FORESEEING BIRTH MODE USING BAGGING ENSEMBLER CLASSIFIER

A PROJECT REPORT

Submitted by

LAVANYA A (2116210701132)

in partial fulfillment for the award of

the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

MAY 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Thesis titled **“Foreseeing birth mode using bagging ensembler classifier**” is the bonafide work of “**LAVANYA A (2116210701132)”** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr . S Senthil Pandi M.E.,Ph.D.,

PROJECT COORDINATOR

Professor

Department of Computer Science and Engineering

Rajalakshmi Engineering College

Chennai - 602 105

Submitted to Project Viva-Voce Examination held on__________________

Internal Examiner **External Examiner**

TABLE OF CONTENTS

ABSTRACT:

In many developing nations, especially in rural areas, significant delivery challenges include maternal mortality and childbirth-related complications. Identifying these risks early on can substantially lower the rates of death for expecting mothers.. Some studies have found the appliance of machine learning methods to anticipate the mode of delivery, whether natural or cesarean section. Naive Bayes, Decision Tree, K-Nearest Neighbor, and Support Vector Machine are commonly employed techniques in this regard. This paper proposes a novel approach to predicting birth modes: employing Bagging Ensemble Classifiers built upon traditional ML algorithms. The study assesses the effectiveness of four ML algorithms—DT-Bagging, KNN-Bagging, NB-Bagging, and SVM-Bagging—utilizing bagging ensemble classifiers. The findings suggest that in this area, bagging ensemble models perform better than traditional ones. The study also finds links between important variables and C-sections. Through the use of these hidden patterns in the data, our research may help design a decision support system that lowers the number of cesarean deliveries in Bangladesh.

I. INTRODUCTION

Asthma risk rises by 79% and diabetes risk rises by 20%, both of which are higher than rates seen in normal deliveries. Determining risk variables is essential to lowering the frequency of cesarean deliveries, which helps allay these worries. Pattern extraction is made easier by machine learning (ML) approaches, which provide a practical method for this task. The use of ML in healthcare is growing quickly, especially in the areas of prediction and classification. Medical data analysis is one of the many fields where automating decision-making has been the focus of numerous research studies. The purpose of this research is to identify the variables responsible for the rise in cesarean birth rates and to develop a prediction model that can differentiate between participants who had vaginal and cesarean section deliveries. The following are the study's objectives:

- Examine how well a new birth mode prediction method performs in comparison to current approaches.

- Look into the relationship between important variables and cesarean sections.

It is possible to avoid complications that harm women and newborns by identifying important characteristics linked to cesarean deliveries. By classifying the manner of delivery, the prediction model will offer crucial data to help prevent cesarean sections by being proactive.

The following is the format of the remaining portions of this paper: After discussing similar works in Section II, our experiment's is covered in Section III, and the final conclusions are said in Section IV.

This effort is finally concluded in Section V.

II. LITERATURE REVIEW

Machine learning techniques have been used in many medical investigation projects, with what comes out being used for forecasting, prediction, and comparative analysis. This section looks at certain studies that are relevant to our investigation. For example, machine learning methods were applied to foresee the modes of childbirth in a study conducted in Bangladesh [6]. However, our research covers demographic data from other regions in Bangladesh, whereas this study was limited to a specific area.

Kowser et al. analyzed 13,527 patient records with 21 variables per using a variety of machine learning classifier techniques. The above details were taken from Bangladesh's Tarail Upazilla. Nine machine learning algorithms, including impact learning and ANN, were implemented.

S. Abbas. established a connection between maternal age and cesarean section. Their findings revealed that the majority of cesarean sections occurred in age groups lesser than 20 (86%) and older than 36 (96%). Additionally, the investigation indicated that ladies who underwent vaginal deliveries exhibited lower blood pressure compared to those who underwent cesarean deliveries.

Robu and Holban performed the Naive Bayes technique to a dataset comprising 2086 patient records. Additionally, they implemented K-NN, Random Forest, AdaBoosting, SVM, KRipp, LogitBoost, REPTree, and Simple Cart.

Sunantha developed the CPD-NN algorithm, a modified form of nearest neighbor analysis that was designed for predicting the risks that accompany cesarean sections. Sunantha Sodsee used closest neighbor analysis to predict cesarean sections.

Khan Jahidur and associates employed machine learning techniques in Bangladesh to forecast childhood anemia via using results from late 2011 Bangladesh socio-Demographic and Health Review. The focus of their inquiry ended up being restricted to a subset of kids from 2013. Logistic regression, K-NN, SVM, CART, and linear discriminant analysis were used in the construction of predictive models. Moreover, they investigated a number of salient factors and their influence on children's anemia..

The study of the literature taken out in the sections before the current one highlights the incidence of cesarean section is believed to be shaped by a number of physical and social factors linked to maternal well being It is also impossible to overstate the significance of applying machine learning methodologies to many different health diagnoses and predicted birth categorization.

III. METHODOLOGY

This provides an outline of our proposed study. Beginning with dataset collection, we proceeded with feature selection and data cleaning. Subsequently, machine learning algorithms were applied to construct prediction models, and the outcomes were tested. Additionally, we found the correlation between the factors and cesarean sections.

A. Data Set

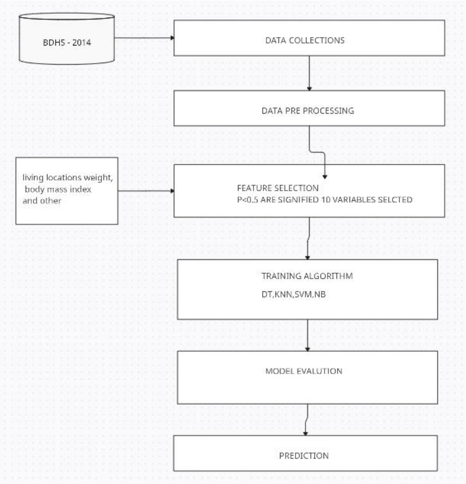
Cross-sectional data from the Bangladesh Demographic and Health Survey (BDHS-2014) dataset were utilized in this research. The BDHS-2014 survey was conducted nationally between June 28 and November 9, 2014. Employing a two-level stratified cluster sampling methodology based on household samples and enumeration areas, BDHS-2014 aimed to ensure representativeness. Additional details regarding the research design and data collection techniques can be found in the BDHS report. The study encompassed a total of 4493 sets, comprising women aged 15 to 49 who had given birth within the preceding five years.

B. Pre-Processing

After the removal of records with missing or incorrect parameter values, our final dataset comprised 4252 records. In our proposed approach, missing values were addressed through imputation techniques before model training, but despite this, the performance of the algorithms did not improve compared to simply removing missing values. Consequently, we opted to exclude missing variables from our models during the final generation phase.

Drawing primarily from prior research, we selected 15 characteristics related to delivery mode [1, 8]. These studies emphasized the significance of various elements in cesarean section deliveries. To ascertain the significance of each feature, we calculated its p-value. A character was judged important if its p-value was less than 0.05.

In our investigation, ten factors were ultimately selected based on their p-values. These factors encompassed a range of sociodemographic variables along with the child's birth weight, wealth index, bmi, age at birth, number of infants, mother's and husband's education levels, type of dwelling, living places (division), and number of previous births.



A. Methods Used for Classification

In order to categorize respondents as either having a cesarean section or giving birth normally, we implemented four known classifiers that are frequently used in this field in addition to four bagging ensemble classifiers. Decision Tree, DT-Bagging, Naive Bayes , NB-Bagging, K-Nearest Neighbor , SVM,NB-Bagging included among the classifiers.

The Decision Tree classifier generates a predictive model by analyzing features within a tree-shaped structure following logical principles. It is made up of three different sorts of nodes: decision, leaf, and root. Positioned at the top or root of the tree, decision nodes provide outputs per decision rules. The tree moves on to leaf nodes after all choices have been made.Support Vector Machine is beneficial for both regression and classification tasks. SVM divides each data element in the n-dimensional feature space and constructs a hyperplane to separate specific items into their respective classes.

Although it is non-parametric, the K-Nearest Neighbors think about does not thoroughly assume anything about the data's distribution. Rather, it uses a labeled training set of input data. When manufacturing output, the K-Nearest Neighbors algorithm Segregates each item based on the majority vote of neighbors.

The Bayes theorem, which simplifies by assuming conditional independence between each couples of attributes given the value of the class variable, forms the basis of the Naive Bayes classifier. This method hinges on the idea that an attribute's presence or absence within a set has no influence on the existence or absence of other attributes within the same dataset.

The bagging ensemble classifier seeks in boosting the accuracy and dependability of machine learning algorithm. It achieves this by aggregating classifications from randomly generated training sets to produce a final prediction. These methods are typically employed to reduce variances by randomly assigning them during the construction process and then utilizing the ensemble. The Bagging classifier has garnered significant attention due to its improved accuracy and implementation speed. By employing a smoothing technique called bagging, regression or classification trees can be made more predictive efficient. The fundamental assumption of this ensemble technique is a collection of slow learners should combine to form a fast learner. Bagging promotes numerous beneficial trees. All of these decision trees contribute to an ensemble learner, with each tree being assessed as a weak learner. A new instance is classified through analyzing it across all trees in the ensemble. Every tree provides a "vote" for a class; the class with the greatest number of votes dictates the final class prediction for the new instance.

IV. RESULTS & DISCUSSION

We used 20% of each group as the validation set and 80% of the people in each group for training in the course of the research. Table 1 shows the outcome of our analysis of eight different classification algorithms. We used a number of performance parameters in our scrutiny, including F1 Score, Accuracy, Precision, and Recall. A bigger area under the ROC curve (AUC) shows better distinction between cesarean and normal scenarios. We also used Receiver Operating Characteristic curves to display performance of the model.

The performance of the classifiers is presented in Table 1, showing that collaborative approaches such as bagging far surpass classical machine learning algorithms. Decision Tree (DT), for instance, achieves an overall accuracy of 0.86 with an F1 score of 0.61, while DT-Bagging achieves the lowest accuracy of 0.82 with an F1 score of 0.59. More specifically, KN-Bagging significantly improves performance with an accuracy of 0.86 versus 0.79 for KNN and an F1 score of 0.49 versus 0.34 for KNN. Additionally, SVM-Bagging and NB-Bagging perform noticeably better in traditional events than DT and NB.

As seen in Figure 2, SVM-Bagging outscored SVM with an AUC of 0.87 as an alternative to 0.85. While KNN showed an AUC of 0.67, KN-Bagging performed significantly more effectively, showing an AUC of 0.83. For both NB and NB-Bagging, the AUC values are somewhat close. Furthermore, DT-Bagging produced an AUC of 0.86, which was more than DT's AUC of 0.83. Therefore, it is clear that bagging ensemble approaches constantly outperform their conventional counterparts when taking into account F1 score, AUC, Accuracy, Precision, and Recall. In fact, the analysis shows that bagging classifiers perform better than classical classifiers in most cases. However, we saw some differences in the algorithms' Precision and Recall ratings. KNN (Bagging), for instance received a recall score of 0.36 but a precision score of 0.84.In the same manner, the SVM obtained a 0.84 precision and 0.36 recall score. These differences reflect that while some classifiers perform superbly in recall, they may perform bad in precision, and vice versa. The significance of taking note of several performance measures in assessing classifier performance thoroughly is highlighted by these subtleties.

TABLE 1 PERFORMANCE OF ALGORITHMS

	Correctness	Precision	Regain	F1
DT (Bagging)	**0.88**	**0.79**	**0.48**	**0.62**
KNN (Bagging)	**0.87**	**0.89**	**0.35**	**0.48**
SVM (Bagging)	0.84	0.72	**0.52**	**0.61**
NB (Bagging)	**0.84**	**0.69**	0.55	**0.64**

Furthermore, we analyzed the connection between numerous factors and cesarean sections (CS), as shown in Tables 2, 3, and 4. This analysis has major implications for more studies in this field and clarifies the possibility of using machine learning methods in predicting cesarean sections based on medical and sociodemographic factors.

As per our data, Dhaka had the second-highest percentage of cesarean deliveries (33.61%), behind the Khulna region with 35.2%. Likewise, a relationship has been established between greater maternal and paternal education levels and an increased tendency to select cesarean deliveries. Interestingly, our study showed that women in employment had babies via cesarean section at a higher rate than those in unemployment. Moreover, compared to women from low-income (5.7%) and middle-income (19.7%) backgrounds, the rates of cesarean deliveries were much higher among affluent women (53.5%).

TABLE 2 ASSOCIATION BETWEEN VARIABLES AND C- SECTION.

Variables	**teams**	**Sample**	**CS (%)**	**P- value**
Divisions	Dhaaka	741	249 (33.61%)	< 0.001
	Chitagong	814	161 (19.8%)	
	Barisal	515	104 (20.1%)	
	Khulna	509	179 (35.2%))	
	Rajshahi	519	147 (28.3%)	
	Rangpur	532	106 (19.9%)	

	Sylhet	622	84 (13.5%)	
Residence	Hi tech	1361	504 (37.3%)	< 0.01
	Rural area	2891	522 (18.5%)	

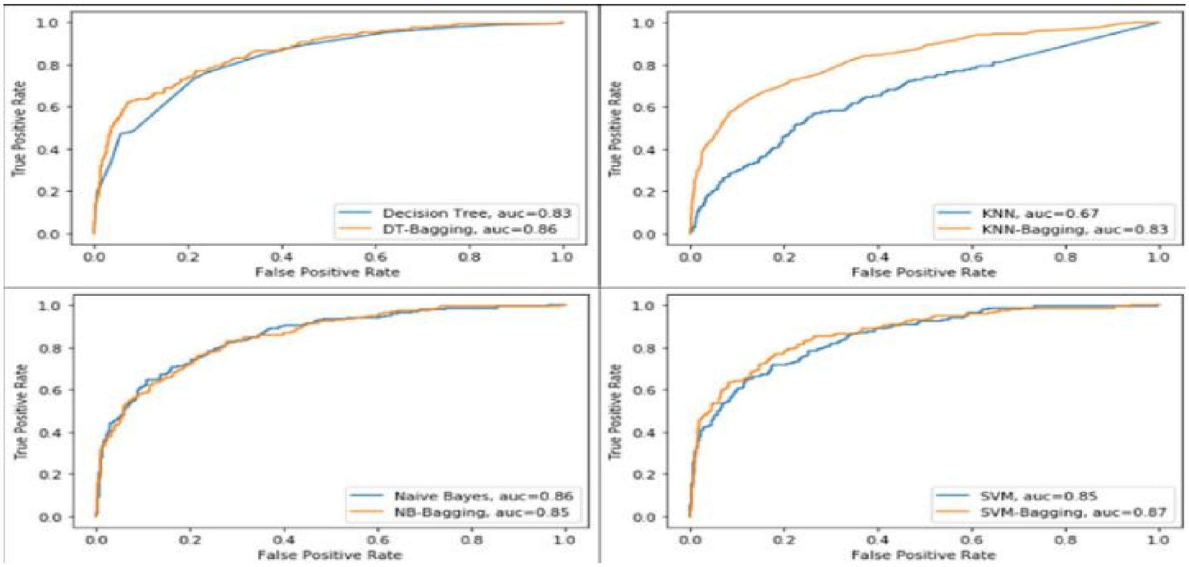


TABLE 3 : RELATION BETWEEN SOCIOECONOMIC FACTORS AND C-SECTION.

Variables	teams	Sample	CS	p-value
Working	No	3339	854 (25.7%)	0.026
	Yes	914	173 (18.9%)	
Age at first birth	≤ 20 years	3401	646 (18.9%)	0.001
	> 20 years	851	357 (41.9%)	
No. of children	1-2 children	3018	860 (28.5%)	< 0.001
	> 2 children	1234	160 (12.9%)	
Birth weight	Normal	3673	852 (23.2%)	< 0.001

	Large	579	172 (29.7%)	
BMI	Under(BMI ≤ 18.4))	1041	137(13.1%)	< 0.001
	Normal(18.4 < BMI)	2517	566 (22.5%)	
	Over (24 < BMI < 31)	595	260(43.7%)	
	Obese (BMI ≥ 30 kg)	99	57(57.6%)	

V. CONCLUSION

In conjunction with addressing problems during childbirth, improved mother and child health outcomes are gained. The significance of demographics and health-related variables in comprehending and forecasting labor outcomes is emphasized by this study. It demonstrates the reliability of bagging ensemble classifiers and pinpoints critical factors causing cesarean sections. Additionally, the new method of classifying births offers helpful details for developing preventative strategies that might decrease the rate of cesarean deliveries and improve the general health of mothers and their children. Additionally, the results of this study open the door to customized healthcare plans that address the unique requirements of expectant mothers and their babies, improving the quality of pregnancy and delivery experiences.

VI. REFERENCES

[1] M. N. Khan, M. M. Islam, A. A. Shariff, M. M. Alam, and M. M. Rahman, "Socio-demographic predictors and average annual rates of caesarean section in Bangladesh between 2004 and 2014," Plos One, vol. 12, no. 5, 2017.

[2] J. Souza, A. Gülmezoğlu, P. Lumbiganon, M. Laopaiboon, G. Carroli, B. Fawole, and P. Ruyan, "Caesarean section without medical indications is associated with an increased risk of adverse short-term maternal outcomes: the 2004-2008 WHO Global Survey on Maternal and Perinatal Health," BMC Medicine, vol. 8, no. 1, 2010.

[3] A. Sana, S. Razzaq, and J. Ferzund, "Cause Analysis Automated Diagnosis and of Cesarean Section Using Machine Learning Techniques," International Journal of Machine Learning and Computing, pp. 2012.

[4] E. Symonds and I. Symonds "Essential Obstetrics and Gynaecology," 4th ed. Oxford: Churchill Livingstone, 2003.

[5] M. J. Patwary and X. Z. Wang, "Sensitivity analysis on initial classifier accuracy in fuzziness based semi-supervised learning," Information Sciences, vol. 490, pp. 93–112, 2019.

[6] M. Kowsher, N. J. Prottasha, A. Tahabilder, and M. B. Islam, "Machine Learning Based Recommendation Systems for the Mode of Childbirth," Cyber Security and Computer Science Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pp. 295–306, 2020.

[7] S. A. Abbas, R. Riaz, S. Z. H. Kazmi, S. S. Rizvi and S. J. Kwon, "Cause Analysis of Cesarean Sections and Application of Machine Learning Methods for Classification of Birth Data," in IEEE Access, vol. 6, pp. 67555-67561, 2018.