be viewed in health care field to predict the class labels. Classification technique has infinite potential to utilize healthcare data more efficiently for predicting diseases and to identify inefficiencies and best practices that improve care and reduces cost. The objective of classification is to classify new data based on datasets or collected records of patients that can be used by healthcare organizations for increasing the capability of health care system [12].

Classification is a two-step process where the dataset further gets divided into two datasets viz. training and testing. The training dataset can be used for model construction and the test partition can be used for model evaluation [13]. The model construction is also known as learning or training step where the training dataset is analyzed for the construction of model. The model thus obtained can be represented in the form of classification rules, decision trees or in the form of mathematical formulas. In model evaluation, the obtained model can be examined by testing dataset in terms of performance measures like accuracy. If the accuracy of dataset is acceptable, then the developed model can be used for classifying new data. In the present research work, the classification of districts of Jammu and Kashmir State, India has been carried out into priority and non-priority districts to make awareness and necessary changes to improve healthcare system.

IV. NAIVE BAYES CLASSIFIER

Naive Bayes is a simplest probabilistic classifier but extremely fast and most effective techniques relative to other classification algorithms. It is a type of supervised machine learning technique that shows the predictable state for probability of each attribute. It is an efficient learning

algorithm which makes use of Bayes theorem by calculating posterior probability [14]. Posterior probability is a conditional probability of an event and can be defined as the probability assigned to the first event before observing the second event. Bayes theorem is a mathematical formula for calculating the probability when certain other probabilities are known. Let D be training set of tuples and each tuple is represented by an n-dimensional attribute which calculates the probability of instances d = (d1, d2, . . . , dn) being in class labels Cj for m classes representing the prior event, C1, C2, . . .Cm. This can be stated as:

$$P(Cj|d)=P(d|Cj)*P(Cj)/P(d) \quad [15]$$

Posterior = Likelihood x Prior/Evidence

Where P(Cj|d) is the posterior probability of class predictor, P(Cj) is the prior probability of class(Cj), P(d|Cj) is the likelihood of d conditioned on Cj, P(d) is the evidence of d.
During testing time, the conditional probability of some particular feature becomes zero, thus, making class-conditional probabilities equal to zero. In order to avoid the problem of zero probabilities, an additional smoothing term can be added and this value will be added to each count by which a negligible amount of difference is estimated. This variant of additive smoothing is called Laplace Smoothing or Expert Parameter.

V. METHODOLOGY

To develop the classification model, Naïve Bayes classifier has been used for predicting the class label for all 22 districts of Jammu and Kashmir State, India.

A. Data Source

The data used for model building is related to maternal health and has been taken from Health Management Information System (HMIS) portal of Ministry of Health and Family Welfare (MoHFW), Government of India for the year 2016-17.

B. Data Set Description

The dataset of maternal health initially contains 57 features and by making use of predictor importance technique of IBM SPSS Modeler, the important features have been selected for building the model. The class label used for the modeling is Focus District (FD) with two values i.e. priority district (PD) and non-priority district (NPD). As per NRHM classification, six districts including Doda, Kishtwar, Leh, Poonch, Rajouri and Ramban are classified as PD whereas sixteen districts including Anantnag, Badgam, Bandipora, Baramula, Ganderbal, Jammu, Kargil, Kathua, Kulgam, Kupwara, Pulwama, Reasi, Samba, Shopian, Srinagar and Udhampur are classified as NPD. The important features that have been identified for modeling are listed in the table 1.

Table 1. Description of the Important Features used for Analysis

S. No.	Features	Description of the Features
1	T_HD_TNSBA	Total number of deliveries conducted at home and attended by trained or non-trained SBA.
2	N_NB_24_HD	Number of newborns visited within 24 hours of delivery for deliveries conducted at home.
3	N_HD_NSBA	Number of deliveries conducted at home and attended by non trained SBA i.e. trained TBA or relatives etc.
4	N_ASHA_JSYI_I	Number of ASHAs paid JSY incentive for deliveries conducted at public institutions.
5	PNC_CS_PRF	PNC maternal complications attended.
6	W_CS_PRF	Women getting a post partum check up between 48 hours and 14 days after delivery.
7	N_CS_CHC	Number of caesarean (C-section) deliveries performed at CHCs (Community Health Centre).
8	N_MO_JSYI_PI	Number of mothers paid JSY incentive for deliveries conducted at public institutions.
9	T_PW_1TRI	Number of pregnant women who registered within first trimester.
10	PW_TT1	Number of pregnant women given TT1 during current pregnancy.
11	FD	Two values i.e. priority district (PD) and non-priority district (NPD).

VI. Experimental Results and Discussions

In the present study, the complete KDD process has been followed to make the predictions i.e. whether a district deserves to be priority or non-priority as per dataset by applying Naive Bayes classifier. The classification model application has been developed by using Java NetBeans 8.2 IDE. Knowledge Discovery from Data (KDD) is an iterative process that is used for discovering useful knowledge from datasets. The various steps [16] of KDD are:

A. Data Selection

Data Selection is the first step of KDD in which appropriate dataset is selected from collected dataset as shown in figure 1. In the present research study, maternal health dataset of Jammu and Kashmir State for the year 2016-17 has been selected.

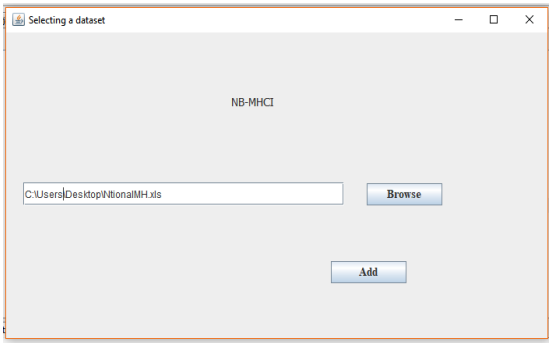

Figure 1: Data Selection

B. Data Preprocessing

Data preprocessing includes basic operations such as data cleaning, integration and reduction. Data Cleaning can be used for removing errors, ensures consistency and filling missing values. Data Integration includes the merging of data from multiple data cubes, flat files and databases. Data Reduction can be applied to obtain a reduced representation of voluminous dataset. This application has an option to fill the missing values with zero fill, one fill, central tendencies like mean, mode and median and even some constant value as shown in figure 2. In the present study, the missing values have been filled with the mean option.

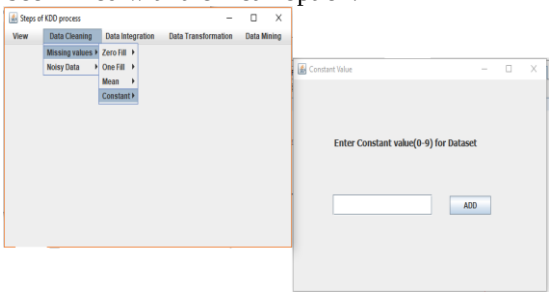

Figure 2: Data Preprocessing

C. Data Transformation

In this step, the data gets transformed into the forms appropriate for mining by using transformation techniques. Two techniques have been used in the application viz. Min-Max Normalization and Z-Score Normalization as shown in figure 3 and they can be applied as per the requirement of the application.

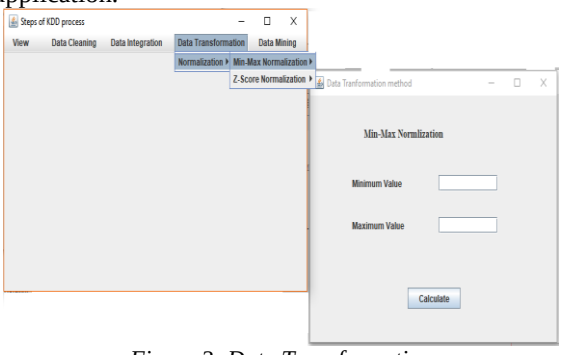

Figure 3: Data Transformation

D. Data Mining

Data Mining is the process of discovering interesting patterns from large amount of data stored in repositories, databases and data warehouse [17]. For applying Naive Bayes algorithm, an input screen shall be appeared after choosing data mining option on the application as shown in figure 4. Input screen contains the actual predictors and the corresponding textboxes for their values. The calculate probability option can be used to calculate the probability on which output of entered data depends. The class label with maximum probability shall be selected.

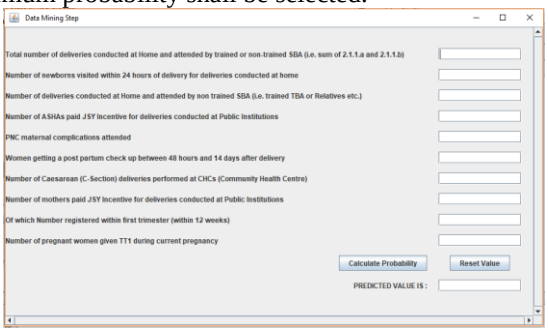

Figure 4: Input Screen of Application

The output screen as shown in figure 5 displays the predicted value as PD or NPD after calculating the probability. The class label with higher probability shall be assigned to the district for which calculations can be performed.

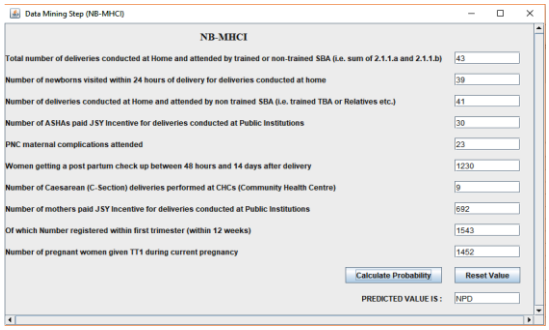

Figure 5: Output Screen of Application

E. Interpretation and Evaluation

These are the most important steps in calculating correctness and accuracy of obtained result in terms of calculating precision, recall, accuracy, F-measure etc. and visualization or representation techniques for knowledge representation.

VII. Measuring Performance

The accuracy of classification can be determined by calculating the percentage of tuples that are correctly classified. The performance of algorithm can be calculated with the aid of confusion matrix as shown in table 2.

Table 2: Confusion Matrix

Classes	PD	NPD	Total
PD	5 (TP)	01 (FN)	06 (P)
NPD	0 (FP)	16 (TN)	16 (N)
Total	5 (P')	17 (N')	22 (P+N)

The true positive value is 5, false negative value is 1 whereas true negative value is 16 and false positive value is 0. The correctly classified instances are represented by diagonal values i.e. 5+16=21 and incorrectly classified instance is only 1. Various performance parameters including accuracy, precision, recall, specificity etc. have been calculated using confusion matrix entries as given below.

A. Accuracy

It is the percentage of correctly classified instances in a test set or simply the ratio of correctly predicted observations to the total observations [18]. Accuracy is measured as: Accuracy = (TP+TN) / (P+N) = (5+16) / (6+16) = 0.9545 i.e. 95.45%.

B. Error Rate

This is simply the incorrect predictions of the dataset and the present classifier has only 4.55 error rate.

C. Specificity

The proportion of negative instances that are correctly classified as negative is known as true negative rate or specificity. Specificity = TN / (FP+TN) = 16 / (0+16) = 1 i.e. 100%.

D. Precision

It is the ratio of correctly predicted positive observations to the total number of positive predictions and also known as positive predictive value. Precision = TP / (TP+FP) = 5 / (5+0) = 1 i.e. 100% that is the best precision value and shows very less false positive values in the dataset.

E. Recall

Recall is a measure which gives the number of relevant results returned i.e. ratio of number of positive predictions to the total number of positives. Recall = TP / (TP+FN) = 5 / (5+1) = 0.833 i.e. 83% which shows less false negative tuples that are correctly identified.

F. F-Measure

It represents the harmonic mean between Precision and Recall values and takes both false positives and false negatives into account. F-Measure = 2 * (precision * recall) / (precision + recall) = 2 * (1 * 0.833) / (1 + 0.833) = 1.666 / 1.833= 0.908 i.e. 90.80.

G. Kappa

The kappa statistic is used to control only those instances that may have been correctly classified by chance [19]. This can be calculated using both the observed (total) accuracy and the random accuracy. Kappa can be calculated as:

Kappa = (total accuracy – random accuracy) / (1- random accuracy).

Here, the obtained Kappa value is 0.872 which lies between 0.0 and 1.0 and gives very good agreement.

The outcome of different performance parameters on the basis of confusion matrix entries are shown in figure 6.

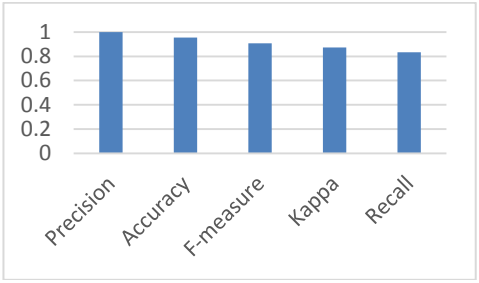

Figure 6: Performance Parameters

H. Curve using Naive Bayes

Area under the Curve (AUC) is the area enclosed by Receiver Operating Characteristic (ROC) curve which is used to determine analysis of model that predicts the classes best. ROC curve is a two-dimensional graph in which False Positive Rate is plotted against True Positive Rate. The value of AUC lies between 0 and 1 where perfect classifier is when the value of AUC is 1 i.e. 100% true positive rate and random classifier is when AUC = 0.5 i.e. 0 % false positive rate [20, 21].

Figure 7: AUC for Maternal Health Data

The value of AUC under ROC comes out to be 0.9063 which lies between 0.90 - 1.00 i.e. result is excellent.

I. Gini Coefficient

It is used to measure inequality or normalize the AUC so that perfect classifier scores 1 and random classifier scores 0. The value of gini coefficient lies between -1 and 1.

Gini Coefficient = 2*AUC-1 = 2*0.9063-1 = 0.8126.

VIII. Conclusion

Data mining and Machine Learning techniques play a vital role and has infinite potential for discovering new trends from voluminous amount of stored information which are then used to build predictive models. The present study classified the maternal health data of the districts of Jammu and Kashmir State, India for the year 2016-17. An application was developed using the process of Knowledge Discovery from Data (KDD) where Naive Bayes classifier has been used for predicting the districts. Only one district was found misclassified and the application gave an overall accuracy of 95% that shows a good accuracy rate. Additionally, precision value of 100%, recall value of 83% and F-measure value of 90.80% shows that the existing model works well for the purpose of prediction. By making use of this predictive KDD application, the districts can be easily classified into priority and non-priority districts in terms of maternal health data. The aforementioned evaluation measures for estimating the performance of the application shows good results.

ACKNOWLEDGEMENTS

The authors would like to express their sincere thanks to the Ministry of Health and Family Welfare (MoHFW), Government of India for providing important healthcare data online through HMIS portal.

REFERENCES

[1] S. Sharmilan and H. T. Chaminda, "Pregnancy Complications Diagnosis using Predictive Data Mining", In the Proceedings of the 2017 International Conference on Computational Modeling & Simulation (IC2MS), 2017.

[2] S. N. Khandale and K. Kedar, "*Analysis of maternal mortality: a retrospective study at tertiary care centre*", International Journal of Reproduction, Contraception, Obstetrics and Gynecology, Vol.6, Issue.4, pp. 1610-1613, 2017.

[3] The Hindustan Times. [Online]. Available: https://www.hindustantimes.com. [Accessed 10 Jan 2019].

[4] W. L. Moreira et al., "An Inference Mechanism using Bayes-based Classifiers in Pregnancy Care", In the Proceedings of the IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom), pp. 305-309, 2016.

[5] A. K. Singha et al., "Application of Machine Learning in Analysis of Infant Mortality and its Factors", *Working Paper*, 2016.

[6] S. Vijayarani and S. Deepa, "*Naïve Bayes Classification for Predicting Diseases in Haemoglobin Protein Sequences*", International Journal of Computational Intelligence and Informatics, Vol.3, Issue.4, pp. 278-283, 2014.

[7] S. J. Hickey, "*Naive Bayes Classification of Public Health Data with Greedy Feature Selection*", Communications of the IIMA, Vol. 13, Issue.2, pp. 87-98, 2013.

[8] C. Sundar, M. Chitradevi and G. Geetharamani, "*An Analysis on the Performance of Naive Bayes Probabilistic Model Based Classifier for Cardiotocogram Data Classification*", International Journal on Computational Sciences & Applications, Vol. 3, Issue.1, pp. 17-21, 2013.

[9] S. Saiyed et al., "*A Survey on Naive Bayes Based Prediction of Heart Disease Using Risk Factors*", International Journal of Innovative and Emerging Research in Engineering, Vol.3, Issue.2, pp. 111-115, 2016.

[10] A. Kamat, V. Oswal and M. Datar, "*Implementation of Classification Algorithms to Predict Mode of Delivery*", International Journal of Computer Science and Information Technologies, Vol.6, Issue.5, pp. 4531-4534, 2015.

[11] A. R. Borkar and P. R. Deshmukh, "*Naïve Bayes Classifier for Prediction of Swine Flu Disease*", International Journal of Advanced Research in Computer Science and Software Engineering, Vol.5, Issue.4, pp. 120-123, 2015.

[12] S. Patel and H. Patel, "*Survey of Data Mining Techniques used in Healthcare Domain*", International Journal of Information Sciences and Techniques, Vol.6, Issue.1/2, pp. 53-60, 2016.

[13] S. S. Nikam, "*A Comparative Study of Classification Techniques in Data Mining Algorithms*", Oriental Journal of Computer Science & Technology, Vol.8, Issue.1, pp. 13-19, 2015.

[14] J. Han, M. Kamber and J. Pei, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, 2012.

[15] A. Alemu, Y. Berhanu and M. Mahalkshmi, "*Assessment of Breastfeeding practices in Ethiopia using different data mining techniques*", Indian Journal of Computer Science and Engineering, vol.7, Issue.1, pp. 1-6, 2016.

[16] N. Rikhi, "*Data Mining and Knowledge Discovery in Database*", International Journal of Engineering Trends and Technology, Vol.23, Issue.2, pp. 64-70, 2015.

[17] K. K. Manjusha, K. Sankaranarayanan and P. Seena, "*Prediction of Different Dermatological Conditions Using Naïve Bayesian Classification*", International Journal of Advanced Research in Computer Science and Software Engineering, Vol.4, Issue.1, pp. 864-868, 2014.

[18] T. Shaikh and D. Deshpande, "*Feature Selection Methods in Sentiment Analysis and Sentiment Classification of Amazon Product Reviews*", International Journal of Computer Trends and Technology, Vol.36, Issue.4, pp. 225-230, 2016.

[19] B. F. Chimieski and R. D. R. Fagundes, "*Association and Classification Data Mining Algorithms Comparison over Medical Datasets*", Journal of Health Informatics, Vol.5, Issue.2, pp. 44-51, 2013.

[20] M. Vuk and T. Curk, "*ROC Curve, Lift Chart and Calibration Plot*", Metodoloski Zzvezki, Vol.3, Issue.1, pp. 89-108, 2006.

[21] 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, Issue.6, pp. 41-49, 2018.

Authors Profile

Paramjit Kour has completed her Masters in Computer Applications from Kathua Campus, University of Jammu. She has qualified National Eligibility Test (NET) in 2018. She has published 2 research papers in international journals. Her main research interests are Data Science, Artificial Intelligence and Deep Learning.

Sourabh Shastri is pursuing Ph.D. 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 17 research papers in international journals. He has more than 10 years of teaching experience and his main research work

focuses on Algorithms, Data Mining and Knowledge Discovery.

Arun Singh Bhadwal has completed his Masters in Computer Applications from University of Jammu. He has also qualified NET and GATE exams. He has 3 years of teaching experience at University level and has published 4 research papers in international journals. His main research interests are Analysis and Design of Algorithms, Compiler Design and Machine Learning.

Sachin Kumar is pursuing Masters in Computer Applications from University of Jammu. His main research interests are Analysis and Design of Algorithms, Internet of Things and Machine Learning.

Kuljeet Singh has completed his Masters in Computer Applications from University of Jammu. He has published 1 research paper in international journal. His main research interests are Time Series Analysis, Machine Learning and Image Processing.

Monika Kumari has completed her Masters in Computer Applications from University of Jammu. She has published 4 research papers in international journals. Her main research interests are Data Mining and Analysis and Design of Algorithms.

Anand Sharma is working as an Assistant Professor at UCCA, Guru Kashi University, Talwandi Sabo, Punjab. He has completed his Ph.D. in Data Mining from University of Jammu. He has published 26 research papers in international journals. His main research work focuses on Data Mining, Knowledge Discovery and Health Care Management.

Vibhakar Mansotra is the senior most Professor in Department of Computer Science and IT, University of Jammu. He is also Dean Mathematical Sciences and Director IT, University of Jammu. He has been conferred with the Best Teacher award in Information Technology by the Amar Ujala B-School Excellence Awards. His research interests are Data Mining, Information Retrieval and Computer Graphics. He is a life member of Computer Society of India.

