Trees and Nearest Neighbors Based Models for Predicting Cesarean Section Delivery

Bushra
Lecturer at Department of Statistics, Abdul Wali Khan University Mardan.
bushra_bakhtiar@awkum.edu.pk

Nosheen Faiz
Assistant Professor at Department of Statistics
Abdul Wali Khan University Mardan.
nosheenfaiz@awkum.edu.pk

Almas
Assistant Professor at Department of Physics
Abdul Wali Khan University Mardan.
almas@awkum.edu.pk

Beenish Khurshid
Lecturer at Department of Biochemistry
Abdul Wali Khan University Mardan.
beenish_khurshid@awkum.edu.pk

Saira Farman
Lecturer at Department of Biochemistry
Abdul Wali Khan University Mardan.
sairafarman@awkum.edu.pk

Zahida Parveen
Associate Professor at Department of Biochemistry
Abdul Wali Khan University Mardan.
zahida@awkum.edu.pk

Date of Submission:10th December 2021 Revised: 25th January 2022 Accepted: 26th March 2022

Abstract : -Several women undergo major surgery due to complications in their pregnancy and suffer severe health issues. Therefore, knowing the reasons that why it happens, is of significant importance. This work aims to assess various risk factors causing c-section surgery using machine learning algorithms. A case study is carried out for known risk factors associated with c-sections across selected hospitals in districts Mardan and Peshawar, Khyber Pakhtunkhwa, Pakistan. The study is based on pregnant women who had registered their pregnancies in Mardan and Peshawar districts to predict c-section for future mothers. Machine learning models such as *k*-NN, Weighted *k*-NN, SVM, Random Forest, and Optimal Trees Ensemble (OTE) are trained and assessed, based on an independent set of observations and the best performing methods are identified.

Keywords C-section, trees, nearest neighbors, risk factors

1 Introduction

Cesarean section is one of the major abdominal surgery in women. When women become pregnant, she mentally prepares herself for both cases either vaginal delivery or cesarean section. Due to its importance, this is essential to identify risk factors associated to c-section. Supervised machine learning (ML) methods are commonly applied for classifying the data of pregnant women based on the way of delivery either c-section or a vaginal delivery [1, 2]. Various machine learning algorithms for example, decision tree and artificial neural networks can be used for childbirth categorization such as c-section or normal delivery, and various health check factors are recognized [3]. A case study shows that if a woman had a previous c-section might be able to deliver her second child vaginally where different ML tools are implemented to find the rate of virginal birth after c-section [4, 5]. These techniques may also be used to recognize risk factors such as, weight, blood pressure and maternal age for Preeclampsia [6]. According to [7], decision tree model can be employed to predict delivery type and the hazard factors accompanying with c-section surgery. In [8], the efficiency of ML algorithms regarding the classification of childbirth using boosting and bagging classification methods has been discussed. It is also suggested to build prognostication models on k-Nearest Neighbors (*k*-NN), stacking classification, decision tree, Support Vector Machine (SVM), and Random Forest (RF) methods for an efficient type of delivery [9]. Least Absolute Shrinkage and Selection Operator (LASSO), Classification and Regression trees, and Random Forest (FR) may possibly be applied to identify risk units, rank variables, and choose variable quantity for a model, support to an effect [10, 11]. Recent studies related to the factors affecting c-section is conducted to evaluate the implementation of ensemble machine learning methods and Deep Neural Networks on the c-sectional dataset [12, 13].

The aim of this work is to use primary data related to c-section, collected through a self-designed questionnaire. For data collection, a convenience sampling procedure has been applied. Since, the data is collected through questionnaire, and is categorical data where different questions have been asked from pregnant women. We aim to identify potential risk factors that lead to c-section for the women in Mardan and Peshawar by implementing the most appropriate up-to-date machine learning model(s) for predicting c-section, for instance, k-Nearest Neighbors (k-NN), Optimal Trees Ensemble (OTE), Random Forest (RF), Support Vector Machines (SVM), Weighted k-Nearest Neighbours. Furthermore, we have also calculated Accuracy, Sensitivity, Specificity, and Brier Score for validating these methods.

2 C-Section Rate Globally and in Pakistan

C-section rates have climbed dramatically in recent years compared to historical periods. The amount has increased between 2003 and 2018. The growth rate is 21%. The rate of C-sections in certain regions of South America is close to 60%. Its rate is 5% in Africa while it was 26% in Canada in 2005. Australia had the highest rate of caesarean sections in 2007 (31%). The World Health Organization officially renounced its former recommendation of a 15% c-section rate in June 2010. Similar to how China has 46% whereas Asia only has 25% of c-sections [3].

Like to other nations, Pakistan is experiencing a sharp rise in the number of caesarean sections. Despite the fact that having a c-section had several drawbacks, many people still preferred it. The rate of caesarean sections in Pakistan was about 3.2% in 1990, and it rose quickly to reach 19.2% in 2018. C-section rates are relatively high, with the most recent demographic survey reporting a rate of nearly 22% [9].

2.1 Target Population

The research is related to factors affecting C-Section so for this purpose our target population is all government, semi-government, and private hospitals. We have employed an elegant questionnaire method to gather the data. Gynecologists and psychologists assisted in the creation of the questionnaire to gather crucial information for the investigation's problem.

2.2 Dataset Description

A sample of 930 patients with 25 factors/variables is selected from different hospitals of Mardan and Peshawar. Data was collected for socio-demographic variables and physical characteristics including age, height, monthly income, residence, number of babies, previous c-section, gestational disease, baby weight, baby position, baby condition etc.

3 Objectives of the Study

The objectives of this study are as follows.

1. To determine potential risk factors for C-Section in Peshawar and Mardan women.
2. To identifythe most suitable current machine learning model (s) for C-Section prediction.

4 Experimental setup

Machine learning methods such as, *k*-NN, Weighted *k*-NN, Random Forest (RF), Support Vector Machines with Gaussian Kernel (SVMG), Support Vector Machines with Polynomial Kernel (SVMP), Support Vector Machines with Linear Kernel (SVML) and Optimal Trees Ensemble (OTE) are trained by using 70%, 50% and 30% training parts in three different cases, while the corresponding 30%, 50% and 70% parts are used as testing data, respectively. Experiments are repeated 500 times and the final results are averaged. The results are shown for all three cases in Tables 1, 2 and 3. Important features are selected via LASSO [14]. For comparison of the methods accuracy, sensitivity, specificity, and Brier Score are used as comparison metrics. Furthermore, for visualization of the results the boxplots are also constructed. In this study, R-software is used for data analysis. "kernlab" package [15] is implemented for Support Vector Machine. *k*-Nearest Neighbour is implemented by using R library "e1071" [16]. We have applied Optimal Trees Ensemble (OTE) by using the default values in the R package "OTE" [17]. Similarly, Random Forest is calculated via R package "Random Forest" [18].

4.1 **Results and Discussion**

A number of machine learning algorithms are used to analyze the dataset i.e., *k*-Nearest Neighbors (*k*- NN), Weighted *k*-Nearest Neighbors (W*k*-NN), Random Forest (RF), Optimal Tree Ensemble (OTE) and Support Vector Machine (SVM) with three different kernels. By using these techniques, the accuracy, sensitivity, specificity and Brier Score are calculated. Table 1 show results for 70% training and 30% testing data on 5, 10, 15 and 20 number of top selected features. We observe that *k*-NN and Weighted k-NN give maximum and equal accuracy for all four different number of features as compared to other methods. In the case of Sensitivity, Random Forest gives maximum value i.e., 0.962 for 5 features. SVMG gives maximum value for 10 features as compared to other methods. Thus, for 15 features SVMP and SVML give best result among all. However, SVMP outperforms when we have 20 features. For Specificity, *k*-NN gives largest value when we have 5 features and Weighted *k*-NN outperforms when 10 features are selected, while k-NN and weighted k-NN gives optimal and same results in case of 15 number of features. Similarly, Random Forest outperformed in case of 20 number of features. In terms of Brier Score, *k*-NN outperforms in all four scenarios for different number of features. Similar conclusions for Table 2 and Table 3 can be drawn.

Table 1 Accuracy, Sensitivity, Specificity and Brier Score for k-NN, Weighted k-NN, SVM (Linear, Polynomial, Gaussian), Random Forest (RF), and Optimal Trees Ensemble (OTE) using 70% training and 30% testing data while taking different number of features such as 5, 10, 15, 20, respectively.

Metric	**Methods**	**Number of Features**			
		5	**10**	**15**	**20**
Accuracy	*k*-NN	**0.894**	**0.899**	**0.897**	**0.898**
	Weighted *k*-NN	**0.894**	**0.899**	**0.897**	**0.898**
	RF	0.892	0.897	0.894	0.895
	SVM Gaussian	0.875	0.896	0.895	0.895
	SVM Polynomial	0.824	0.888	0.894	0.894
	SVM Linear	0.824	0.888	0.894	0.894
	OTE	0.893	0.896	0.895	0.896
Sensitivity	*k*-NN	0.941	0.936	0.936	0.943
	Weighted kNN	0.941	0.935	0.936	0.943
	RF	**0.961**	0.946	0.939	0.933
	SVM Gaussian	0.959	**0.954**	0.956	**0.963**
	SVM Polynomial	0.935	0.951	**0.959**	0.949
	SVM Linear	0.935	0.951	**0.959**	0.949
	OTE	0.952	0.94	0.936	0.934
Specificity	*k*-NN	**0.7**	0.746	**0.741**	0.718
	Weighted *k*-NN	0.698	**0.747**	**0.741**	0.716
	RF	0.612	0.694	0.714	**0.742**
	SVM Gaussian	0.528	0.656	0.647	0.624
	SVM Polynomial	0.369	0.626	0.63	0.67
	SVM Linear	0.369	0.626	0.63	0.67
	OTE	0.655	0.714	0.727	0.742

Table 2 Accuracy, Sensitivity, Specificity and Brier Score for k-NN, Weighted k-NN, SVM (Linear, Polynomial, Gaussian), Random Forest (RF), and Optimal Trees Ensemble (OTE) using 50% training and 50% testing data while taking different number of features such as 5, 10, 15, 20, respectively.

Metric	**Methods**	**Number of Features**			
		5	**10**	**15**	**20**
Accuracy	***k*-NN**	0.893	**0.899**	**0.899**	**0.898**
	Weighted *k*-NN	0.893	**0.899**	**0.899**	**0.898**
	RF	0.89	0.894	0.892	0.891
	SVM Gaussian	0.875	0.893	0.893	0.893
	SVM Polynomial	0.827	0.89	0.895	0.891
	SVM Linear	0.827	0.89	0.895	0.891
	OTE	**0.896**	0.893	0.894	0.894
Sensitivity	***k*-NN**	0.939	0.935	0.935	0.938
	Weighted *k*-NN	0.938	0.936	0.934	0.938
	RF	**0.962**	0.947	0.936	0.926
	SVM Gaussian	0.957	0.952	0.949	**0.954**
	SVM Polynomial	0.936	**0.958**	**0.962**	0.945
	SVM Linear	0.936	**0.958**	**0.962**	0.945
	OTE	0.953	0.942	0.925	0.932
Specificity	***k*-NN**	0.705	**0.753**	0.754	0.734
	Weighted *k*-NN	**0.707**	0.75	**0.756**	0.735
	RF	0.594	0.679	0.712	**0.749**
	SVM Gaussian	0.539	0.655	0.663	0.641
	SVM Polynomial	0.382	0.612	0.623	0.668
	SVM Linear	0.382	0.612	0.623	0.668
	OTE	0.664	0.699	0.767	0.742

Table 3 Accuracy, Sensitivity, Specificity and Brier Score for k-NN, Weighted k-NN, SVM (Linear, Polynomial, Gaussian), Random Forest (RF), and Optimal Trees Ensemble (OTE) using 30% training and 70% testing data while taking different number of features such as 5, 10, 15, 20, respectively.

Metric	**Method**	**Number of Features**			
		5	**10**	**15**	**20**
Accuracy	*k*-**NN**	**0.894**	**0.899**	**0.897**	**0.898**
	Weighted *k*-**NN**	**0.894**	**0.899**	**0.897**	**0.898**
	RF	0.892	0.897	0.894	0.895
	SVM Gaussian	0.875	0.896	0.895	0.895
	SVM Polynomial	0.824	0.888	0.894	0.894
	SVM Linear	0.824	0.888	0.894	0.894
	OTE	0.893	0.896	0.895	0.896
Sensitivity	*k*-**NN**	0.941	0.936	0.936	0.943
	Weighted *k*-**NN**	0.941	0.935	0.936	0.943
	RF	**0.961**	0.946	0.939	0.933
	SVM Gaussian	0.959	**0.954**	0.956	**0.963**
	SVM Polynomial	0.935	0.951	**0.959**	0.949
	SVM Linear	0.935	0.951	**0.959**	0.949
	OTE	0.952	0.94	0.936	0.934
Specificity	*k*-**NN**	**0.7**	0.746	**0.741**	0.718
	Weighted *k*-**NN**	0.698	**0.747**	**0.741**	0.716
	RF	0.612	0.694	0.714	**0.742**
	SVM Gaussian	0.528	0.656	0.647	0.624
	SVM Polynomial	0.369	0.626	0.63	0.67
	SVM Linear	0.369	0.626	0.63	0.67
	OTE	0.655	0.714	0.727	0.742

From the above tables, it is found that *k*-NN procedure outperformed the other methods in majority of the cases in three different training-testing partitions with different number of selected features, while weighted *k*-NN is the second method, which gives the best results as compared to the other methods. SVML and SVMP also perform well in some cases but the other two i.e., RF and OTE do not give satisfactory results. It shows that tree structure methods do not provide plausible results for predicting c-section data in this case study. Both *k*-NN and Weighted *k*-NN are based on nearest neighbourhood search, methods of similar kind are best suited for analyzing the dataset.

Furthermore, the work also assessed the significant factors associated with c-section. Factors such as mother disability, height, age, baby condition at birth, baby weight, and gross income are recorded significant factors, which cause cesarean section. The significant factors used in this study reveals that the outcomes agree with majority of the studies conducted in the literature.

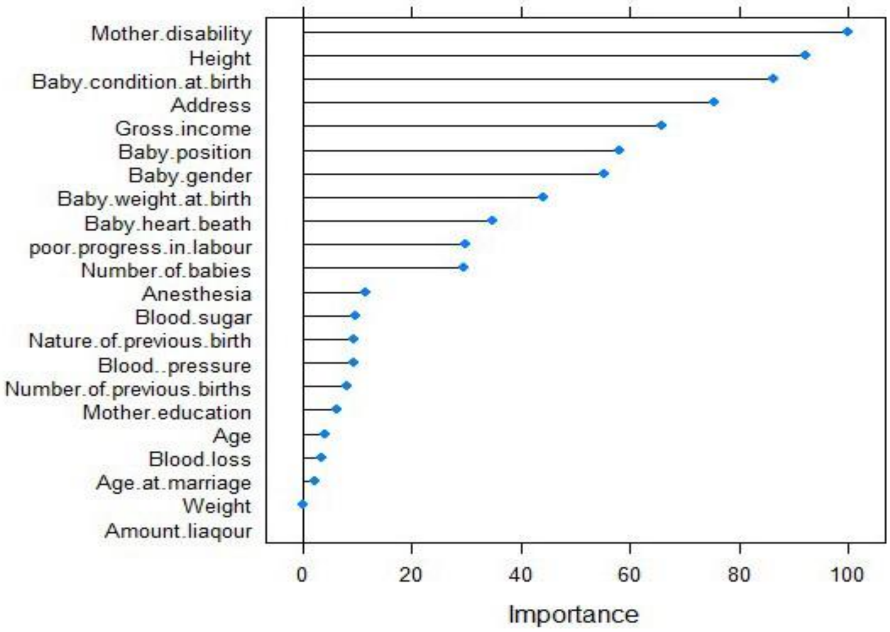

Fig. 1: Variable’s importance plot for all the variables.

Brier Score are calculated for all techniques with varying training and testing sizes and with different number of features selected by LASSO. Furthermore, boxplots are also constructed for showing the calculated results for the purpose of visualization of the machine learning algorithms used in the analysis. The boxplot in Figure 2 shows the result of Brier Score for 70% training and 30% testing data. Similarly, Figure 3 gives results of Brier Score for 50% training and 50% testing while Figure 4 presents the results of Brier Score for 30% training and 70% testing. A small value of Brier Score indicates that the computed probabilities are captured the same as the true probabilities that are not available. A machine learning method with the smallest Brier Probability forecast has been predicted by using Brier Score. A machine learning method having a smallest Brier Score is considered the best probability estimate among other techniques. The boxplots show that *k*-NN attains the smallest Brier Score in all the three scenarios.

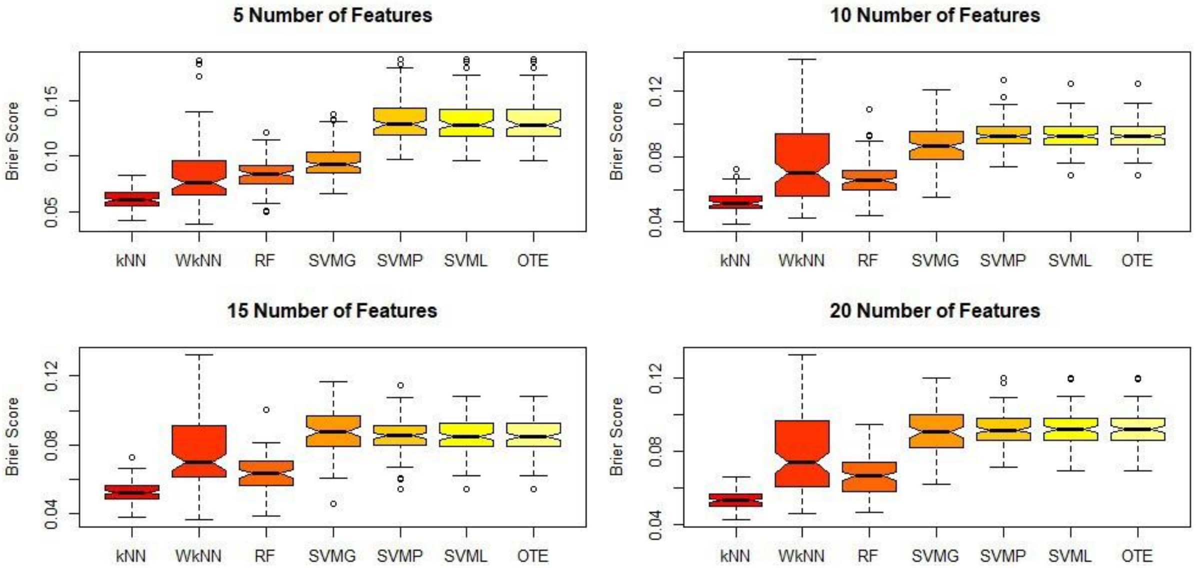

Fig. 2: Brier Score for *k*-NN, Weighted *k*-NN, SVM (Linear, Polynomial, Gaussian), Random Forest (RF), and Optimal Trees Ensemble (OTE) using 70% training and 30% testing data while taking different number of features such as 5, 10, 15, 20, respectively.

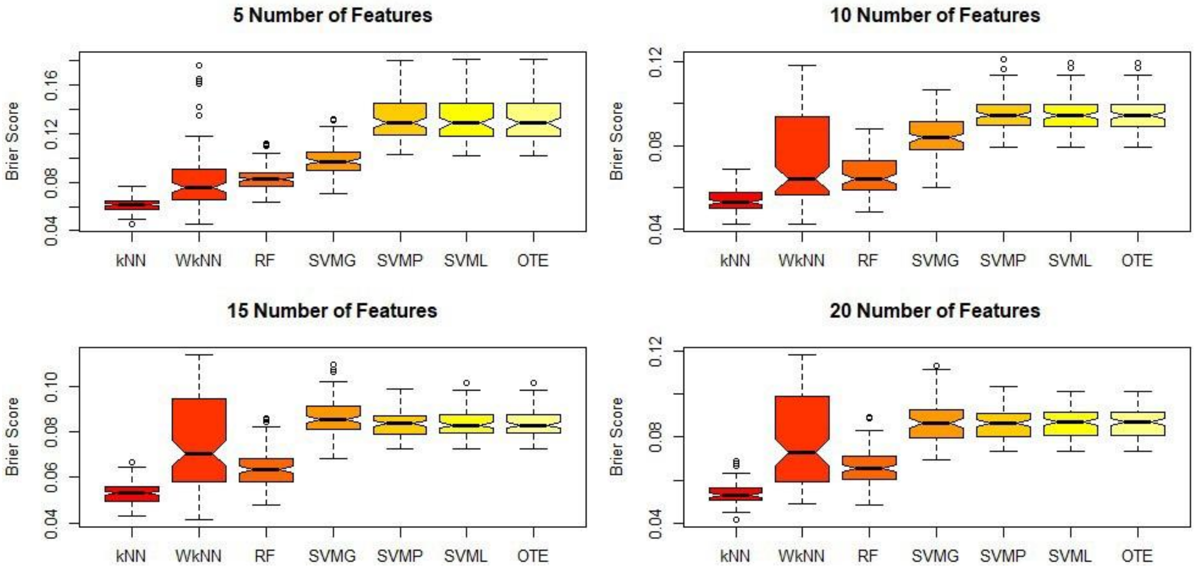

Fig. 3: Brier Score for *k*-NN, Weighted *k*-NN, SVM (Linear, Polynomial, Gaussian), Random Forest (RF), and Optimal Trees Ensemble (OTE) using 50% training and 50% testing data while taking different number of features such as 5, 10, 15, 20, respectively.

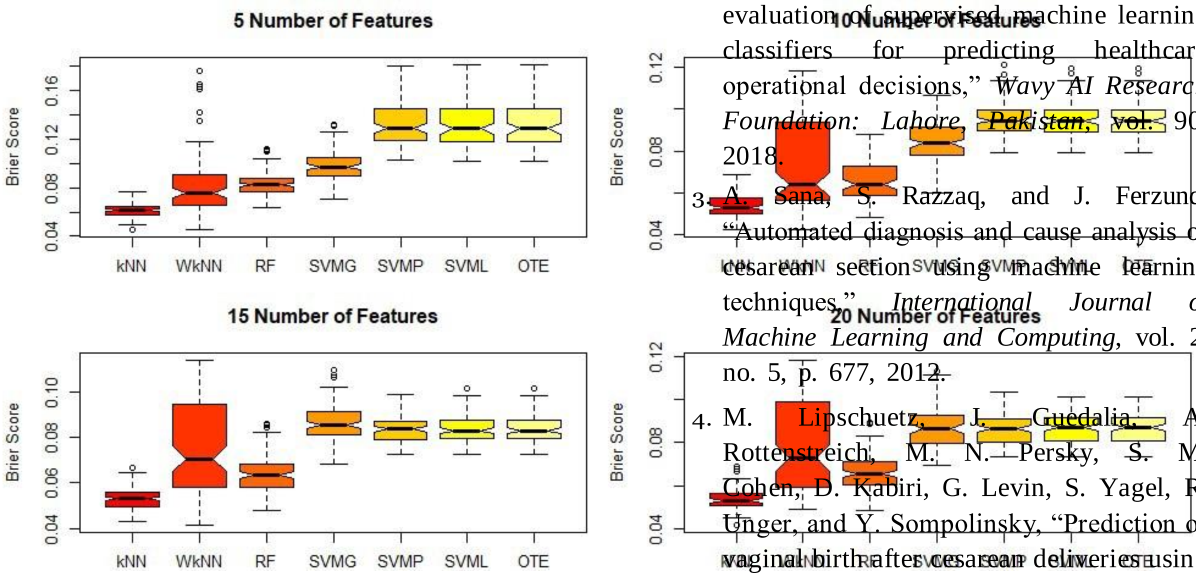

Fig. 4: Brier Score for *k*-NN, Weighted *k*-NN, SVM (Linear, Polynomial, Gaussian), Random Forest (RF), and Optimal Trees Ensemble (OTE) using 30% training and 70% testing data while taking different number of features such as 5, 10, 15, 20, respectively.

5 Conclusion

The principal objective of the study is to detect the risk factors of highest importance associated to c-section with the help of machine learning algorithms for providing an accurate and reliable results to predict new cases. For further work in this direction, other districts of the country could be included for a more generic conclusion. Various other machine learning methods could also be used for further improving prediction performance. Using various feature methods could be used to select the most regulatory risk factors.

References

1. Z. Ullah, F. Saleem, M. Jamjoom, and B. Fakieh, “Reliable prediction models based on enriched data for identifying the mode of childbirth by using machine learning methods: development study,” *Journal of Medical Internet Research*, vol. 23, no. 6, p. e28856, 2021.
2. M. Amin and A. Ali, “Performance evaluation of supervised machine learning classifiers for predicting healthcare operational decisions,” *Wavy AI Research Foundation: Lahore, Pakistan*, vol. 90, 2018.
3. A. Sana, S. Razzaq, and J. Ferzund, “Automated diagnosis and cause analysis of cesarean section using machine learning techniques,” *International Journal of Machine Learning and Computing*, vol. 2, no. 5, p. 677, 2012.
4. M. Lipschuetz, J. Guedalia, A. Rottenstreich, M. N. Persky, S. M. Cohen, D. Kabiri, G. Levin, S. Yagel, R. Unger, and Y. Sompolinsky, “Prediction of vaginal birth after cesarean deliveries using machine learning,” *American journal of obstetrics and gynecology*, vol. 222, no. 6, pp. 613–e1, 2020.
5. C. Lindblad Wollmann, K. D. Hart, C. Liu, A. B. Caughey, O. Stephansson, and J. M. Snowden, “Predicting vaginal birth after previous cesarean: using machine-learning models and a population-based cohort in sweden,” *Acta obstetricia et gynecologica Scandinavica*, vol. 100, no. 3, pp. 513–520, 2021.
6. S. Li, Z. Wang, L. A. Vieira, A. B. Zheutlin, B. Ru, E. Schadt, P. Wang, A. B. Copperman, J. Stone, S. J. Gross, *et al.*, “Improving pre-eclampsia risk prediction by modeling individualized pregnancy trajectories derived from routinely collected electronic medical record data,” *medRxiv*, pp. 2021–03, 2021.
7. D. Kavitha and T. Balasubramanian, “International journal of engineering sciences & research technology predicting the mode of delivery and the risk factors associated with cesarean delivery using decision tree model,”
8. S. A. Abbas, A. U. Rehman, F. Majeed, A. Majid, M. S. A. Malik, Z. H. Kazmi, and S. Zafar, “Performance analysis of classification algorithms on birth dataset,” *IEEE Access*, vol. 8, pp. 102146–102154, 2020.
9. M. N. Islam, T. Mahmud, N. I. Khan, S. N. Mustafina, and A. N. Islam, “Exploring

machine learning algorithms to find the best features for predicting modes of childbirth," *IEEE Access,* vol. 9, pp. 1680–1692, 2020.

10. R. R. Clark and J. Hou, "Three machine learning algorithms and their utility in exploring risk factors associated with primary cesarean section in low-risk women: A methods paper," *Research in nursing & health,* vol. 44, no. 3, pp. 559–570, 2021.

11. S. Maroufizadeh, P. Amini, M. Hosseini, A. Almasi-Hashiani, M. Mohammadi, B. Navid, and R. Omani-Samani, "De- terminants of cesarean section among primiparas: a comparison of classification methods," *Iranian journal of public health,* vol. 47, no. 12, p. 1913, 2018.
12. A. Karacı, "Evaluation of deep neural network and ensemble machine learning methods for cesarean data classification," in *Deep Learning for Biomedical Applications*, pp. 301–313, CRC Press, 2021.
13. N. I. Khan, T. Mahmud, M. N. Islam, and S. N. Mustafina, "Prediction of cesarean childbirth using ensemble machine learning methods," in *Proceedings of the 22nd international conference on information integration and web-based applications & services*, pp. 331–339, 2020.
14. R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 267–288, 1996.
15. A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab-an s4 package for kernel methods in r," *Journal of Statistical Software*, vol. 11, no. 9, pp. 1–20, 2004.
16. D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2019. R package version 1.7-3.
17. Z. Khan, A. Gul, A. Perperoglou, O. Mahmoud, W. Adler, Miftahuddin, and B. Lausen, *OTE: Optimal Trees Ensembles for Regression, Classification and Class Membership Probability Estimation,* 2015. R package version 1.0.
18. A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.