International Journal of Combinatorial Optimization Problems and Informatics, 11(3), Sep-Dec 2020, 121-129. ISSN: 2007-1558.

www.editada.org

Problems in pregnancy, modeling fetal mortality through the Naïve Bayes classifier

Israel Campero-Jurado, Daniel Robles-Camarillo, Eric Simancas-Acevedo

[1] Universidad Politécnica de Pachuca, Carretera, Carretera Ciudad Sahagún-Pachuca Km. 20, Ex-Hacienda de Santa Bárbara, 43830 Hgo. Mexico.

Abstract. Fetal mortality represents a problem in society, deaths are related to different causes. However, there are also a large number of unknown causes. Studies around the world, broadly speaking, have found that a total of 275,914 strong fetuses occurred in the United States between 2004 and 2014 compared to 7,571 maternal deaths. The Gaussian Naïve Bayes classifier is presented to assist in the fetal reduction mortality with four variables obtained through INEGI which are: gender, gestational age, maternal age, fetuses (a single pregnancy or a multiple pregnancy). The sample has a total of 2000 data where there is a balance of classes with 50% of cases of fetal death against 50% of births without complications in the year 2017. As a result the number of mislabeled points out of a total 400 points were 17. With a percentage of 96%, 96% and 96% for measures precision, recall and f1-score respectively.

Keywords:

Article Info
Received June 11, 2019
Accepted Dec 11, 2019

1. Introduction

In general, intrauterine fetal mortality is defined as death that occurs before the expulsion of the mother or the complete removal of the product of conception, whatever the duration of the gestation [1]. Fetal loss has been defined as the cessation of fetal life after 20 weeks of gestation and weighing more than 500 grams.

Fetal deaths from a combination of known and unknown causes are a problem that significantly affects the population. Many fetal deaths are related to different disorders [2] such as pregnancy-induced hypertension, poor maternal nutrition, addictions, etc.

Also, fetal deaths are largely related to maternal deaths [3], which represents a significant problem for society. This is due to inconsistency in comparing the frequency of fetal death between different populations, such as lack of agreed or uniformly accepted definition, use of different rates (fetal mortality rate, perinatal mortality rate), mode of estimation of gestational age and non-comparable research designs [1].

Nowadays, more and more new tools have been chosen in the field of computer science, which solve more complex problems. Such data can often have an economic, social or biological origin, as it is the case in the present work. We propose a Gaussian Naïve Bayes classifier capable of anticipating through different descriptive characteristics when a pregnancy is at risk of abortion for unknown or known causes.

The purpose of this paper is to provide an overview of the work done in different institutions, then detail the methodology used, model the dataset and finally show the results obtained.

2. State of the art

In 2001, Rish [4] analyzed the impact of the distribution entropy on the classification error, and he demonstrated that naïve Bayes worked well for certain nearly functional feature dependencies, thus reaching its best performance in two different cases: completely independent features and functionally dependent features.

In 2012, Saurabh Mukherjee and Neelam Sharma identified important reduced input features in the building of the intrusion detection system (IDS) that is computationally efficient and effective. They [5] investigated the performance of three standard feature selection methods using Correlation-based Feature Selection, Information Gain and Gained Ratio. Also, they proposed a method Feature Vitality Based Reduction Method, to identify important reduced input features.

In April 2009, Jingnian Chen et al. [6] presented two feature evaluation metrics for the Naïve Bayesian classifier applied on multi-class text datasets: Multi-class Odds Ratio (MOR), and Class Discriminating Measure (CDM). Experiments of text classification with Naïve Bayesian classifiers were carried out on two multi-class texts collections.

In 2011, Ting et al. [7] proposed an article aimed to highlight the performance of employing Naïve Bayes in document classification. Results show that NaïveBayes was the best classifiers against several common classifiers (such as decision tree, neural network, and support vector machines) in term of accuracy and computational efficiency

In 2010, Srinivas et al. [8] examined the potential use of classification based data mining techniques such as Rule-based, Decision tree, Naïve Bayes and Artificial Neural Network to massive volume of healthcare data. Moreover, using medical profiles such as age, sex, blood pressure and blood sugar, it could predict the likelihood of patients getting a heart disease.

In 2012, Shadab Adam Pattekari and Asma Parveen [9] developed an Intelligent System using data mining modeling technique, namely, Naïve Bayes. It was implemented as a web-based application in this user answered the predefined questions. It retrieved hidden data from stored database and compared the user values with trained data set. It could answer complex queries for diagnosing heart disease and thus assisted healthcare practitioners to made intelligent clinical decisions which traditional decision support systems could not. By providing effective treatments, it also helps to reduce treatment costs.

In 2017, Tejaswinee A. Shinde and Jayashree R. Prasad [10] proposed system aimed to assess the data mining techniques and apply them to Animal database to establish meaningful relationships. This study focused on the Naïve Bayes Classification method of data mining to classify animal sensor data. The proposed system consists of animal health care benefiting the farmers by using Wireless Sensor Network technology and IoT applications.

In 2015, Vembandasamy et al. [11] analysed a few parameters and predicts heart diseases, thereby suggests a heart diseases prediction system (HDPS) based on the data mining approaches.

In July 1999, Cheng and Greiner [12] empirically evaluated algorithms for learning four types of Bayesian network (BN) classifiers - Naïve-Bayes, tree augmented Naïve Bayes, BN augmented Naïve Bayes and general BNs, where the latter two were learned using two variants of a conditional-independence (CI) based BN-learning algorithm. Experimental results showed the obtained classifiers, learned using the CI based algorithms, were competitive with the best-known classifiers, based on both Bayesian networks and other formalisms: and that the computational time for learning and using these classifiers was relatively small.

In 2010, Binal. A. Thakkar et al. [13] developed prototype Intelligent Swine flu Prediction software (ISWPS). They used Naïve Bayes classifier for classifying the patients of swine flu into three categories (least possible, probable or most probable). Also, they have used 17 symptoms of Swine flu and collected 110 symptoms sets from various hospitals and medical practitioners. Using ISWPS, they have achieved an accuracy of nearly 63.33%. It was implemented on the JAVA platform.

In 2016, Langarizadeh and Moghbeli [14] reviewed published evidence about the application of Naive Bayesian networks (NBNs) in predicting disease, and it tried to show NBNs as the fundamental algorithm for the best performance in comparison with other algorithms. Finally, the results were reported in terms of Accuracy, Sensitivity, Specificity and Area under ROC curve (AUC).

In 2003, Price et al. [15] developed a decision support system (DSS) for the histological interpretation of these lesions. Knowledge and uncertainty were represented in the form of a Bayesian belief network that permitted the storage of diagnostic knowledge and, for a given case, the collection of evidence in a cumulative manner that provided a final probability for the possible diagnostic outcomes. The network comprised 8 diagnostic histological features (evidence nodes) that were each independently linked to the diagnosis (decision node) by a conditional probability matrix.

In 2016, Ben-Assuli and Leshno [16] evaluated the impact of an electronic health record on emergency department physicians' diagnosis and admission decisions. A decision-analytic approach using a decision tree was constructed to model the admission decision process to assess the added value of medical information retrieved from the electronic health record. Using a Bayesian

statistical model, this method was evaluated on two coronary artery disease scenarios. The results show that the cases of coronary artery disease were better diagnosed when the electronic health record was consulted and led to more informed admission decisions.

In 1997, Nir Friedman et al. [17] evaluated approaches for inducing classifiers from data, based on the theory of learning Bayesian networks. These networks were factored representations of probability distributions that generalize the naive Bayesian classifier and explicitly represent statements about independence. Among these approaches, they singled out a method they called Tree Augmented Naïve Bayes (TAN), which outperforms naïve Bayes, yet at the same time maintains the computational simplicity and robustness that characterize naïve Bayes.

In 2012, Muralidharan and Sugumaran [18] presented the use of Naïve Bayes algorithm and Bayes net algorithm for fault diagnosis through discrete wavelet features extracted from vibration signals of good and faulty conditions of the components of centrifugal pump. The classification accuracies of different discrete wavelet families were calculated and compared to find the best wavelet for the fault diagnosis of the centrifugal pump.

In 2015, Wolfson et al. [19] proposed an adaptation of the well-known Naïve Bayes machine learning approach to time-to-event outcomes subject to censoring. They compared the predictive performance of their method with the Cox proportional hazards model which was commonly used for risk prediction in healthcare populations and illustrated its application to prediction of cardiovascular risk using an electronic health record dataset from a large Midwest integrated healthcare system.

In 2015, Benndorf et al. [20] developed and validated a decision support tool for mammographic mass lesions based on standardized descriptor terminology (BI-RADS lexicon) to reduce the variability of practice. They used separate training data (1,276 lesions, 138 malignant) and validation data (1,177 lesions, 175 malignant). They created naïve Bayes classifiers from the training data with tenfold cross-validation. Their model comprised BI-RADS categories, BI-RADS descriptors, and age as predictive variables; their descriptor model comprised BI-RADS descriptors and age. The resulting Naïve Bayes classifiers were applied to the validation data. They evaluated and compared classifier performance with ROC-analysis.

In 2012, William Klement et al. [21] developed a prediction model that achieved more balanced performance (in terms of sensitivity and specificity) than the Canadian Assessment of Tomography for Childhood Head Injury (CATCH) rule, when predicting the need for computed tomography (CT) imaging of children after a minor head injury. They compared the proposed ensemble model to other ensemble models employing rule-, tree- and instance-based member classifiers. Their prediction model demonstrated the best performance in terms of AUC, G-mean and sensitivity measures. In the second phase, using a bootstrapping experiment similar to that reported by the CATCH investigators, they showed that the proposed ensemble model achieved a more balanced predictive performance than the CATCH rule with an average sensitivity of 82.8% and an average specificity of 74.4% (vs 98.1% and 50.0% for the CATCH rule respectively).

In 2012, Spilka et al. [22] analyzed fetal heart rate of normal and acidemic fetuses. They used conventional and nonlinear features for the signal analysis. Addition of nonlinear features improved accuracy of classification. The best classification results were achieved using a combination of conventional and nonlinear features with a sensitivity of 73.4%, specificity of 76.3%, and F-measure of 71.9%. The best selected nonlinear features were: Lempel Ziv complexity, Sample entropy, and fractal dimension estimated by Higuchi method.

In July 2011, Wei Wei et al. [23] evaluated the performance of an algorithm that predicts patient outcomes from genome-wide data by efficiently model averaging over an exponential number of naïve Bayes models. This model-averaged naïve Bayes (MANB) method was applied to predict late-onset Alzheimer's disease in 1411 individuals whom each had 312,318 SNP measurements available as genome-wide predictive features. Its performance was compared to that of a naïve Bayes algorithm without feature selection and with feature selection (FSNB).

In February 2008 data of 142 brain tumour patients irradiated from 2000 to 2005 were analyzed by Kazmierska and Malicki [24] Ninety-six attributes related to disease, patient and treatment were chosen. Attributes in binary form consisted of the training set for Naïve Bayes Classifier learning. Naïve Bayes calculated an individual conditional probability of being assigned to: relapse or progression (1), or no relapse or progression (0) group. Accuracy, attribute selection and quality of classifier were determined by comparison with actual treatment results, leave-one-out and cross-validation methods, respectively. High classification accuracy (84%), specificity (0.87) and sensitivity (0.80) were achieved, both for classifier training and in progressive clinical evaluation.

Finally, In October 2007 Chun et al. [25] utilized a Bayesian risk prediction model to predict the incidence of breast cancer in a high-risk population. 10-fold cross-validation was performed using a Naïve Bayes classifier. The area under the ROC curve (AUC) was used to measure prediction accuracy. These results were then compared to the ROC curve (AUC) results of the Gail Model Risk Assessment Tool.

3. Methodology

The methodology used is divided into 4 phases [26]: understanding the data, cleaning it up, modeling it and evaluation, they are described below:

Understanding the data

The data collection was obtained from the National Institute of Statistics and Geography (INEGI) from the microdata section. Two data sets were merged. One is focusing on birth rates and the other on fetal mortality. A random sample of 1000 units was collected from each of the two databases.

The merged data consists of 5 variables which are:

- Gender
- Gestational age
- Maternal age
- Fetuses (a single pregnancy or a multiple pregnancy)
- Living children

And the label, which is the dependent variable.

Data cleaning and processing

As previously proposed by [26] the data were cleaned according to common problems such as:

Missing values: solved with the Clamp transformation, see Eq. 1:

$$a_i = \begin{cases} lower & if\ a_i < lower \\ upper & if\ a_i > upper \\ a_i & Otherwise \end{cases} \quad (1)$$

As it is possible to see the outliers are shown in Fig. 1 through boxplot for the variables gender, gestationa age and maternal age.

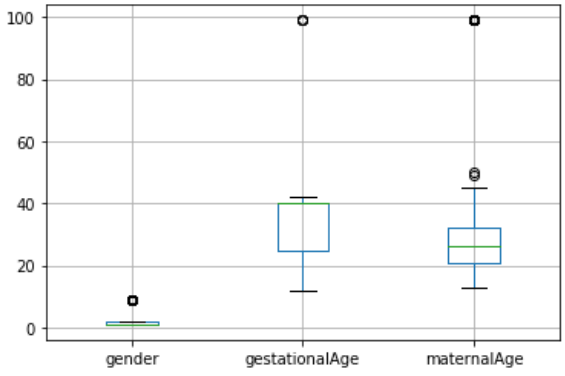

Fig. 1 Boxplot, 3 variable outlier for diagnosing fetal death

In the same way in Fig.2 the boxplot for the fetuses and living children variables are established. Here it can be see that the living children variable contains a great part of outliers. However, the information of the INEGI establishes that these values belong to missing values, Kelleher set [26] that in the absence of 60% of data it is recommended to eliminate the variable and since you have 64% of information missing living children was removed.

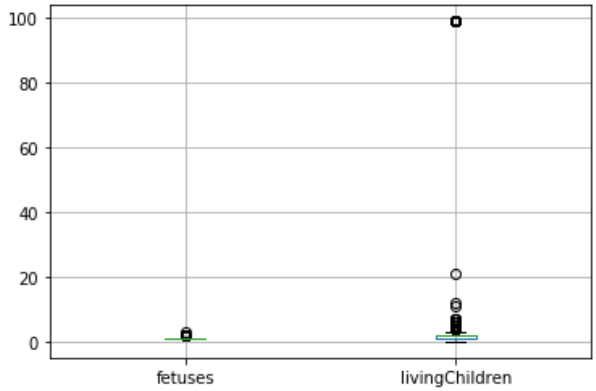

Fig. 2 Boxplot, two variable outliers for diagnosing fetal death.

Similarly, the number of fetuses is not considered for outliers as such information may be highly relevant for finding gestational problems [3].

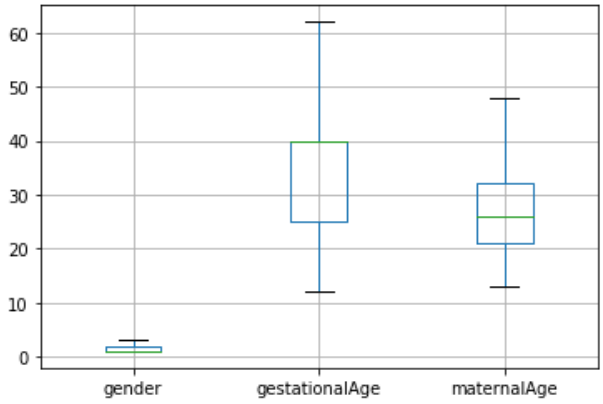

Fig. 3 Box graph, data without outliers

It was found that there was a variation in the independent linear relationship through a correlation matrix. Therefore, the 4 selected data are used to model fetal mortality.

Data modelling

Naïve Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem. The mathematical basis of the Bayes' theorem is described below [27] [28] [29]:

Given the class variable y and the dependent characteristic vector x_1 through x_n(see Eq. 2):

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n|y)}{P(x_1, \ldots, x_n)} \tag{2}$$

it can be assumed that (given the naive conditional Independence in Eq. 3):

$$P(x_i \mid y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y), \tag{3}$$

for the *i*-th, this is simplified as follows in Eq. 4:

$$P(y|x_1, \ldots, x_n) = \frac{P(y)\Pi_{i=1}^{n} P(x_i|y)}{P(x_1, \ldots, x_n)} \tag{4}$$

Given that $P(x_1, \ldots, x_n)$ is static given the input, we can use the following sorting rule [30], Eq. 5 and Eq. 6:

$$P(y \mid x_1, \ldots, x_n) \; \alpha \; P(y)\Pi_{i=1}^{n} P(x_i|y) \Downarrow \tag{5}$$

$$\hat{y} = \arg\max_{y} P(y) \, \Pi_{i=1}^{n} P(x_i|y), \tag{6}$$

and can use Maximum A Posteriori (MAP) estimation to estimate $P(y)$ and $P\,(x_i|y)$;

Naïve Bayes learners and classifiers can be extremely fast compared to more sophisticated methods. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one-dimensional distribution. This in turn helps to alleviate problems stemming from the curse of dimensionality.

The model tested was used in Python ©, where the sklearn library was used. Sklearn works with the GaussianNB variation. It implements the Gaussian Naïve Bayes algorithm for classification. The probability of the features is assumed to be Gaussian (see Eq. 7):

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\,(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}) \tag{7}$$

The parameters σ_y and μ_y are estimated using maximum likelihood.

4. Results

Evaluation

The evaluation of the algorithms was done using the Area Under the Curve (AUC) with curves that evaluate how specific the model is against its Receiver Operating Characteristic (ROC) Curve sensitivity. This is given by the ratio of true positives, true negatives, false positives and false negatives measures, all described in Table 1:

Table 1. Measurement ratio

		Predicted label	
		Positive	Negative
Actual Label	Positive	True Positive Rate (TPR)	False Negative (FN)
	Negative	False Positive Rate (FPR)	True Negative (TN)

The model obtained the performance shown in Table 2 where the number of mislabeled points out of a total of 400 points were 17. With a percentage of 96% accuracy.

Table 2. Accuracy obtained for Gaussian Naive Bayes Classifier

	Precisión	*Recall*	*F1-Score*	*Support*
0	0.95	0.97	0.96	201
1	0.97	0.94	0.96	199
Accuracy			0.96	400
Macro avg	0.96	0.96	0.96	400
Weighted avg	0.96	0.96	0.96	400

The ROC curve for Gaussian Naïve Bayes Classifier is shown in Fig. 3. Where the observation in Table 2 is confirmed.

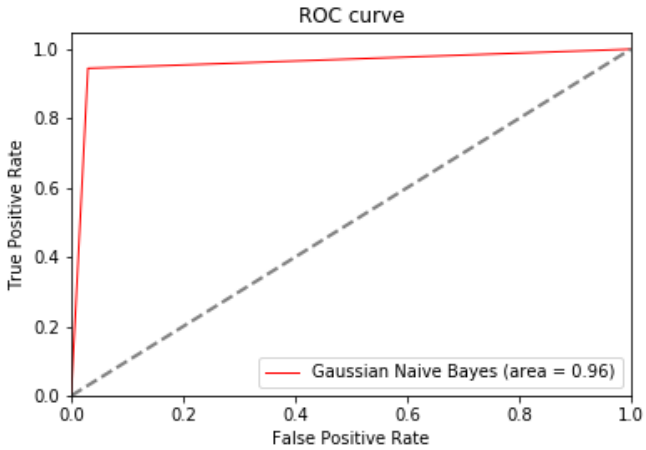

Fig. 4 Gaussian Naive Bayes Classifier for fetal deaths

5. Conclusions

Hornbuckle et al. [31] compared Bayesian classifiers to identify hypertensive disorders in pregnancy (which is one of the main problems of risk of abortion and fetal mortality). Their results showed that Naïve Bayes classifier had an excellent performance, presenting better precision and F-measure, compared to the other experimented classifiers.

Based on this premise, a Gaussian Naïve Bayes model is proposed to anticipate when a pregnancy is at risk of abortion. To this end, four variables are proposed and considered in the literature. A percentage in the model performance of 96% was obtained over a balanced data set. The ROC curves confirm what was obtained at the time of the evaluation. The present model can be embedded and generate a useful application for medical specialists.

Acknowledgements

To CONACYT for its financial support and to the Polytechnic University of Pachuca.

References

1. Vogelmann, R. A., Sánchez, J. E., Sartori, M. F., & Speciale, J. D. (2008). Muerte fetal intrauterina. *Revista de Posgrado de la VIa Cátedra de Medicina*, *188*, 10-7.
2. American College of Obstetricians and Gynecologists. (2013). Hypertension in pregnancy. Report of the American College of Obstetricians and Gynecologists' task force on hypertension in pregnancy. *Obstetrics and gynecology*, *122*(5), 1122.

3. Shah, K. H., Simons, R. K., Holbrook, T., Fortlage, D., Winchell, R. J., & Hoyt, D. B. (1998). Trauma in pregnancy: maternal and fetal outcomes. *Journal of Trauma and Acute Care Surgery*, *45*(1), 83-86.
4. Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46).
5. Mukherjee, S., & Sharma, N. (2012). Intrusion detection using naive Bayes classifier with feature reduction. *Procedia Technology*, *4*, 119-128.
6. Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*, *36*(3), 5432-5435.
7. Ting, S. L., Ip, W. H., & Tsang, A. H. (2011). Is Naive Bayes a good classifier for document classification. *International Journal of Software Engineering and Its Applications*, *5*(3), 37-46.
8. Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering (IJCSE)*, *2*(02), 250-255.
9. Pattekari, S. A., & Parveen, A. (2012). Prediction system for heart disease using Naïve Bayes. *International Journal of Advanced Computer and Mathematical Sciences*, *3*(3), 290-294.
10. Shinde, T. A., & Prasad, J. R. (2017). IoT based animal health monitoring with naive Bayes classification. *IJETT*, *1*(2).
11. Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart diseases detection using Naive Bayes algorithm. *International Journal of Innovative Science, Engineering & Technology*, *2*(9), 441-444.
12. Cheng, J., & Greiner, R. (1999, July). Comparing Bayesian network classifiers. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (pp. 101-108). Morgan Kaufmann Publishers Inc..
13. Thakkar, B. A., Hasan, M. I., & Desai, M. A. (2010, October). Health care decision support system for swine flu prediction using naïve Bayes classifier. In *2010 International Conference on Advances in Recent Technologies in Communication and Computing* (pp. 101-105). IEEE.
14. Langarizadeh, M., & Moghbeli, F. (2016). Applying naive bayesian networks to disease prediction: a systematic review. *Acta Informatica Medica*, *24*(5), 364.
15. Price, G. J., McCluggage, W. G., Morrison, M. L., McClean, G., Venkatraman, L., Diamond, J., & Hamilton, P. W. (2003). Computerized diagnostic decision support system for the classification of preinvasive cervical squamous lesions. *Human pathology*, *34*(11), 1193-1203.
16. Ben-Assuli, O., & Leshno, M. (2016). Assessing electronic health record systems in emergency departments: using a decision analytic Bayesian model. *Health informatics journal*, *22*(3), 712-729.
17. Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, *29*(2-3), 131-163.
18. Muralidharan, V., & Sugumaran, V. (2012). A comparative study of Naïve Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis. *Applied Soft Computing*, *12*(8), 2023-2029.
19. Wolfson, J., Bandyopadhyay, S., Elidrisi, M., Vazquez-Benitez, G., Vock, D. M., Musgrove, D., & O'Connor, P. J. (2015). A Naive Bayes machine learning approach to risk prediction using censored, time-to-event data. *Statistics in medicine*, *34*(21), 2941-2957.
20. Benndorf, M., Kotter, E., Langer, M., Herda, C., Wu, Y., & Burnside, E. S. (2015). Development of an online, publicly accessible naive Bayesian decision support tool for mammographic mass lesions based on the American College of Radiology (ACR) BI-RADS lexicon. *European radiology*, *25*(6), 1768-1775.
21. Klement, W., Wilk, S., Michalowski, W., Farion, K. J., Osmond, M. H., & Verter, V. (2012). Predicting the need for CT imaging in children with minor head injury using an ensemble of Naive Bayes classifiers. *Artificial intelligence in medicine*, *54*(3), 163-170.
22. Spilka, J., Chudáček, V., Koucký, M., Lhotská, L., Huptych, M., Janků, P., ... & Stylios, C. (2012). Using nonlinear features for fetal heart rate classification. *Biomedical signal processing and control*, *7*(4), 350-357.
23. Wei, W., Visweswaran, S., & Cooper, G. F. (2011). The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. *Journal of the American Medical Informatics Association*, *18*(4), 370-375.
24. Chun, J., Schnabel, F., & Ogunyemi, O. (2007, October). Assessing a Bayesian risk prediction model in a high-risk breast cancer population. In *AMIA Annual Symposium Proceedings* (pp. 913-913).
25. Kelleher, J. D., Mac Namee, B., & D'arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT press.
26. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge university press.
27. Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006, July). Spam filtering with naive bayes-which naive bayes?. In *CEAS* (Vol. 17, pp. 28-69).

28. McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (Vol. 752, No. 1, pp. 41-48).
29. Zhang, H. (2004). The optimality of naive Bayes. *AA*, *1*(2), 3.
30. Moreira, M. W., Rodrigues, J. J., Oliveira, A. M., Saleem, K., & Neto, A. V. (2016, September). An inference mechanism using bayes-based classifiers in pregnancy care. In *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)* (pp. 1-5). IEEE.
31. Lindley, D. V. (1958). Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society. Series B (Methodological)*, 102-107.
32. Moaddab, A., Clark, S. L., Dildy, G. A., Belfort, M. A., Sangi-Haghpeykar, H., & Davidson, C. (2019). Maternal and fetal death on weekends. *American journal of perinatology*, *36*(02), 184-190.