International Journal of
Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: 10.31695/IJASRE.2020.33794

Volume 6, Issue 4
April - 2020

K-Means Bootstrap Analysis in the Birth Weight Classification of the Born Babies

I Ketut Putu Suniantara[1], Ni Putu Nanik Hendayanti[2], Gede Suwardika[3]

[1-2]Department of Information Systems

[1-2]Institute of Technology and Business STIKOM Bali

[3]Department of Statistics, Faculty of Mathematics and Sciences, Terbuka University

Denpasar, Bali

Indonesia

ABSTRACT

Low birth weight babies are caused by poor nutritional status before and during pregnancy. The impact of low birth weight is the slowing of the growth of infants seen in weight gain that does not reach normal levels when one-year-old. Classification of birth weight babies will be grouped into several groups, namely, low birth weight, normal baby weight, and excess baby weight. Classification of birth weight is done by cluster analysis. Cluster analysis is one of the statistical analyzes which aims to group objects based on the similarity of characteristics between those objects. Clusters of K-Means begin by determining in advance the number of clusters desired, so the measure of inaccuracy must be measured again. The inaccurate size of the K-Means cluster analysis can be done with the application of Clustering Bootstrap. This study aims to classify the birth weight of a baby using a cluster and a bootstrap cluster. The results showed a grouping of birthweight babies grouped into three clusters, cluster I had a baby's birth weight above 3130.171 kilograms, cluster II had a baby's weight under 3033.327 kilograms and cluster III had a baby's birth weight below 2299.994 kilograms. K-Means Bootstrap results produce relatively the same cluster. The best method is K-Means Bootstrap B = 75, with a total accuracy rate (TAR) of 0.7993 and a total error rate (TER) of 0.2007. K-Means Bootstrap can be used as an alternative way to determine the group of birth weight babies.

Key Words: *Cluster K-Means, K-Means Bootstrap, Classification, Weight Of Born Babies.*

1. INTRODUCTION

Infant Mortality Rate is the number of babies who die before reaching the age of 1 (one) year expressed in 1,000 live births in the same year. The age of the baby is a condition that is vulnerable to both illness and death. Infant Mortality Rate is an indicator that is usually used to determine the degree of public health. Therefore many health efforts have been made in reducing infant mortality. Infant Mortality Rate is one indicator that is sensitive to the availability, utilization, and quality of antenatal care and poor nutritional status [1-3].

One of the main causes of infant mortality is low birth weight babies or prematurity. Low birth weight babies are caused by poor nutritional status before and during pregnancy. The impact of low birth weight is the slowing of the growth of infants seen in weight gain that does not reach normal levels when one-year-old. Birth weight is influenced by two factors, namely internal and external factors. Internal factors consist of maternal, fetal, and uterine-placental factors. External factors consist of social and environmental factors. Maternal factors consist of maternal characteristics (age, parity, the distance of pregnancy, upper arm circumference, height, and nutritional status) and other supporting factors namely gestational age, weight gain, antenatal care (ANC), Hemoglobin (Hb), Fe supplementation , blood pressure, education level, and pregnancy visit [4].

Cluster Analysis was first used by Tyron in 1939. Cluster Analysis aims to allocate groups of individuals to independent groups so that individuals in the same group are similar to each other, whereas individuals in different groups are not similar. In grouping it uses a measure that can explain the similarity or closeness between data to explain a simple group structure of complex data, that is a measure of distance or similarity and a measure of distance that is often used is a measure of distance called the euclidean distance [5-7].

The use of cluster analysis especially K-Means with distance measurement techniques only provides one solution in its solution, which is based on the distance technique used. So the statistical significance value and accuracy measurement in a grouping is not found with this classical method. To improve the significance and measurement of accuracy in grouping a method is needed so that the solution provided is no longer one. One of the grouping improvements is using the K-Means Cluster Bootstrap method.

Multiscale bootstrap approach to provide a measure of uncertainty in the classical hierarchical clustering method. This method works with the bootstrap resampling approach for each group (cluster). So that the validation measure obtained is the p-values informing groups (clusters) that have similarities with one another [8]. There are two p-values in the bootstrap resampling approach, namely probability bootstrap (BP) value, and approximately unbiased (AU) p-values. Multiscale bootstrap resampling is used to calculate the Approximately unbiased (AU) p-values, which provide a better estimate of bias resolution. Some research on classification with the bootstrap approach includes [9] introducing Approximately unbiased (AU) p-values in gene classification. [10] uses a bootstrap approach for classification or grouping of microarray data, which is used to measure grouping reliability.

In addition to the above research, other research related to K-Means bootstrapping was also carried out by [11] with the multiscale bootstrap approach in clustering hierarchical cluster analysis, which is said to be able to provide a measure of uncertainty in grouping that is similar to one another with Approximately unbiased (AU) p-values more or equal to 95%. Whereas conducted by [12] in increasing the accuracy of K-Means with the K-Means bootstrap method in grouping the nutritional status of children under 5 years of age results that the K-Means bootstrap with B = 75 is able to make the grouping accuracy better.

Based on the things that have been explained above, the formulation of the problem in this study is how to apply the K-Means and K-Means Bootstrap method to the case of a baby's birth weight? and how do you compare the accuracy of the K-Means classification with the K-Means Bootstrap?

2. METHOD

The data used in this study are secondary data in the form of medical records of babies born. The research variable used is the dependent variable of birth weight. The independent variables are the age of pregnant women, gestational age, the order of pregnancy, type of birth and length of education of pregnant women [13]. The steps to achieve the research objectives are as follows:

a. The weight grouping of babies born with cluster analysis
b. Determine the baby's birth weight for each group as a result of cluster analysis, group determination is based on the characteristics of the members of each group.
c. Bootstrap clusters for 25, 50, 75 and 100 bootstrap replication to get misclassified e_B.
d. Determine the accuracy of bagging classification.
e. Compare the results of bagging classification with the accuracy of the classification of a single model.

3. ANALYSIS OF RESULTS

2.1 General Description of Data

The object of observation in this study was patients giving birth from 2015 - 2017. In this study, there were independent variables with nominal, ordinal, and continuous scales, while the dependent variables were continuous scale. The number of samples taken was 202 samples. Most patients were patients with normal birth types of 70.3% (142 samples) and the lowest was 29.7% (60 samples) with cesarean section (Figure 1).

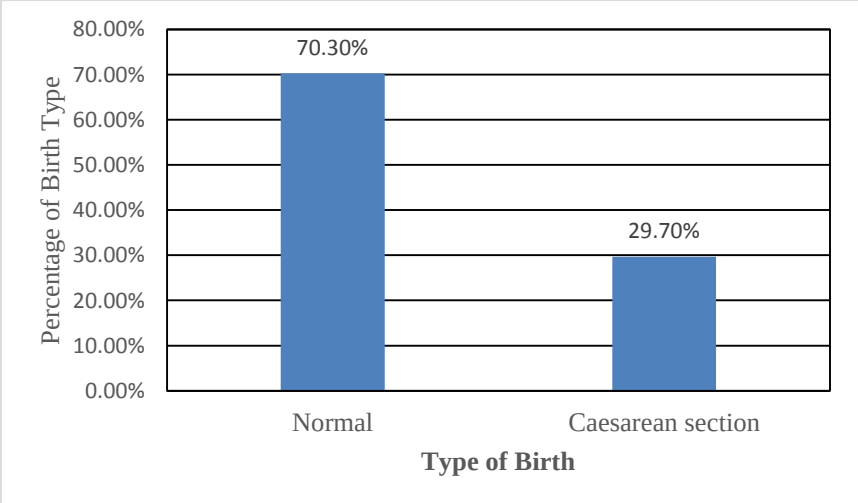

Figure 1. Birth Type Category

DOI: 10.31695/IJASRE.2020.33794

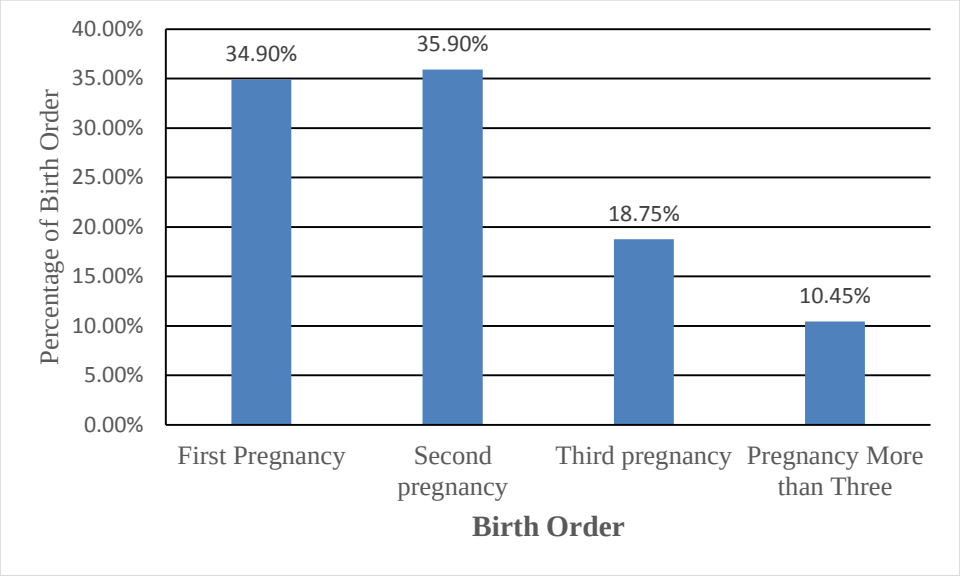

Figure 2. Histogram of Birth Order

Based on Figure 2, it can be seen that the highest number of patients in the second pregnancy is 35.90%. Furthermore, patients with the first pregnancy sequence were 34.90%, the third pregnancy sequence was 18.75%, the pregnancy order was more than three, 10.45%.

Table 1. Size of Centering Data

	Minimum	Maximum	Mean	Std. Deviation
Birth Weight	1000	4000	3092.57	470.782
Age of Pregnant Women	17	45	28.34	5.819
Gestational Age	29	43	39.59	1.569
Length of Education of Pregnant Women	0	16.5	11.406	2.0415

Sources: Primary data is processed, 2020

Based on Table 1, it can be seen that the age of the mother has a mean value of 28.34 years with a standard deviation of 5.819 years. The highest maternal age is 45 years and the lowest is 17 years. Gestational age has an average value of 39.59 weeks with a standard deviation of 1.569 weeks. The highest gestational age is 43 years and the lowest is 29 years. Based on the length of education, the average length of education of a mother is 11.406 years with a standard deviation of 2.0415 years. The highest education is 16.5 years (undergraduate) and the lowest is 0 years (never received formal education). In the baby birth weight variable (in grams) which is measured when the newborn is born, it appears that the birth weight of the baby has a mean value of 3,092.57 grams with a standard deviation of 470.782 grams. The highest baby's weight was recorded at 4000 grams and the lowest was 1000 grams.

2.2 K-Means Cluster Analysis

Before cluster analysis is carried out, it is necessary to standardize the variables because there are significant unit differences between the variables studied. The results of Clusters with K-Means formed with 3 clusters with the clustering process carried out as many as 9 iterations. The minimum distance between cluster centers that occurs from the iteration results is 7.339. The final results of the clustering process can be seen in Table 2.

DOI: 10.31695/IJASRE.2020.33794

Table 2. Cluster Center

Variable	Cluster		
	1	2	3
Birth Weight	0.080	-0.126	-1.684
Age of Pregnant Women	-0.188	1.151	0.113
Gestational Age	0.156	0.073	-4.730
The Order of Pregnancy	-0.278	1.637	0.474
Type of Birth	-0.047	0.153	0.641
Length of Education of Pregnant Women	0.239	-1.505	0.046

Sources: Primary data is processed, 2020

The cluster center is still related to the previous data standardization process above which refers to the z-score with the provisions 1) a negative value (-) means the data is below the total average and 2) a positive value (+) means the data is above the total average. Calculation of cluster average values as follows:

$$\bar{X} = \pi + \sigma \times Z$$

$$\bar{X} = 3092.57 + 470.782 \times 0.080 = 3130.171$$

The results of calculating the average value of each variable and cluster can be seen in table 3.

Table 3. Average Value of Each Cluster

Variable	Cluster		
	1	2	3
Birth Weight	3130.171	3033.327	2299.994
Age of Pregnant Women	27.247	35.036	28.998
Gestational Age	39.835	39.705	32.168
The Order of Pregnancy	1.801	3.996	2.663
Type of Birth	0.688	0.780	1.002
Length of Education of Pregnant Women	11.894	8.333	11.500

Sources: Primary data is processed, 2020

Based on table 3, it is known that cluster 1 is a cluster with birthweight with a value above 3.1 kg, gestational age above 39.83 weeks, maternal age under 27.274 years with the length of education above 11.895 years. In cluster 2, the baby's body weight was born under 3.03 kg with a gestational age above 39.7 weeks and normal birth. Whereas in cluster 3, the baby's birth weight is below the average of 2.3 kg with a gestational age of fewer than 32.12 weeks with a cesarean birth. The division of clusters is explained in the following description:

1. Cluster – 1
   Cluster-1 contains objects that have a baby's birth weight, gestational age, length of education, birth order and type of birth that are more than the average population of the object under study. Thus, it can be assumed that Cluster-1 is a grouping of birth weight babies that have a large average value or can be categorized as overweight babies born.
2. Cluster – 2
   Characteristics of birth weight babies included in the Cluster-2 grouping are all variables having an average birth weight of babies born less than the total average, with a length of education under 8 years. Thus, it can be assumed that a set of intermediate objects are in Cluster-2.
3. Cluster – 3
   While the characteristics of objects that are clustered in cluster-3 are the weight variables of babies born under 2.3 kg with gestational age under 32 weeks. In this cluster, it is thought to be a low birth weight baby group.

Differences in the variables in each cluster formed were tested by analysis of variance (Anova). Analysis of the variables in the formed cluster can be seen in Table 4.

DOI: 10.31695/IJASRE.2020.33794

Table 4. ANOVA Analysis of the Best Clusters

Variable	Cluster		Error		F	Sig
	Mean Square	df	Mean Square	Df		
Birth Weight	9.256	2	0.917	199	10.093	0.000
Age of Pregnant Women	20.895	2	0.800	199	26.117	0.000
Gestational Age	69.263	2	0.314	199	220.620	0.000
The Order of Pregnancy	43.419	2	0.574	199	75.687	0.000
Type of Birth	1.737	2	0.993	199	1.750	0.176
Length of Education of Pregnant Women	35.407	2	0.654	199	54.122	0.000

Sources: Primary data is processed, 2020

Based on Table 4, it appears that all variables are significant except the variable birth type which is not significant. This means that the variables of birth weight, age of pregnant women, gestational age, the order of pregnancy and length of education show differences in cluster formation. Cluster 1 consisted of 169 samples, cluster 2 consisted of 27 samples and cluster 3 consisted of 6 samples.

Table 5. Accuracy in K-Means Classification

		Actual Cluster			Total
		1	2	3	
Prediction Cluster	1	145	20	4	169
	2	22	3	2	27
	3	2	4	0	6
Total		169	27	6	202

Sources: Primary data is processed, 2020

In table 5, it is used to measure classification accuracy by calculating the value of Total Accuracy Rate (TAR) and Total Error Rate (TER). The calculation results show that the TAR is 0.7327 or in other words, the right sample data is classified as a whole as much as 73.27% and misclassified by 26.73%.

$$TAR = \frac{145 + 3 + 0}{202} = 0.7327$$

$$TER = \frac{22 + 2 + 20 + 4 + 4 + 2}{202} = 0.2673$$

2.3 K-Means Bootstrap Analysis

Bootstrap is a sampling technique with the return of an original sample, which aims to obtain parameter estimates based on minimal data with the help of a computer. Classification with K-Means bootstrap method is done by resampling 25 times, 50 times, 75 times and 100 times. The K-Means bootstrap classification results can be seen in the following Table 6:

Table 6. K-Means Bootstrap Classification

	Boostrap 25			Boostrap 50		
	1	2	3	1	2	3
Cluster members	167	29	6	167	29	6
Birth Weight	3059.92	2963.08	2229.74	3059.92	2963.08	2229.74
Age of Pregnant Women	24.00	31.79	25.75	24.00	31.79	25.75
Gestational Age	38.06	37.92	30.39	38.06	37.92	30.39
The Order of Pregnancy	1.46	3.66	2.32	1.46	3.66	2.32
Type of Birth	0.56	0.65	0.88	0.56	0.65	0.88
Length of Education of Pregnant Women	10.63	7.07	10.24	10.63	7.07	10.24
	Boostrap 75			Boostrap 100		
	1	2	3	1	2	3

DOI: 10.31695/IJASRE.2020.33794

Cluster members	168	28	6	169	27	6
Birth Weight	3128.92	3032.08	2298.74	3128.83	3031.98	2298.65
Age of Pregnant Women	26.48	34.27	28.23	26.46	34.25	28.21
Gestational Age	39.12	38.99	31.45	39.12	38.99	31.45
The Order of Pregnancy	1.24	3.43	2.10	1.24	3.43	2.10
Type of Birth	0.61	0.71	0.93	0.61	0.71	0.93
Length of Education of Pregnant Women	11.63	8.07	11.24	11.63	8.07	11.24

Sources: Primary data is processed, 2020

Based on Table 6, the results of grouping infant weight using the K-Means Bootstrap Method B = 25, B = 50, B = 75 and B = 100 showed relatively similar results with uniform cluster dispersion rates. The number of members in the cluster also shows relatively the same conditions, with the most cluster members in cluster 1. This condition shows the number of Bootstrap replication is not different so that replication 25, 50, 75 and 100 will show the same condition. The classification accuracy with the K-Means Bootstrap Method can be seen in Table 7.

Table 7. The Accuracy of K-Means Bootstrap Method Classification

Indicator	**Classification Method**			
	Bootstrap B = 25	**Bootstrap B = 50**	**Bootstrap B = 75**	**Bootstrap B = 100**
Total Accuracy Rate	0.7327	0.7819	0.7993	0.7992
Total Error Rate	0.2673	0.2181	0.2007	0.2008

Sources: Primary data is processed, 2020

TAR value goes up, shows better results. While the TER value is increasing, it shows that the results are getting worse. The greater the bootstrap resampling, the greater the TAR value and the smaller the TER value. This shows that the results are getting better. Based on Table 7, shows that the most optimal classification accuracy using the K-Means Method with Bootstrap 75 times. Because the method shows the highest TAR value or accuracy with the lowest error value

3. CONCLUSION AND SUGGESTION

Based on the results of the analysis of the birth weight classification of babies born with K-Means and K-Means Bootstrap clusters, it can be concluded as follows:

a. Results of the Application of K-Means Clusters on birth weight of babies resulted in Clustering of baby's weight grouped into 3 clusters with a classification accuracy of 73.27%, consisting of cluster I with infant weight above 3130.171 kg, cluster II with body weight babies born under 3033,327 kg and cluster III with birth weight babies weighing under 2299,994 kg.
b. K-Means Bootstrap results produce clusters that are relatively similar to the uniform distribution of clusters with each accuracy in the form of TAR and TER values for B = 25 namely 73.27% and 26.73%, B = 50 is 78.19% and 21.81%, B = 75, 79.93% and 20.07% and B = 100, 79.92% and 20.08%.
c. The best method on K-Means Bootstrap B = 75 times with a total accuracy rate (TAR) of 0.7993 and a total error rate (TER) of 0.2007 so that it can be said K-Means Bootstrap can improve the accuracy of classification.

Suggestions that can be given in this study are to add other independent variables to get better classification results, and research that uses Bagging Methods, bootstrap replication can be done to get better improvements and can be used other methods for classification problems.

REFERENCES

[1] I. U. Tarigan, T. Afifah, and D. Simbolon, "Faktor-faktor yang berhubungan dengan pelayanan bayi di indonesia: pendekatan analisis multilevel," *J. Kesehat. Reproduksi*, vol. 8, no. 1, pp. 103–118, 2017.

[2] Abdiana, "Determinan Kematian Bayi di Kota Payakumbuh," *J. Kesehat. Masy. Andalas*, vol. 9, no. 2, pp. 88–92.

[3] A. P. Rachmadiani, M. A. Shodikin, and C. Komariah, "Faktor-Faktor Risiko Kematian Bayi Usia 0-28 Hari di RSD dr. Soebandi Kabupaten Jember," *J. Agromedicine Med. Sci.*, vol. 4, no. 2, pp. 60–65, 2018.

DOI: 10.31695/IJASRE.2020.33794

[4] T. Hollingworth, *Differential Diagnosis in Obstetrics and Gynecology*. Great Britain: Edward Arnold, 2008.

[5] R. A. Johnson and D. . Winchern, *Applied Multivariate Statistical Analysis*. USA: Prentice Hall. Inc, 2007.

[6] B. F. . Manly, *Multivariate Statistical Methods: A primer, Third edition*. Chapman and Hall, 2005.

[7] N. H. Timm, *Applied Multivariate Analysis*. New York: Springer, 2002.

[8] R. Suzuki and Hidetoshi Shimodaira, "An application of multiscale bootstrap resampling to hierarchical clustering of microarray: How accurate are these clusters," in *In proceedings by the fifteenth International Conference on Genome Informatics*, 2004, p. 34.

[9] H. Shimodaira, "Approximately unbiased tests of regions using multistepmultiscale bootstrap resampling," *Ann. Stat.*, vol. 32, no. 6, pp. 2616–2641, 2004.

[10] M. K. Kerr and G. A. Churchill, "Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments," in *Proceedings of the National Academy of Sciences of the United States of America*, 2001, pp. 8961–8965.

[11] G. Anuraga., "Hierarchical Clustering Multiscale Bootstrap Untuk Pengelompokan Kemiskinan Di Jawa Timur," *J. Stat.*, vol. 3, no. 1, pp. 27–33, 2015.

[12] H. Prasetyo, Kuntoro, W. Purnomo, Soenarnatalina, M. Adriani, and Bambang Wijanarko, "Penerapan Clustering Bootstrap dengan Metode K-Means," *J. Biometrika dan Kependud.*, vol. 3, no. 1, pp. 43–49, 2014.

[13] I. K. P. Suniantara, I. G. E. W. Putra, and G. Suwardika, "Peningkatan Ketepatan Klasifikasi dengan Metode Bootstrap Aggregating pada Regresi Logistik Ordinal," *INTENSIF J. Ilm. Penelit. dan Penerapan Teknol. Sist. Inf.*, vol. 3, no. 1, pp. 32–42, 2019.

DOI: 10.31695/IJASRE.2020.33794