Feature Selection on Pregnancy Risk Classification Using C5.0 Method

Yufis Azhar*[1], Riz Afdian[2]
[1,2]Universitas Muhammadiyah Malang
yufis@umm.ac.id*[1], afdianriz@webmail.umm.ac.id[2]

Abstract

The maternal mortality rate in Indonesia is still relatively high. This is caused by several factors, including the ignorance of pregnant women about the risk status of pregnancy. Several methods are proposed for early detection of the risk of a mother's pregnancy. However, no one has highlighted what features are most influential in the process of classifying the risk of pregnancy. In this research, we use data of pregnant women in one of the health centers in Malang, Indonesia, as a dataset. The dataset has 107 features, therefore, feature selection is needed for the classification process. We propose to use the C5.0 method to select important features while classifying dataset into low, high, and very high risk of pregnancy. C5.0 was chosen because this method has a better pruning algorithm and requires relatively smaller memory compared to C4.5. Another classification method (SVM, Naive Bayes, and Nearest Neighbor) is then used to compare the accuracy values between datasets that use all features with datasets that only use the selected features. The test results show that feature selection can increase accuracy by up to 5%.

Keywords: *Pregnancy Risk Classification, Feature Selection, C5.0*

1. Introduction

Women's deaths due to pregnancy problems are still often found in developing countries such as Indonesia. This case often appears in rural areas with limited access to health services. Malang is one of the regions in Indonesia with a high maternal mortality rate. In 2010, there were 92 mothers who died from 100,000 births. Most maternal deaths are caused by congenital diseases such as hypertension during pregnancy, diabetes mellitus, or preeclampsia [1]. Therefore, a preventive effort is needed to reduce this mortality rate.

The method often used to reduce maternal mortality is by early detection of the risk of pregnancy. Early detection is done by analyzing the inherent features of a pregnant woman, including medical history, previous pregnancy history, and current health conditions. Because of the many features that must be analyzed, this early detection process becomes quite complicated. Several studies in the health sector proposed several different features to determine the risk of pregnancy. Tejayanti, et al. [1] conducted observations on the cases of maternal mortality in Malang Regency, East Java. The results obtained show that most cases of maternal mortality are caused by hypertension during pregnancy, other non-specific causes (such as accidents), and postpartum bleeding. Utama, et al [2], in his study with subjects taking at Raden Mattaher Hospital in Jambi, said that in addition to hypertension, the mother's age during pregnancy also had an effect on increasing the risk of pregnancy. Utama said that a safe age for mothers to get pregnant is in the range of 20-35 years old. Meanwhile, there is no strong evidence to suggest that gravida status affects the increased risk of pregnancy. On a broader scale, Afifah, et al. [3] conducted a study taking data from the Indonesian population census data in 2010. The results showed that the most cases of death in pregnant women occurred in the age range under 15 years. The cause of death is postpartum hemorrhage, edema, proteinuria, and hypertension. The results of his research also show that the main causes of maternal mortality vary between regions.

Because the cause of death of pregnant women varies, the number of features that medical personnel must observe to determine the level of risk for pregnancy is also increasing. In the city of Malang itself, some health services use a pregnant mother's card to record all features related to the risk of pregnancy. These features will later be used to classify pregnant women into three categories, namely pregnancies with low risk, high risk, and very high risk. Although the number of features recorded in the card is quite a lot, in reality not all features are important features.

Azhar, Y., & Afdian, R. (2018). Feature Selection on Pregnancy Risk Classification Using C5.0 Method. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 3(4). doi:http://dx.doi.org/10.22219/kinetik.v3i4.703
Receive September 18, 2018; Revise September 28, 2018; Accepted September 29, 2018

There are several features that might be ignored, so the analysis process becomes simpler, but still maintains the accuracy of the classification results.

Feature selection is not a new problem in the field of data mining. Several similar studies on feature selection have been carried out. Among them in Moustakidis's research, this study proposes a feature selection method called SVM-FuzCocs [4]. This method overcomes high-dimensional feature space with feature quality assessment based on fuzzy membership output from SVM. And the results show a satisfactory classification and reduction dimensions. In addition, this method has a fairly low computing requirement.

The research conducted by Dong applied t-test and p-value to educate the feature space [5]. And the results of this study indicate that with the application of these two features selection can increase the speed of the classification process without reducing the classification results. This proves that the use of feature selection is not only intended to improve classification performance but also reduces the classification computing load.

Hasan's research applied Principal Component Analysis (PCA) feature extraction and in more detail, three best algorithms from PCA, namely Screen Test, Cumulative Variance and KG rule, used as feature selection and Artificial Neural Network (ANNs) used as the classifier [6]. In this study, the best average classification accuracy achieved by the selection of Cumulative Variance features is 95.68%. This proves that the three best feature selection algorithms owned by PCA are able to improve the classification accuracy with the ANNs method.

Research conducted by Osareh applies several classification methods and feature selection methods including Support Vector Machines (SVM), K-nearest neighbors and probabilistic neural networks classifiers will be combined with signal-to-noise ratio feature ranking, and sequential forward selection as a selection feature. principal component analysis feature extraction [7]. The results of this study indicate an accuracy of achievement between 98.80% and 96.33% with SVM as the dominant classifier.

In this paper, we use the C5.0 algorithm, a classification algorithm that can be used to select features and form a rule set for early detection of the risk of pregnancy. The C5.0 method is chosen because this method can prune better than its predecessor methods such as ID3 and C4.5 [8]. Because the pruning process is better, this method will produce fewer rule sets, but the accuracy is better than ID3 or C4.5. This method also saves more memory compared to other decision tree methods [9]. The features that were successfully selected by the C5.0 method were quite good. In his research, Ojha, et al. [10] used C5.0 to select features to classify types of breast cancer. And the results obtained show that the features chosen by the C5.0 algorithm are able to obtain better accuracy when tested with other classification methods, such as KNN or Naïve Bayes. The main motivation of this paper is to show that the use of features with a small amount, but important, is better than a large number of features.

2. Research Method

2.1 Data Source

The data used in this study are primary data in the form of pregnant women card data as many as 300 data taken from Cipto Mulyo Health Center Malang from 2016 to 2017. This data was originally in the form of physical files, therefore the data was moved one by one manually into Command format Separated Value (CSV). While the features used are attributes that exist on pregnant women cards with a total of 107 features. These features are obtained from 4 types of checks.

1. History of previous pregnancy

   Examination of the history of pregnancy, childbirth and family planning needs to be done especially for women who are pregnant more than once. Because if there are complications in previous pregnancies, for example, a pregnant woman has a history of abortion or miscarriage in a previous pregnancy there will be indications that it can happen again in the current pregnancy. Some of the features obtained from this examination include a type of labor (abortion, premature, normal, cesarean), fetal condition (life, death), birth control, the frequency of pregnancy, etc.

2. Conditions of pregnancy now

   Examinations carried out in pregnancy now aim to find out if there are dangerous indications for the mother's pregnancy going forward. For example, if pregnant women suffer from hypertension during pregnancy at this time it will be risky for the mother to develop preeclampsia. Some of the features obtained from this examination include menstrual cycle, history of maternal illness, husband's illness history, family history, maternal habits, and so on

3. Physical Examination

   This examination is carried out to determine the physical condition of a pregnant woman and the condition of the fetus. Some of the features obtained from this examination include weight, hip shape, fetal location, body temperature, etc.

4. Laboratory Examination

   Laboratory tests are performed to determine hemoglobin levels and protein content in urine. Low hemoglobin levels can lead to anemia. While high urine protein content can be indicated preeclampsia. Some features obtained from this examination include hemoglobin level, urine albumin, urine reduction, etc.

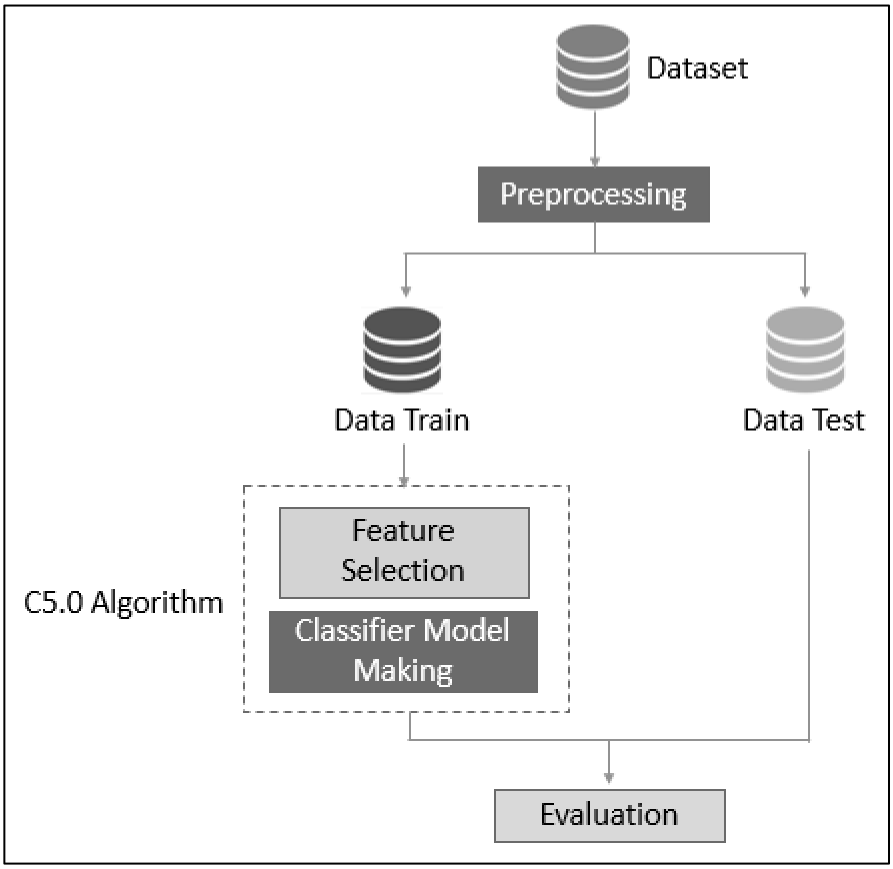

Figure 1. System Workflow

2.2 Data Preprocessing

As can be seen in Figure 1, the first step that needs to be done is preprocessing. Preprocessing data is a stage in data mining that is carried out to convert raw data into data that is ready to be processed. In this paper, we use 2 types of preprocessing data techniques.

1. Handling of missing values

   Of the 300 data obtained, there are several data that have a blank value on the feature. Features that have blank data include menstrual cycle, menstrual period, pre-pregnancy weight, height, weight, systolic blood pressure, diastolic blood pressure, upper arm circumference, temperature, pulse, breathing, gestational age, heart rate, hemoglobin, urine albumin, urine reduction, and blood type. This can be caused by the negligence of the medical officer to record the patient's complete data, or it could be because the examination is not necessary because the medical officer considers the condition of the pregnancy of the patient to be fine. To handle missing values, substitution methods are carried out. This method will fill in the data on an empty feature with an average value from other data in the same class. Table 1 and Table 2 show how the process of handling missing values in this study was conducted.

Table 1. Data with a Missing Value

Data	Height	Systolic	Diastolic	Class
Patient1	150	100	70	Low Risk
Patient2	151	110	?	High Risk
Patient3	146	90	70	Low Risk
Patient4	?	100	60	Low Risk
Patient5	153	120	80	High Risk
Patient6	160	?	70	High Risk

Table 2. Data after Preprocessing

Data	Height	Systolic	Diastolic	Class
Patient1	150	100	70	Low Risk
Patient2	151	110	75	High Risk
Patient3	146	90	70	Low Risk
Patient4	148	100	60	Low Risk
Patient5	153	120	80	High Risk
Patient6	160	115	70	High Risk

In Table 1, there are three data that have missing values, namely Patient2, Patient4, and Patient6. For example, to fill the height data in Patient2, it can be done by calculating the average height of Patient1 and Patient3. Why Patient1 and Patient3? Because both patients have the same class as Patient2, namely Low Risk.

2. Data Normalization

   The next preprocessing model that we do is data normalization. The purpose of this process is to balance the data by giving certain limits so that the data is in a certain range. The normalization method used is the z-score method.

2.3 Feature Selection and Classification Using C5.0 Method

The final result of this research is a classifier model in the form of a decision tree (decision tree). The shape of the tree is chosen as a model because this form can show the most dominating features for determining classification results. This dominant feature is called root. By knowing the most dominant features, the medical personnel or people around the patient can focus on observing this dominant feature. So that when a problem occurs which can cause an increased risk of pregnancy, it can be detected as early as possible. In addition, the tree-shaped model will automatically perform feature selection. Where non-influential features will not appear in the tree. Or if it appears, its position is usually near the leaf so that it is not too problematic if one day is done pruning. With this tree model, medical personnel simply ask questions in accordance with the sequence of features that are in the decision tree. Starting from the feature that is in the root position, then down to the successor feature that is below it.

To make the model the appropriate method is needed. The C5.0 algorithm is chosen because it is the development of the C4.5 algorithm. The C4.5 algorithm itself is one of the algorithms that is quite widely used to handle classification problems and proven reliability. C5.0 can classify data that has parameters with unknown values. C5.0 can also do better pruning compared to C4.5 and lower memory usage [9]. This is because C50 supports the boosting method that improves the trees and gives them more accuracy. In addition, one of the advantages of C5.0 is that this method can select features while classifying data properly. To apply the C5.0 method, we use R as a programming language. R was chosen because of its reliability to manage data with a large number of features and the availability of libraries to perform statistical operations [11].

3. Results and Discussion

3.1 Classification Evaluation

Of the 117 features contained in the dataset, C5.0 selects 20 features which are the most influential features in the process of classifying the risk of pregnancy. The twenty features of the biggest to smallest influence are SC, bleeding, UI, location of fetus at age> 36 weeks, Ab, Hb, Hypertension, Pulmonary Disease, IUFD, Tools, Herbs, TB, frequency of pregnancy, blood type, pregnancy pause, complications, nausea, vomiting, P / L, abdominal pain, and fetal motion.

The model produced is then tested using 50 data test data. Which consists of 20 data for the "Low Risk" class, 17 data for the "High Risk" class, and 13 data for the "Very High Risk". From the test results, obtained confusion matrix table as can be seen in Table 3.

Table 3. Confusion Matrix of the Model

		Prediction		
		Low Risk	High Risk	Very High Risk
Actual	Low Risk	19	1	0
	High Risk	2	14	1
	Very High Risk	0	2	11

From the confusion matrix table above, the accuracy rate or success rate of the resulting model is 0.88 or 88%. How to calculate accuracy is as seen Equation 1.

$$\text{accuracy} = \frac{\text{TP}_{\text{lowrisk}} + \text{TP}_{\text{highrisk}} + \text{TP}_{\text{veryhighrisk}}}{\text{total\_data}} \quad (1)$$

$$= \frac{19 + 14 + 11}{50} = 0.88$$

Where TP is True Positive, which is a value that shows how much data is classified correctly (according to the actual data).

3.2 Feature Selection Evaluation

To evaluate the C5.0 selection features, accuracy testing is performed using several different classification methods. For implementation, we use 250 data as data train, and 50 data as data test. Two test scenarios are used as follows:

1. The preprocessed dataset (still has 107 features), was tested using the SVM, Naïve Bayes, and Nearest Neighbor (NN) methods. 10-fold cross validation is also used to get consistent accuracy values.
2. The dataset with 20 selected features using C5.0 was tested using the SVM, Naïve Bayes, and Nearest Neighbor (NN) methods. 10-fold cross validation is also used to get consistent accuracy values.

The results of the two tests can be seen in Table 4 below.

Table 4. Accuracy Test Result

Method	Accuracy	
	107 features	20 features
SVM	0.91	0.91
Naïve-Bayes	0.89	0.88
Nearest Neighbor	0.80	0.85

From the table, it can be observed that the selection results feature using C5.0 is able to maintain the accuracy value when tested using several different methods, even though the number of features used is only 18% of the total features available. Even for the nearest neighbor method, the selection results are able to produce better accuracy values than if we use the whole feature.

4. Conclusion

From the test results, it can be concluded that the C5.0 algorithm can perform feature selection while classifying the risk of pregnancy with an accuracy rate of 88%. The feature selection carried out by C5.0 also proved to be quite good, and was able to increase the accuracy of results by up to 5% (in nearest neighbor method) when compared to the accuracy obtained when using the entire feature.

References

[1] T. Tejayanti, D. Bisara, and L. Pangaribuan, *“Penyebab kematian maternal di Kabupaten Malang Provinsi Jawa Timur tahun 2010,”* Jurnal Kesehatan Reproduksi, Vol. 6, No. 1, Pp. 1–9, 2015.

[2] S. Y. Utama, *“Faktor Risiko yang Berhubungan dengan Kejadian Preeklampsia Berat pada Ibu Hamil di RSD Raden Mattaher Jambi Tahun 2007,”* Jurnal Ilmiah Universitas Batanghari Jambi, Vol. 8, No. PREEKLAMSIA, Pp. 52–58, 2008.

[3] T. Afifah *et al.*, *“Maternal Death in Indonesia: Follow-Up Study of the 2010 Indonesia Population Census,”* Jurnal Kesehatan Reproduksi, Vol. 7, No. 1, Pp. 1–13, 2016.

[4] S. P. Moustakidis and J. B. Theocharis, *“SVM-Fuzcoc: A Novel SVM-Based Feature Selection Method Using a Fuzzy Complementary Criterion,”* in Pattern Recognition 43, Pp. 3712–3729, 2010.

[5] A. Dong and B. Wang, *“Feature Selection and Analysis on Mammogram Classification,”* in Communications, Computers and Signal Processing, Pp. 731–735, 2009.

[6] H. Hasan and N. M. Tahir, *“Feature Selection of Breast Cancer Based on Principal Component Analysis,”* in Signal Processing and Its Applications (CSPA), Pp. 1–4, 2010.

[7] A. Osareh and B. Shadgar, *“Machine Learning Techniques to Diagnose Breast Cancer,”* in In Health Informatics and Bioinformatics (HIBIT), Pp. 114–120, 2010.

[8] S. A. Selvi, S. Sowmiya, and R. Sangeetha, *“C5 Causal Decision Tree,”* International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Vol. 3, No. 3, Pp. 876–879, 2018.

[9] R. Pandya and J. Pandya, *“C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning,”* International Journal of Computer Applications, Vol. 117, No. 16, Pp. 18–21, 2015.

[10] U. Ojha, M. Jain, G. Jain, and R. K. Tiwari, *“Significance of Important Attributes for Decision Making Using C5.0,”* 8th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2017, Op. 3–6, 2017.

[11] R Core Team, *“R: A language and environment for statistical computing,”* URL http://www.R-project.org/.