Feature Selection Based on Naive Bayes for Caesarean Section Prediction

T Desyani1*, A Saifudin2, Y Yulianti3
1,2,3Informatics Engineering, Pamulang University, Jalan Raya Puspitek 46, Banten 15310, Indonesia
Email : *dosen00839@unpam.ac.id

Abstract. Data mining using machine learning algorithms can be used to help analyze historical data to predict the need for a caesarean section. The dataset used for predicting caesarean section has many features, but those features have the possibility of redundancy or irrelevance that can cause a decrease in classifier performance. This research proposes a model that implements feature selection to select relevant features and can provide improved performance predictions for caesarean section. Some proposed feature selection techniques are Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), Sequential Forward Floating Selection (SFFS), Sequential Forward Floating Selection (SBFS), and selectKBest. The classification algorithm used to classify is Naive Bayes. The model that gives the best performance value is the model that applies the SelectKbest as feature selection.

1. Introduction

Excessive interpretation of Cardiotocography (CTG) is a common and direct cause of unnecessary cesarean section (CS)[1]. Cardiotocography (CTG) is the most common method used to monitor the fetus during the early stages of delivery and clinical decisions are made using the visual inspection of CTG traces. Caesarean section is a surgical procedure that is usually done in the process of giving birth to prevent maternal and child deaths due to certain medical conditions[2][3]. Caesarean section has improved, but is this the right decision? Predictions and judgment about the need for a caesarean are very important because the cesarean section has a risk in mothers who have had a cesarean for a long time. Before performing a cesarean section, considerations of many things are needed. In certain conditions decision-making must be taken quickly. Not all doctors have the same knowledge and experience, so they can make the wrong decision. The short-term adverse associations of caesarean delivery for the mother, such as infection, haemorrhage, visceral injury, and venous thromboembolism[4].

Before performing a cesarean section, considerations of many things are needed. In certain conditions decision-making must be taken quickly. Not all doctors have the same knowledge and experience, so they can make the wrong decision.

Data mining is a method that can be used by computers to find out the pattern of a set of data. By using computer data mining can be used to assist in decision making. One data mining method that can be used to help decision making is classification. This method is used to classify new data into certain groups based on the data set provided. So this method can be used to classify whether a mother who is going to deliver a cesarean section or not based on her condition.

The dataset used for predicting caesarean section has many features, but these features have the possibility of redundancy or irrelevance that can cause a decrease in the performance of classifiers[5].
This research proposes to implement feature selection to choose the relevant features and can provide improved performance models for caesarean section predictions. Some proposed feature selection techniques are Sequential Forward Selection (SFS), Sequential Backward Selection (SBS)[6], Sequential Forward Floating Selection (SFFS), Sequential Forward Floating Selection (SBFS), Sequential Backward Floating Selection (SBFS)[7]. The classification algorithm used to classify is Naive Bayes.

2. Method

This work is experimental research, the experiment was carried out by making a program using the Python language to implement the feature selection algorithm and the classification algorithm use. This experiment carried out by proposing caesarean prediction models, then applying to the caesarean dataset. In this experiment, a general dataset was obtained from https://archive.ics.uci.edu/ml/datasets/Caesarian+Section+Classification+Dataset. The dataset consists of 6 attributes shown in Table 1. This dataset was collected and used in the research of Gharehchopogh, Mohammadi, and Hakimi[8] as well as in Amin and Ali’s research[9]. The results of model performance measurements are compared to get the best model.

| Attributes      | Type  | Value                                      |
|-----------------|-------|--------------------------------------------|
| Age             | Numerical | Mother’s age                               |
| No. of Pregnant | Numerical | {1, 2, 3, 4}                              |
| Delivery Time   | Nominal | {0, 1, 2} -> {0=timely, 1=prenature, 2 latecomer} |
| Blood Pressure  | Nominal | {0, 1, 2} -> {0=low, 1=normal, 2=hight}    |
| Heart Status    | Nominal | {0, 1} -> {0=apt, 1=inept}                |
| **Caesarean**   | Label  | {0, 1} -> {0=No, 1=Yes}                    |

The proposed model of the caesarean prediction in this work is shown in Figure 1. The first is to standardize the Caesarean dataset. Then the features are selected using the proposed feature selection algorithm. The dataset of feature selection results is then used to train and test the model using a 10-fold cross validation technique. The test results are entered into a confusion matrix table to measure the performance of the model. Naïve Bayes algorithm is used as a basic classification because it has been proven to have an efficient[10], effective[11], and good performing [12] based on previous studies.

Based on the proposed model, there will be 6 models, namely NB, SFS, SBS, SFFS, SBFS and SelectKBest. The performance of the five models was compared to get the best model. SFS is a deterministic feature selection method that uses hill climbing search to add and assess all possible single attribute extensions for the current subset[13]. While SBS works in the opposite direction to SFS[14]. SFS and SBS select features in one-way, so the features that have been evaluating cannot be selected again, but these weaknesses avoided in SFFS and SBFS[15]. SelectKBest is a module in the scikit learn library that select k feature that has the highest score. The score is calculated based on univariate statistical analysis, which is an analysis of variables one by one.
The test results are entered in the confusion matrix table and calculate the performance of classifiers is carried out in the form of accuracy and AUC (Area Under the Curve). The confusion matrix is a very useful tool for analysing the performance of classifying models and being able to recognize tuples and features from different classes \([13]\). Analysis using a confusion matrix is done by counting the number of objects that are predicted correctly and incorrectly to determine the performance of the model \([14]\). Validation values that have been entered into the confusion matrix are used to calculate the Accuracy or AUC value of each model and measure the performance of the model. In this study using the python programming language that has provided many libraries that can be used, including to measure the performance of the model in the form of a matrix confusion, accuracy, and AUC.

Naïve Bayes algorithm is a machine learning method that analyzes data based on calculating probabilities. When used to analyze large-sized datasets, Naïve Bayes can provide high accuracy and speed. Naïve Bayes can also be used even if it doesn't have enough datasets with accurate results \([15]\). The Naïve Bayes equation is based on the Bayes theorem as follows:

\[
P(C|x) = \frac{P(C)P(x|C)}{P(x)}
\]  

In the above equation, \(C\) is a class whose probability value will be calculated because it is influenced by the value of feature \(x\). When feature \(x\) has a continuous value, it is assumed to have a Gaussian distribution with averages \((\mu)\) and standard deviations \((\sigma)\) \([16]\). So, the equation is as follows:
\[ P(x|C) = \frac{1}{\sqrt{2\pi} \sigma_r} e^{-\frac{(x-\mu)^2}{2(\sigma_o)^2}} \] (2)

3. Results and Discussion

Based on the model proposed in Figure 1, to find out the performance of the basic model applied by the Naïve Bayes algorithm as a classification without being optimized. The other model applied by integrated Naïve Bayes and feature selection, namely SFS, SBS, SFFS, SBFS, and SelectKBest.

Base on experiment, the selected feature shown in Table 2. The performance are shown in Table 3 for accuracy and Table 4 for AUC.

| Model      | Selected Feature                                                                 |
|------------|----------------------------------------------------------------------------------|
| Naïve Bayes| Heart Problem, Delivery time, Heart Problem, Delivery time, Blood of Pressure,    |
|            | Heart Problem                                                                      |
| SFS        | Heart Problem, Delivery time, Heart Problem, Delivery time, Blood of Pressure,    |
|            | Heart Problem                                                                      |
| SBS        | Heart Problem, Delivery time, Heart Problem, Delivery time, Blood of Pressure,    |
|            | Heart Problem                                                                      |
| SFFS       | Heart Problem, Delivery time, Heart Problem, Delivery time, Blood of Pressure,    |
|            | Heart Problem                                                                      |
| SBFS       | Heart Problem, Delivery time, Heart Problem, Delivery time, Blood of Pressure,    |
|            | Heart Problem                                                                      |
| SelectKBest| Heart Problem, Delivery time, Heart Problem, Delivery time, Heart Problem, Delivery|
|            | number, Delivery time, Blood of Pressure, Heart Problem                            |
| Feature Number | Naïve Bayes | SFS | SBS | SFFS | SBFS | SelectKBest |
|---------------|-------------|-----|-----|------|------|-------------|
| 1             | 61.25%      | 65.00% | 65.00% | 65.00% | 65.00% | 65.00%      |
| 2             | 61.25%      | 58.75% | 58.75% | 58.75% | 58.75% | 58.75%      |
| 3             | 61.25%      | 61.25% | 61.25% | 61.25% | 61.25% | 63.75%      |
| 4             | 61.25%      | 60.00% | 60.00% | 60.00% | 60.00% | 63.75%      |
| 5             | 61.25%      | 61.25% | 61.25% | 61.25% | 61.25% | 61.25%      |
| Average       | 61.25%      | 61.25% | 61.25% | 61.25% | 61.25% | 62.50%      |

Based on the performance in table 4 and table 5, the highest performance is obtained when using one feature, namely Heart Problem. The highest performance value is the accuracy is 65% and the AUC value is 0.673. The performance results are then visualized using the graph shown in Figure 2 and Figure 3.
Based on the values in the tables and performance graphs of the model it can be stated that the model applying the SelectKBest feature selection can provide the best value. From the results of the application of feature selection shows that not all features are relevant and can positively influence the performance of the model.

4. Conclusion

The decision to make a caesarean section or give birth normally is very important, because the main goal is to prevent maternal and child mortality. But cesarean section also has effects that can harm the mother and child, so it should be avoided except in certain conditions that can endanger the health of the mother and child. By using a dataset that has been collected and analyzed using machine learning can help make decisions in caesarean section. The experimental results show that the application of feature selection can help choose relevant features and can improve the performance of the model. The best feature selection algorithm for predicting the need for a caesarean is SelectKBest.

References

[1] P. Fergus, A. Hussain, D. Al-Jumeily, D. S. Huang, and N. Bouguila, 2017. Classification of caesarean section and normal vaginal deliveries using foetal heart rate signals and advanced machine learning algorithms, *Biomed. Eng. Online*, 16(1), pp. 1–26
[2] E. C. Romanis, 2019. Why the Elective Caesarean Lottery is Ethically Impermissible,” *Heal. Care Anal.*, 27(4), pp. 249–268
[3] K. D. Konlan, E. K. Baku, M. Japiong, K. Dodam Konlan, and R. M. Amoah, 2019. Reasons for Women’s Choice of Elective Caesarian Section in Duayaw Nkwanta Hospital, *J. Pregnancy*, 2019
[4] O. E. Keag, J. E. Norman, and S. J. Stock, 2018.Long-term risks and benefits associated with cesarean delivery for mother, baby, and subsequent pregnancies: Systematic review and meta-analysis,” *PLoS Med.*, 15(1), pp. 1–22.
[5] H. Turabieh, M. Mafarja, and X. Li, 2019. Iterated feature selection algorithms with layered
recurrent neural network for software fault prediction, *Expert Syst. Appl.*, **122**, pp. 27–42.

[6] S. Paul and S. Das, 2015. Simultaneous feature selection and weighting - An evolutionary multi-objective optimization approach, *Pattern Recognit. Lett.*, **65**, pp. 51–59.

[7] K. Homsapaya and O. Sornil, 2018. Modified Floating Search Feature Selection Based on Genetic Algorithm,” *MATEC Web Conf.*, **164**, p. 01023

[8] F. SoleimanianGharehchopogh, P. Mohammadi, and P. Hakimi, 2012. Application of Decision Tree Algorithm for Data Mining in Healthcare Operations: A Case Study,” *Int. J. Comput. Appl.*, **52**(6), pp. 21–26.

[9] M. Z. Amin and A. Ali, 2018. Performance Evaluation of Supervised Machine Learning Classifiers for Predicting Healthcare Operational Decisions Performance Evaluation of Supervised Machine Learning Classifiers for Predicting Healthcare Operational Decisions, no. December,

[10] H. Zhang, L. Jiang, and J. Su, 2005. Augmenting naive Bayes for ranking,” *Proc. 22nd Int. Conf. Mach. Learn. - ICML ’05*, **1**, pp. 1020–1027.

[11] W. Tao, L. Wei-hua, and A. Overview, 2010. Naïve Bayes Software Defect Prediction Model, no. **2006**, pp. 0–3.

[12] T. Hall, S. Beecham, D. Bowes, D. Gray, and S. Counsell, 2011. A Systematic Review of Fault Prediction Performance in Software Engineering, *Softw. Eng. IEEE Trans.*, **PP**(99), p. 1.

[13] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, 2012. *Data Mining: Concepts and Techniques*.

[14] F. Gorunescu, *Data mining: concepts and techniques*. 2011.

[15] E. Alpaydin, *Introduction to Machine Learning*, 2013. **53**(9).

[16] M. Jain and V. Richariya, 2012. An Improved Techniques Based on Naive Bayesian for Attack Detection,” *Int. J. Emerg. Technol. Adv. Eng.*, **2**(1), pp. 324–331.