Data Mining Performance of Toddler Nutrition Classification Based on Family Nutrition Awareness and Human Development Index

Darmansyah, Gede Putra Kusuma

Abstract: ***Nutrition problems that occurred in districts/cities of Central Java province from 2015-2017 were only 1 district city that did not have nutritional problems (good category) in 2015.The rest had acute, chronic or acute chronic nutrition problems. The search for the most influential attributes in toddler nutrition problems using data mining is expected to help health workers to focus more on solving problems based on classification in the area.Therefore, improving the nutritional status of the community can be accelerated. The best parameter search from the selection of features and data mining algorithm using the Optimize Parameters (Grid) operator found in Rapidminer.The feature selection models used are Backward Elimination, Forward Selection, and Optimize Selection. The datamining algorithm used is Naive Bayes, Decision Tree, k-NN, and Neural Network.The merging of the feature selection model and the datamining algorithm resulted in 12 algorithm models used in this study.The best model that was processed using test data with the highest accuracy of 74.19% was obtained from backward-neural network elimination. The attribute that is not very influential based on the model obtained is the condition of the mother who died.***

Keywords**: *Feature Selection, Data Mining, Parameter Tuning.***

I. INTRODUCTION

Toddler is a substitute candidate that must be considered because it is an age group that is vulnerable to nutrition and disease. Cases of malnutrition in infants not only cause physical disorders but also affect intelligence and productivity as adults [1]. Based on the 2017 Nutrition Status Monitoring (NSM) conducted by the Ministry of Health, infants under five years old (toddler) who experienced nutritional problems in 2017 reached 17.8%, the same as the previous year. The amount consists of toddlers experiencing 3.8% malnutrition and 14% malnutrition. Based on Fig. 1 of 514 districts/cities, there are 6 districts/cities with good categories/low nutritional problems [2].

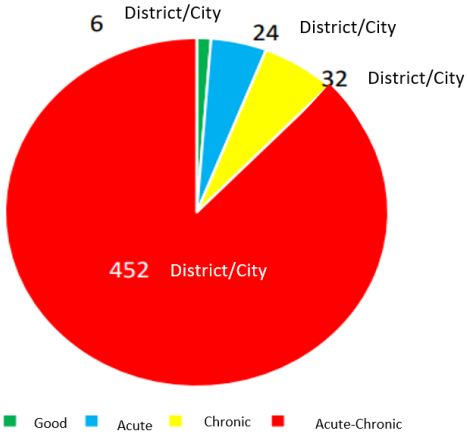
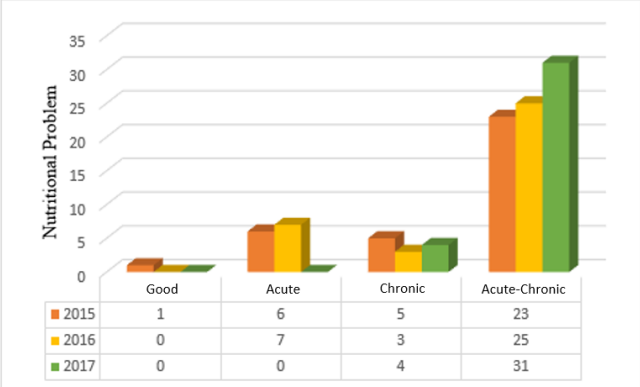


Fig. 1. Distribution of Nutritional Problems of City District 2017 [3]

Central Java Province is one of the provinces that has the largest population in Indonesia after West Java and East Java. As shown inFig. 2, nutrition problems in the province of Central Java in 2017 have increased compared to 2015.A clear change can be seen in acute nutrition problems,which did not exist in 2017. Meanwhile chronic acute nutrition problems in 2015 has increased from 23 districts cities in 2017 to 31 districts/cities experiencing chronic acute nutritional problems [4, 4].



	Good	Acute	Chronic	Acute-Chronic
2015	1	6	5	23
2016	0	7	3	25
2017	0	0	4	31

Fig. 2. Nutrition Problems 2015-2017 Districts/Cities of Central Java Province [3] [5] [6]

The issue of nutritional status is influenced by several direct and indirect factors. The causative factors are directly affected by children's food and infectious diseases, while the causative factors are indirectly affected by food security in the family, child care patterns, health care, environmental health, parents' education level, income level, type occupation, level of knowledge, number of family

Revised Manuscript Received on January 15, 2020
***** Correspondence Author
Darmansyah, Computer Science Department, BINUS Graduate Program - Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480. Email: darmansyah@binus.ac.id
Gede Putra Kusuma*, Computer Science Department, BINUS Graduate Program - Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480. Email: inegara@binus.edu

members and social culture. [7]

The health office as an extension of the government in Central Java Province in handling nutrition problems in the region, in general every year in determining nutrition problems in Central Java Province based on data obtained by health workers in the area and recorded in the provincial health profile. The provincial health profile book includes several data, namely demographic conditions, health facilities, health human resources, health financing, family health, disease control and environmental health [4]. This study will look for the most influential attributes in toddler nutrition problems using data mining. It is hoped that from the attributes obtained from this study, health workers can focus more on solving problems based on classification in the area so that the improvement of people's nutritional status can be accelerated. Based on this background, the researchers consider the need for research to classify nutrition for toddlers based on family nutrition awareness and human development index.

II. RESEARCHMETHODOLOGY

This study consists of 6 steps as shown in Fig.3.

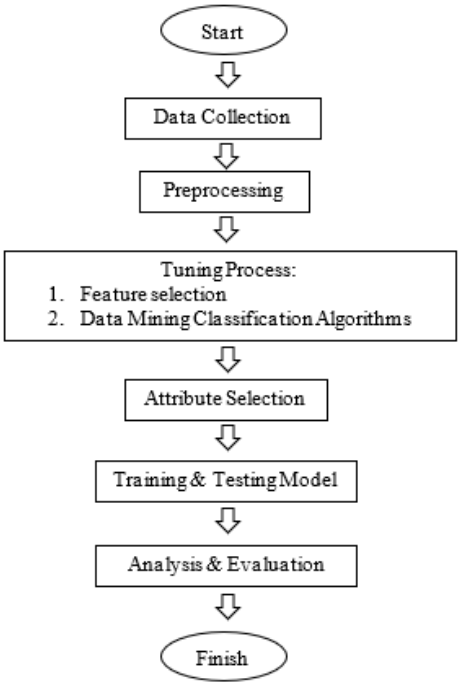


Fig. 3. The Research Methodology

A. Data Collection

The data used in this study came from the Indonesian Center for Statistics and Pusdatin Ministry of Health. The data obtained are the health data of the province of Central Java in 2015 and 2018 in the form of data that affect the nutritional status of children under five, both directly and indirectly. The data obtained is divided into two parts, namely training data and testing data. Training data is obtained from data for 2015 & 2016. While testing data uses data from 2017 & 2018.

B. Preprocessing

The data obtained is integrated to produce data attributes in accordance with research needs as shown in Table 1. The ID used is the ID of the district/city and nutrition problem labels as in Table 2.

Table 1. Data Attributes

No	Attributes	Data Type	Data Source
1	Mother Passed Away	Numeric	Health Profile of Central Java Province
2	TB Children 0-14 years old	Numeric	Health Profile of Central Java Province
3	Toddler pneumonia	Numeric	Health Profile of Central Java Province
4	HIV	Numeric	Health Profile of Central Java Province
5	Vit A Mother	Numeric	Health Profile of Central Java Province
6	FE 1 Pregnant Mother	Numeric	Monitoring the Nutrition Status of Pusdatin Ministry of Health
7	FE 3 Pregnant Mother	Numeric	Monitoring the Nutrition Status of Pusdatin Ministry of Health
8	BBLR	Numeric	Health Profile of Central Java Province
9	ASI Exclusive	Numeric	Monitoring the Nutrition Status of Pusdatin Ministry of Health
10	Immunization Hepatitis B	Numeric	Health Profile of Central Java Province
11	Immunization Tuberculosis	Numeric	Health Profile of Central Java Province
12	Complete Basic Immunization	Numeric	Health Profile of Central Java Province
13	Immunization Polio	Numeric	Health Profile of Central Java Province
14	Immunization Measles	Numeric	Health Profile of Central Java Province
15	Vit A 0-59 Months	Numeric	Monitoring the Nutrition Status of Pusdatin Ministry of Health
16	Toddler Health Services (min 8 times)	Numeric	Health Profile of Central Java Province
17	Households with Clean Behaviour	Numeric	Health Profile of Central Java Province
18	Healthy Home	Numeric	Health Profile of Central Java Province
19	Access to Proper Drinking Water	Numeric	Health Profile of Central Java Province
20	Decent Drinking Water Provider	Numeric	Health Profile of Central Java Province
21	Healthy Latrine	Numeric	Health Profile of Central Java Province
22	Healthy Public Places	Numeric	Health Profile of Central Java Province
23	Hygienic Food Management Place	Numeric	Health Profile of Central Java Province
24	Active Posyandu	Numeric	Health Profile of Central Java Province
25	Poor Population	Numeric	Central Bureau of Statistics
26	Life Expectancy at Birth	Numeric	Central Bureau of Statistics
27	Average length of school	Numeric	Central Bureau of Statistics
28	Per capita expenditure adjusted	Numeric	Central Bureau of Statistics
29	Old School Hopes	Numeric	Central Bureau of Statistics

Table 2. Data Label

Data Type: Polynomial	
No	Nutrition Problems
1	Good
2	Acute
3	Chronic
4	Acute-Chronic

C. Tuning Process

At this stage the best model of feature selection and data mining algorithm will be searched using the Optimize Parameters (Grid) operator found in the Rapidminer [8]. The parameters to be used can be seen in Table 3 [9].

Table 3. Parameters Tuning

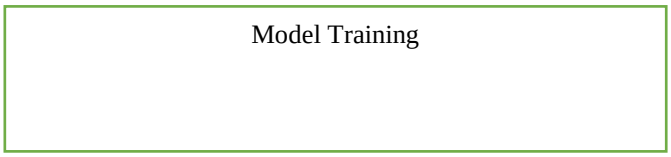
	Model	Parameters
Feature Selection	Backward Elimination	1. Maximal_number_of_eliminations 2. Speculative_rounds
	Forward Selection	1. Maximal_number_of_attributes 2. Speculative_rounds
	Optimize Selection	1. Selection_direction
Algorithm data mining	Naive Bayes	1. Laplace_correction
	Decision Tree	1. Criterion 2. Maximal_depth 3. Confidence 4. Minimal_gain 5. Minimal_leaf_size
	k-NN	1. k 2. Weighted_vote 3. Measure_types
	Neural Net	1. Momentum 2. Learning_rate

D. Attribute Selection

Attribute selection functions to find the most influential attributes of 29 attributes used [10]. At this stage, feature selection and data mining algorithms will have the highest performance from the previous tuning results.

E. Training & Testing Model

After getting the most influential attributes, the training & testing process is carried out using the data that has been prepared. The training model uses 2015 data and 2016 data. The attributes of existing data are filtered based on the results of feature selection. After that, it is normalized with the transformation method range 0 - 1. Normalization results are entered into the data laminating algorithm from the results of the tuning process. The results of the training model are calculated for accuracy of performance and kappa. The difference between testing and training models is that the testing model uses 2017 data to see performance and 2018 for the next classification prediction. The data mining algorithm model of the training process is used for model testing. For more details, the models can be seen in Fig 4.



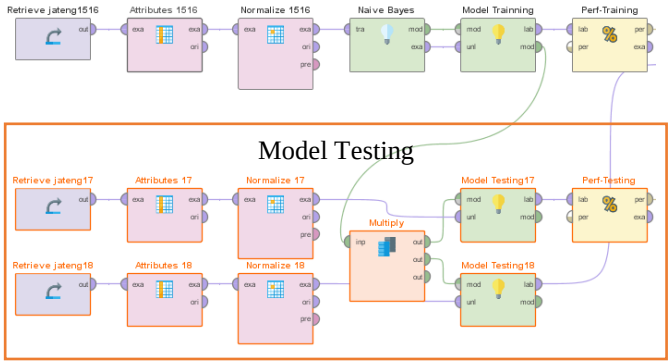


Fig. 4. Training & Testing Model

F. Analysis & Evaluation

The next step to be carried out is to evaluate the results obtained with the Performance Matrix. At this stage the evaluation table will be obtained from the data mining classification as shown in Table 4. The Performance Matrix contains the calculation of the accuracy of the predictive data test object being the actual class, either predicted to be positive or negative.

Table 4. PerformanceMatrix

Accuracy: %			
	True 1	True 2	Class Precision
Prediction 1	TP	FP	%
Prediction 1	FN	TN	%
Class Recall	%	%	

Explanation:

a) TP = True Positive is positive data that is detected correctly
b) FP = False Positive is negative data but detected as positive data
c) FN = False Negative are positive data that is detected incorrectly as negative
d) TN = True Negative is the amount of negative data that is detected correctly

Precision is the percentage comparison of true positive predictions with the results of overall positive predictions. Precision is chosen if wechoose True Positive and really does not want False Positive to occur.

$$Precision = \frac{TP}{(TP + FP)} \; x\; 100\% \qquad (1)$$

Recall is a percentage comparison of true positive predictions with overall positive true data. Recall is selected if having False Positive is better than False Negative.

$$Recall = \frac{TP}{(TP + FN)} \; x\; 100\% \qquad (2)$$

Specificity is the comparison of the percentage of negative predictive truth with overall negative data. Specificity is chosen if you don't want false positives.

$$\text{Specificity} = \frac{TN}{(TN + FP)} \; x\; 100\% \qquad (3)$$

Accuracy is a percentage of the total data identified and assessed. Accuracy is used if the dataset has symmetrical False Negative and Positive False data numbers.

$$Accuracy = \frac{TP + TN}{(TP + +TN + FP + FN)} \; x\, 100\% \qquad (3)$$

F1 score is a comparison of the percentage of precision of weighted average and recall. F1 scores are used if the dataset has a number of False Negatives and False Positive [11].

$$F1\ Score = 2\, x\, \frac{Recall \;\; x\, Precisio\, n}{Recall \;\; + Precision} \; x\, 100\% \qquad (4)$$

III. RESULT AND ANALYSIS

The number of districts/cities in Central Java province is 35. The total data used as training data is 70 and that used as testing data is 35.

Based on the feature selection and data mining algorithm settings as in Tables 5 - 6, there are 7 models that have 100% accuracy, namely the neural network on each feature selection, decision tree and k-NN in the forward selection and Optimize Selection. The model that uses the longest time is the neural network in backward elimination, which is 54658 seconds (15 hours 10 minutes 58 seconds).The shortest time is k-NN at 10 seconds.

Table 5. Tuning Result of Optimize Selection

	Optimize Selection							
	Naive Bayes		**Decision Tree**		**k-NN**		**Neural Net**	
Time	< 1 s		2 h 1 m 37 s		10 s		1 h 42 m 45 s	
Accuracy	81.43%		100.00%		100.00%		100.00%	
Kappa	0.601		1		1		1	
Selection_direction	backward		backward		forward		backward	
Parameter Model	laplace_ correction	TRUE	criterion	gini_index	k	1	momentum	0.1
			max_depth	11	weighted_ vote	TRUE	learning_rate	0.2
			confidence	0.1				
			min_gain	0.1	measure_ types	Mixed Measures		
			min_leaf_size	1				

Table 6. Tuning Result of Forward Selection

	Forward Selection							
	Naive Bayes		**Decision Tree**		**k-NN**		**Neural Net**	
Time	32 s		4 h 13 m 12 s		29 m 39 s		7 h 13 m 4 s	
Accuracy	82.86%		100.00%		100.00%		100.00%	
Kappa	0.588		1		1		1	
Max_number_of_att	8		11		1		32	
Speculative_rounds	2		7		0		4	
Parameter Model	laplace_ correction	TRUE	criterion	info_gain	k	1	momentum	0.1
			max_depth	11	weighted_ vote	TRUE	learning_rate	0.2
			confidence	0.1				
			min_gain	0.1	measure_ types	Mixed Measures		
			min_leaf_size	1				

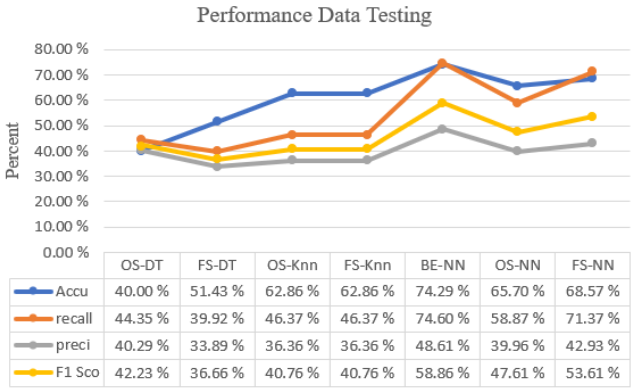
Table 7. Tuning Result of Backward Elimination

	Backward Elimination							
	Naive Bayes		**Decision Tree**		**k-NN**		**Neural Net**	
Time	39 s		2 h 7 m 31 s		45 m 42 s		15 h 10 m 58 s	
Accuracy	82.86%		82.86%		77.14%		100.00%	
Kappa	0.613		0.652		0.36		1	
Max_num_of_elim	5		11		26		11	
Speculative_rounds	3		7		10		2	
Parameter Model	laplace_ correction	TRUE	criterion	info_gain	k	1	momentum	0.5
			max_depth	11	weighted_ vote	TRUE	learning_rate	0.1
			confidence	0.1				
			min_gain	0.1	measure_ types	Mixed Measures		
			min_leaf_size	7				

Training data models (data from 2015-2016) that reach 100% accuracy and kappa 1 value are used to process test data (data from 2017) and produce Fig 5. Although all training models used have 100% F1 accuracy and scores, the test is done with the highest testing accuracy 74.29% in backward elimination - neural network models using 28 selected attributes. Accuracy of at least 40% in the Optimize Selection - Decision Tree model by using 27 selected attributes.

The attribute that is considered to have no effect on backward elimination - the neural network model is the condition of the mother dying when she was born.Because when the biological mother dies, the baby can still grow up healthy and get health care and education from the baby guardian.

In 2017, the test data used is different from the 2015-2016 data.There are no districts/cities that have acute and good nutrition problems. Based on Table 8, the model obtained is able to predict chronic acute nutritional problems as many as 23 out of 31 chronic acute problems or in the percentage of 74.19%. It able to predict chronic nutritional problems in 3 districts/cities out of 4 districts/cities that have chronic nutritional problems.



	OS-DT	FS-DT	OS-Knn	FS-Knn	BE-NN	OS-NN	FS-NN
Accu	40.00 %	51.43 %	62.86 %	62.86 %	74.29 %	65.70 %	68.57 %
recall	44.35 %	39.92 %	46.37 %	46.37 %	74.60 %	58.87 %	71.37 %
preci	40.29 %	33.89 %	36.36 %	36.36 %	48.61 %	39.96 %	42.93 %
F1 Sco	42.23 %	36.66 %	40.76 %	40.76 %	58.86 %	47.61 %	53.61 %

Fig. 5. Performance Data Testing

The correct prediction for acute-chronic nutrition problems is 23 out of the 24 predicted data or around 95.83%. The successful prediction of chronic nutrition problems is 3 out of 6 data or by 50%. For acute nutrition problems, the data model fails to predict, which should not be an acute nutritional problem but the model predicts there are 5 acute data.

Table 8. Result Model Testing Backward Elimination – Neural Network

		True				
		Acute-Chronic	Chronic	Acute	Good	**Precision**
Prediction	Acute-Chronic	23	1	0	0	95.83
	Chronic	3	3	0	0	50.00
	Acute	5	0	0	0	0.00
	Good	0	0	0	0	-
	Recall	74.19	75.00	-	-	

Table 9. Result Model Training

		True				
		Acute-Chronic	Chronic	Acute	Good	**Precision**
Prediction	Acute-Chronic	48	0	0	0	100%
	Chronic	0	13	0	0	100%
	Acute	0	0	8	0	100%
	Good	0	0	0	1	100%
	Recall	100%	100%	100%	100%	

IV. CONCLUSION

Finding the best parameters using the tuning method of the Optimize Parameters (Grid) operator found in Rapidminer is the best way to get the best performance. The training data used in this study are acute-chronic (48 data), chronic (13 data), acute (8 data) and good (1 data). From 12 training data models that were processed, 7 models produced 100% accuracy from the training data. The best model that is processed using test data with the highest accuracy of 74.19% is obtained from backward elimination - neural network. The attribute that is not too influential based on the model obtained is the condition of the deceased mother.

REFERENCES

1. K. Pibriyanti and T. P. Kurniawan, "Studi Status Gizi Bayi Usia 6-12 Bulan Di Desa Kradenan Kecamatan Trucuk Kabupaten Klaten Tahun 2017," Jurnal Kesehatan, p. 10, 2017.
2. Direktorat Gizi Masyarakat, "Hasil Pemantauan Status Gizi (PSG) Tahun 2017," Kementerian Kesehatan Republik Indonesia, 2018.
3. Dinas Kesehatan Provinsi Jawa Tengah , "Profil Kesehatan Provinsi Jawa Tengah Tahun 2016," Dinas Kesehatan Provinsi Jawa Tengah , Semarang, 2017.
4. Dinas Kesehatan Provinsi Jawa Tengah, "Web Dinas Kesehatan Provinsi Jawa Tengah," [Online]. Available: https://dinkesjatengprov.go.id/v2018/profil-kesehatan-2/. [Accessed March 2019].
5. Dinas Kesehatan Provinsi Jawa Tengah , "Profil Kesehatan Provinsi Jawa Tengah Tahun 2015," Dinas Kesehatan Provinsi Jawa Tengah , Semarang, 2016.
6. Dinas Kesehatan Provinsi Jawa Tengah, "Profil Kesehatan Provinsi Jawa Tengah Tahun 2017," Dinas Kesehatan Provinsi Jawa Tengah, Semarang , 2018.
7. T. Priyo, H. M. Par'i and S. Wiyono, Penilaian Status Gizi, Jakarta Selatan: Pusat Pendidikan Sumber Daya Manusia Kesehatan, 2017.
8. RapidMiner, RapidMiner, [Online]. Available: https://docs.rapidminer.com/. [Accessed June 2019].
9. M. F. Nugroho, "Fitur Seleksi Forward Selection Untuk Menetukan Atribut Yang Berpengaruh Pada Klasifikasi Kelulusan Mahasiswa Fakultas Ilmu Komputer UNAKI Semarang Menggunakan Algoritma Naive Bayes," Jurnal Informatika Upgris, vol. 3, no. 1, 2013.
10. V. Chandani and R. Satria Wahono, "Komparasi algoritma klasifikasi Machine Learning dan feature selection pada analisis sentimen review film," Journal of Intelligent Systems, vol. 1, no. 1, pp. 56-60, 2015.
11. j. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques Third Edition, USA: Morgan Kaufmann Publishers, 2012.
12. Badan Pusat Statistik, "Web BPS - Statistics Indonesia," [Online]. Available: https://www.bps.go.id/. [Accessed March 2019].

AUTHORS PROFILE

Darmansyah Received a bachelor's degree from the Physics Department of Hasanuddin University in 2011. He is Currently Pursuing a master's degree program in Bina Nusantara University, Indonesia. His research interest is in the fields of Data Mining, Machine Learning, and Artificial Intelligence.

Gede Putra Kusuma received PhD degree in Electrical and Electronic Engineering from Nanyang Technological University (NTU), Singapore, in 2013. He is currently working as a Lecturer and Research Coordinator in Computer Science Department, Bina Nusantara University, Indonesia. Before joining Bina Nusantara University, he was working as a Research Scientist in I2R – A*STAR, Singapore. His research interests include pattern recognition, machine learning, face recognition, appearance-based object recognition, mobile learning, and gamification of learning.