Preprocessing Using Correlation Based Features Selection on Naive Bayes Classification

Yogiek Indra Kurniawan1,4, Teguh Cahyono1, Nofiyati1, Eddy Maryanto1, Ari Fadli2 and Naisha Rahma Indraswari3

1Informatics Department, Universitas Jenderal Soedirman, Indonesia
2Electrical Engineering Department, Universitas Jenderal Soedirman, Indonesia
3Informatics Department, Universitas Muhammadiyah Surakarta, Indonesia
4yogiek@unsor.ac.id

Abstract. The birth of a healthy baby is generally around 38-42 weeks of pregnancy. However, there are many babies born at an inadequate age of birth and the age of birth that is past its time. This study aims to predict the age of the birth of a patient. The method used is a classification with the Naive Bayes algorithm with input variable (X), the factors experienced by pregnant women in the form of 8 variables X and Y variable in the form of Birth Age. Problems that arise are too many attributes that affect the results of accuracy. To overcome this, preprocessing is used with the Correlation Based Features Selection (CBFS) method. CBFS chose the X variables which had the highest correlation with the Y variable (Birth Age) but had the least correlation between the X variables. From the CBFS that had been done, produced 4 X variables, namely: blood pressure, number of babies, congenital diseases before pregnancy, and problems during pregnancy. The results of the test showed an increase in Precision, recall, and accuracy in the Naive Bayes classification when implemented CBFS. The highest value of accuracy after preprocessing is 67% with an increase of 2 percent compared to before preprocessing.

1. Introduction
The birth rate of babies in Indonesia tends to increase every year. This baby's birth rate is accompanied by a large newborn mortality rate. Many things cause newborns to die, some of them are nutritional deficiencies when in the womb, fetuses that have defects in the body, premature age birth factors, postmature / postdate age birth factors, and others. Death due to premature birth is certainly a pretty serious problem. According to research conducted by WHO in 2010, Indonesia is the fifth-highest country with 675,700 premature births in one year. This figure is still quite low compared to India which reaches 3.5 million premature births per year. This causes India to rank first in the highest number of premature births among 184 other countries. In second place is China with 1.1 million premature births followed by Nigeria and Pakistan. Indonesia ranks 9th in the highest average number of premature births with birth rates of 15.5 per 100 births[1].

More than 15 million premature births occur in the world and continue to increase every year. More than 1 million children die each year due to complications due to premature birth. Premature birth is the cause of death of newborns age less than 4 weeks and is the second leading cause of death in children under 5 years other than pneumonia. Meanwhile, babies who successfully survive
experience disabilities have experience throughout their lives including learning and visual impairments and hearing problems[2].

Based on the problems mentioned above, we must know what factors influence the age of birth in infants so that nothing untoward happens in the future. One solution that can solve that problem is to classify data mining. In the classification process, it is found a set of objects with the same characteristics in a database, then classifies them into some classes that have been determined through the process of building the model. The purpose of classification is to find a model of training data or training data that distinguishes objects into appropriate categories or classes, the model is then used to classify objects whose class has not been previously known. There are several classification methods in Data Mining, including Decision Tree Classifier, Bayesian Classifier, Rule-Based Classifier, etc[3]. This research aims to predict the age of birth into 3 categories based on factors experienced by pregnant women using the classification algorithm. The categories of age at birth in this study include premature, normal, or term and postdate.

The method used to predict the age of birth in this study is classification using the Naïve Bayes algorithm. Naïve Bayes is a simple probabilistic based prediction technique based on the application of Bayes theorem or Bayes rules with strong (naïve) independence assumptions[4]. The Bayes Method is a statistical approach to induction inference on classification problems.

In this study, data taken from pregnant women patients from RSUD. Dr. Moewardi Central Java Province and An-Nisa Primary Clinic as much as 550 data. 8 attributes are used as variable X in the initial data, namely: age, blood pressure, number of babies, history of childbirth, history of abortion, malnutrition, other diseases, and current pregnancy problems. The attribute used as the Y variable is the age of birth which contains normal, premature, and postdate data. However, many attributes can make the accuracy of the Naïve Bayes algorithm worse. Therefore, it takes preprocessing data to eliminate the attributes that affect the Y variable.

The Correlation-based feature selection or abbreviated as CBFS is a simple filter algorithm that ranks the subset based on a correlation-based heuristic evaluation function. Based on the hypothesis that a good attribute is an attribute that has a high correlation with the class and does not correlate to each other. The high correlation between attributes indicates that the attribute is redundant. The attributes which have a low correlation to the class is said to be an irrelevant attribute. The irrelevant attribute and redundant then must be eliminated[5].

Correlation Based Features Selection (CBFS) has been implemented on several classification algorithms, such as K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), Artificial Neural Network and Decision Tree which show increased accuracy in the classification[6]. This research[6] shows that CBFS can be applied to classification and can improve accuracy. However, in previous studies, the CBFS method was not applied to the Naive Bayes algorithm.

In this research, Preprocessing was implemented using the Correlation Based Features Selection (CBFS) method on Naïve Bayes Algorithm. CBFS chooses X variables which have the highest correlation with the Y variable but has the least correlation between X variables[7]. After that, measurements are made between the accuracy before using preprocessing and after implementing preprocessing to find out how much effect the use of CBFS on the Naïve Bayes algorithm.

2. Method
This part of the method will explain the stages carried out in this research from beginning to end. The Method from this research divided into some stages in Figure 1:

![Figure 1. Stages of Research Method](image-url)
2.1. Problem Identification
Problem identification is an initial stage that aims to determine what problems happen[8], especially present in the birth of a baby. Based on the problems in the field, a method for prediction is needed to predict birth ages into 3 age categories: normal, premature, and postdate.

2.2. Data Collection
The data obtained in this study are data obtained from medical records of childbirth patients in hospitals. Dr. Moewardi (Central Java Province, Indonesia) and An-Nisa Primary Clinic (Indonesia), with the amount of data reaching 550 rows.

Data is divided into 2 parts, namely training data and testing data. The amount of testing data is fixed at 100 rows. While the training data is divided into several parts again, namely training data of 100 rows, 200 rows, 300 rows, 400 rows, and finally 500 rows. This is intended to determine the effect of the amount of training data on the value of accuracy.

Explanation of the data taken in the patient's medical record is in table 1.

**Table 1. Data Variables**

| Variable           | Attribute       | Type         | Information                                                                 |
|--------------------|-----------------|--------------|-----------------------------------------------------------------------------|
| X1                 | Mother's age    | Polynomial   | a. Less (<20 years)                                                         |
|                    |                 |              |  b. Sufficient (20-35 years)                                                |
|                    |                 |              |  c. More (> 35 years)                                                       |
| X2                 | Blood pressure  | Polynomial   | a. Low (<90⁄70 mmHg)                                                        |
|                    |                 |              |  b. Normal (90⁄70 - 140⁄90 mmHg)                                            |
|                    |                 |              |  c. Height (> 140⁄90 mmHg)                                                  |
| X3                 | Number of Babies| Binomials    | {1, 2}                                                                      |
| X4                 | Childbirth History| Polynomial | a. Premature history                                                         |
|                    |                 |              |  b. Postmature / postdate history                                           |
|                    |                 |              |  c. History of childbirth of normal age                                      |
|                    |                 |              |  d. Short distance from previous pregnancy (<2 years)                        |
|                    |                 |              |  e. First pregnancy                                                          |
| X5                 | Abortion history| Binomials    | {Yes, No}                                                                   |
| X6                 | Malnutrition    | Polynomial   | {Less, Normal, More}                                                        |
| X7                 | Other Diseases  | Polynomial   | a. Heart                                                                    |
|                    |                 |              |  b. Asthma                                                                  |
|                    |                 |              |  c. Hypertension                                                            |
|                    |                 |              |  d. Anemia                                                                  |
|                    |                 |              |  e. Diabetes mellitus                                                        |
|                    |                 |              |  f. HIV                                                                     |
|                    |                 |              |  g. There is no                                                             |
| X8                 | Problems During | Polynomial   | a. Mild pre-eclampsia                                                       |
| This Pregnancy     |                 |              |  b. Severe pre-eclampsia                                                   |
|                    |                 |              |  c. Gestational hypertension                                                |
|                    |                 |              |  d. Bleeding                                                                |
|                    |                 |              |  e. There is no                                                             |
| Y                  | Age of Birth    | Label        | a. Premature (<37 weeks)                                                    |
|                    |                 |              |  b. Normal (38 - 42 weeks)                                                  |
|                    |                 |              |  c. Postmature / postdate (> 42 weeks)                                      |
2.3. Algorithm Implementation

At this stage, it is divided into 2 phases. The first phase is applying Naive Bayes to classify cases of age of birth. The second phase is implementing CBFS as preprocessing before implementing the Naive Bayes algorithm.

For the first phase, the method used to predict the age of birth in this study is classification using the Naive Bayes algorithm. Naive Bayes is a simple probabilistic based prediction technique based on the application of Bayes theorem or Bayes rules with strong (naïve) independence assumptions[4]. The Bayes Method is a statistical approach to induction inference on classification problems.

The advantage of using the Naïve Bayes classification is this method only requires very little training data to create the (mean and variance of variables) needed for classification. Because the independent variable is assumed, only the variance of each class variable must be determined and not the whole covariance matrix[9].

Bayes’ prediction is based on the Bayes theorem with the general formula contained in the equation 1:[10]

\[
P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}
\] (1)

Information :

\( P (H | E) \): The conditional probability of a hypothesis \( H \) occurs if the evidence is given.
\( P (E | H) \): Probability of evidence \( E \) will affect the hypothesis \( H \).
\( P (H) \): The initial probability of hypothesis \( H \) occurs regardless of any evidence.
\( P (E) \): The initial probability of proof \( E \) occurs regardless of the hypothesis / other evidence.

For the second phase, CBFS was implemented as a preprocessing stage before applying the Naive Bayes algorithm. Following steps are to calculate the correlation value on CBFS:

1. Calculate the correlation coefficient with Symmetrical Uncertainty by using the following equation :

\[
Gain(S, A) = Entropy(S) - \sum_{t=1}^{n} \frac{|S_i|}{|S|} \times Entropy(S_i)
\] (2)

Where \( S \) is a special set, \( A \) is an attribute, \( n \) is a total number of a partition of \( A \) attribute, \( |S_i| \) is a total number of cases i-th partition and \( |S| \) is a total number of cases in \( S \).

\[
SU = 2.0 \times \left( \frac{Gain}{H(X) + H(Y)} \right)
\] (3)

Where \( H \) is entropy attribute, \( X \) is Attribute 1 and \( Y \) is Attribute 2.

2. Calculate the value of the merits by using the equation (4) :

\[
Merits = \left( \frac{k \times \tilde{r}_e}{\sqrt{k + k(r-1)\tilde{r}_f}} \right)
\] (4)

Where \( K \) is the total number of attributes, \( \tilde{r}_{e} \) is the correlation between class and attribute and \( \tilde{r}_{ff} \) is intercorrelation between attribute and attribute.

3. Perform the merits calculation for all possible attribute selection combinations to find the highest Merits

CBFS is done to eliminate the variable \( X \) which is less influential on the variable \( Y \). After elimination, then the Naive Bayes algorithm is applied to the remaining variable \( X \).

2.4. Testing

Testing is done to find out whether the system or algorithm is running[11], to find out the performance of the algorithm that has been implemented and to test calculations in training data by using data
During this testing phase, calculation of the level of precision, recall, and accuracy is also carried out to test the accuracy of the application being built.

A Precision is a calculation of the estimated proportion of true positive cases and is formulated in equation 5[13]:

\[
Precision = \frac{TP}{TP+FP}
\]  

A Recall is a calculation of the estimated proportion of positive cases that are correctly identified and formulated in equation 6:

\[
Recall = \frac{TP}{TP+FN}
\]

An Accuracy is a calculation of the proportion of the total number of correct predictions and is formulated in equation 7:

\[
Accuracy = \frac{TP+TN}{TP+FP+TN+FN}
\]

Where

TP : True Positive
TN : True Negative
FP : False Positive
FN : False Negative

2.5. Analysis and Discussion

At this stage, analysis and discussion of the test results are carried out. The first analysis is about the effect of the amount of training data on the accuracy value of the Naive Bayes algorithm. The next analysis is about a comparison between the accuracy value of the pure Naive Bayes Algorithm and the accuracy value of the Naive Bayes Algorithm that has been applied by CBFS before. After that, conclusions are drawn from the research that has been done.

3. Result and Discussion

This section contains the results of the test as well as analysis and discussion of the results of the tests that have been carried out.

The amount of data in this research is 550 data, divided into 2 parts, namely training data and testing data. The amount of testing data is fixed at 100 rows. While the training data is divided into several parts again, namely training data of 100 rows, 200 rows, 300 rows, 400 rows and finally 500 rows. This is intended to determine the effect of the amount of training data on the value of accuracy.

3.1. Precision, Recall and Accuracy Testing on Naïve Bayes Algorithm

The first testing has been done for the precision, recall, and accuracy of the Naive Bayes algorithm on each training data. The results of the test can be seen in table 2.

| Training Data | Precision | Recall | Accuracy |
|---------------|-----------|--------|----------|
| 100           | 0.567     | 0.58   | 58%      |
| 200           | 0.62      | 0.62   | 62%      |
| 300           | 0.63      | 0.63   | 63%      |
| 400           | 0.68      | 0.64   | 64%      |
| 500           | 0.703     | 0.65   | 65%      |
3.2. **Precision, Recall and Accuracy Testing on CBFS before Applying Naïve Bayes Algorithm**

Before the second testing, CBFS has been implemented on the training data to get optimum variables. From the CBFS that had been done, produced 4 X variables, namely: blood pressure, number of babies, congenital diseases before pregnancy, and problems during pregnancy. After CBFS has been implemented as preprocessing before applied Naïve Bayes, testing has been done for precision, recall, and accuracy. The results of the test can be seen in table 3.

### Table 3. Precision, Recall and Accuracy Testing on CBFS before applying Naïve Bayes Algorithm

| Training Data | Precision | Recall | Accuracy |
|---------------|-----------|--------|----------|
| 100           | 0.484     | 0.55   | 55%      |
| 200           | 0.495     | 0.59   | 59%      |
| 300           | 0.64      | 0.63   | 63%      |
| 400           | 0.70      | 0.65   | 65%      |
| 500           | 0.72      | 0.67   | 67%      |

3.3. **Discussion**

From table 2, the result of the test about Precision and Recall on Naïve Bayes algorithm can be processed into figure 2.

![Figure 2. Precision and Recall Trends on Naïve Bayes Algorithm](image)

From figures 2 and 3, it can be seen that the value of Precision and recall from both testings tends to increase when the amount of training data increases. This is because the more the amount of training data, then the possibility of similarity of data between training data with testing data is also increasing, so it can increase the value of precision and recall.

Higher levels of precision and recall indicate that the data in the application have a collective level of data high enough to measure the closeness between the actual value and the predicted results.

From Tables 2 and 3, the result of the test about accuracy on Naïve Bayes and CBFS before Naïve Bayes can be processed into figure 4.
Based on Figure 4, it can be seen that the value of Accuracy increases with an increasing amount of training data. Also, when the amount of training data is small (100 lines and 200 lines), the accuracy in CBFS Naive Bayes is lower than the accuracy of Naive Bayes, but when the amount of training data increases, starting at 400 lines up to 500 lines of training data, the accuracy value at CBFS Naive Bayes increases and exceeds Naive Bayes accuracy. This is because, in CBFS Naive Bayes, the number of X attributes that affect the value of the Y attribute (gestational age) is getting smaller, so the more training data provided, each attribute X can have more effect for the Y attribute. With the least attribute X, then the probability of calculation is increasing which can increase the accuracy value of Naive Bayes.
4. Conclusion
Based on the results of tests that have been done, it can be concluded that the greater the amount of training data, it can increase the value of Precision, Recall, and Accuracy on Naive Bayes and Correlation Based Feature Selection (CBFS) as preprocessing before Naive Bayes. Also, accuracy values on CBFS Naive Bayes tend to increase beyond accuracy in Naive Bayes when the amount of training data is large.

References
[1] WHO, “Preterm birth,” 2018. https://www.who.int/en/news-room/fact-sheets/detail/preterm-birth (accessed Jul. 14, 2020).
[2] W. March of Dimes, PMNCH, Save the Children, Born to Soon: The Global Action Report on Preterm Birth. Geneva: Eds CP Howson, MV Kinney, JE Lawn. World Health Organization., 2012.
[3] T. M. Tran, X. M. T. Le, H. T. Nguyen, and V. N. Huynh, “A novel non-parametric method for time series classification based on k-Nearest Neighbors and Dynamic Time Warping Barycenter Averaging,” Eng. Appl. Artif. Intell., vol. 78, no. October 2017, pp. 173–185, 2019, doi: 10.1016/j.engappai.2018.11.009.
[4] L. Hou, Y. Guo, and J. Cao, “The Improved Joint Bayesian Method for Person Re-identification Across Different Camera,” vol. 15, no. 4, pp. 785–796, 2019.
[5] S. Adi, Y. Pristyanto, and A. Sunyoto, “The best features selection method and relevance variable for web phishing classification,” in 2019 International Conference on Information and Communications Technology, ICOIACT 2019, 2019, no. June 2020, pp. 578–583, doi: 10.1109/ICOIACT46704.2019.8938566.
[6] Y. Pristyanto, S. Adi, and A. Sunyoto, “The Effect of Feature Selection on Classification Algorithms in Credit Approval,” in 2019 International Conference on Information and Communications Technology (ICOIACT), 2019, doi: 10.1109/ICOIACT46704.2019.8938523.
[7] T. Astuti, R. Wahyudi, U. Hasanah, B. P. Hartato, and Z. Rifai, “The Impact of Features Selection on Performance of Artificial Neural Network in Diagnosis of Diabetic Retinopathy,” in 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2019, doi: 10.1109/ICITISEE48480.2019.9003856.
[8] Y. I. Kurniawan, A. Rahmawati, N. Chasanah, and A. Hanifa, “Application for determining the modality preference of student learning,” in Journal of Physics: Conference Series, 2019, vol. 1367, no. 1, pp. 1–11, doi: 10.1088/1742-6596/1367/1/012011.
[9] V. S and D. S, “Data Mining Classification Algorithms for Kidney Disease Prediction,” Int. J. Cybern. Informatics, vol. 4, no. 4, pp. 13–25, 2015, doi: 10.5121/ijci.2015.4402.
[10] Y. I. Kurniawan, “Perbandingan Algoritma Naive Bayes dan C.45 Dalam Klasifikasi Data Mining,” vol. 5, no. 4, pp. 455–463, 2018, doi: http://dx.doi.org/10.25126/jtiik.201854803.
[11] F. Y. Al Irsyadi, Supriyadi, and Y. I. Kurniawan, “Interactive educational animal identification game for primary schoolchildren with intellectual disability,” Int. J. Adv. Trends Comput. Sci. Eng., vol. 8, no. 6, pp. 3058–3064, 2019, doi: 10.30534/ijatcse/2019/64862019.
[12] Y. I. Kurniawan, E. Soviana, and I. Yuliana, “Merging Pearson Correlation and TAN-ELR algorithm in recommender system,” in AIP Conference Proceedings, 2018, vol. 1977, doi: 10.1063/1.5042998.
[13] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, “A comparison of machine learning techniques for customer churn prediction,” Simul. Model. Pract. Theory, vol. 55, pp. 1–9, 2015, doi: 10.1016/j.simpat.2015.03.003.