Comparison and Analysis of Classification Algorithm Performance for Nutritional Status Data

Herman Yuliansyah
Informatics Department
Universitas Ahmad Dahlan, Indonesia

Sri Winiarti
Informatics Department
Universitas Ahmad Dahlan, Indonesia

Ika Arfiani
Informatics Department
Universitas Ahmad Dahlan, Indonesia

Norma Sari
Faculty of Law
Universitas Ahmad Dahlan, Indonesia

ABSTRACT

Nutritional status data is essential data in analyzing early childhood growth and development. This study conducts experiments based on classification algorithms to predict the nutritional status of early childhood. The nutritional status data is analyzed for early childhood with three class labels are nutritional status based on weight for age, height for age and weight for height. By knowing the best and suitable algorithm, in this case, the algorithm analysis results can be extended to the basis of software development to predict the nutritional status of early childhood. The study results are comparisons of classification algorithms such as Support Vector Machine, K-Nearest Neighbors, Random Forest, Decision Tree, and Naïve Bayes. This study uses the split test method by separating the dataset into training sets and test sets by determining the parameters of the amount of training data that is 10%, 20%, 30%, 40%, and 50%. The experimental results show that the Decision Tree algorithm is superior for weight data for age and height for an age while K-Nearest Neighbors is superior for weight data for height.

General Terms

Data Mining, Classification Algorithm

Keywords

Nutritional Status, Early Childhood, Support Vector Machine, K-Nearest Neighbors, Random Forest, Decision Tree, Naïve Bayes

1. INTRODUCTION

In developing countries, the problem of malnutrition can disrupt the country's economic development so that the number of cases of nutrition must be well anticipated through early prevention and detection. Bodyweight, height and age are a parameter used to measure nutritional status by using classification[1]. In previous research, the nutritional status is identified in toddlers[2] and analyzed the nutritional content of packaged food products[3] based on a data clustering approach. This data clustering is the basis for classifying data. Data clustering aims to separate data based on the proximity of the data. If a cluster has been formed later with the help of experts, the data labelling process will be carried out. So that obtained data that has been clustered and labelled. In addition to going through the data clustering process, data labels can also be taken from real data from particular research objects. This study tries to analyze new data obtained from public health service data. Health experts who have expertise in the processing of the data have labelled this data.

Classification is one technique that usually used in data mining[4] besides association rules[5]–[7] and clustering[8][9]. Classification approach has been implemented to solve various problems such as image classification[10] in soil images[11], breast cancer images[12], Mango classification[13], brainwave signal from EEG single-sensor[14] and motor imagery EEG signal[15], cancer that causes death in children[16], transportation classification in-vehicle pattern analysis[17] and street lighting conditions[18], traffic incident detection[19], student data analysis[20], scattering mechanisms information[21], agriculture classification in honey botanical origin[22] and network security classification for for detecting DoS flooding attacks[23]. Classification also can be used to predict data[24].

The comparison and analysis advantage is to explore the performance of the algorithms for the classification of nutritional status data. There are several classification algorithms in data mining with its advantages and disadvantages. There is no best algorithm in the classification problem, but all depends on the right case and data characteristics. This paper conducts a comparison of several algorithms, such as Support Vector Machine, K-Nearest Neighbors, Random Forest, Decision Tree, and Naïve Bayes. The comparison is needed before implementing in software development in web-based or mobile-based based on the best algorithm after get the comparison results. So, the usage of single algorithms or hybrid algorithms is needed to process the nutritional status dataset.

The main objectives of this article are resumed as follows:

- To present the comparison accuracy of classification algorithm of the nutritional status dataset.
- To present the confusion matrix measurement from the best accuracy of the classification algorithm.
- To present the classification report such as precision, recall and f1-score.

This article organized as follows: Section 2 describes the step of methodology research. Section 3 explores the results and discussions based on finding in this research and give an analysis of the finding. Finally, the conclusion of the report and discuss and future work of nutritional status in Section 4.

2. METHODOLOGY

There are six steps in classification the nutritional status data set, as shown in Figure 1 as follow:

The Preparing data. This process also called data preprocessing. The nutritional status dataset collected from public health service centre. The data is consist of the data feature, i.e., toddler name, gender, date of birth, parent name, address, age, body weight, body height, nutritional status for weight per age, nutritional status for height per age, and nutritional status for weight per height. Then from these data feature, the next step is removing not needed feature such as toddler name, gender, date of birth, parent name, and address. Then remove the noise data or invalid data to get fix dataset. The last step is to split the dataset into three-labelled dataset based in nutritional status for weight per age, nutritional status for height per age, and nutritional status for weight per height.

Split training set and testing set. The three datasets from step 1 divided into the data training set and testing set. We conduct five split scenarios by determining the parameters of the amount of training data that is 10%, 20%, 30%, 40%, and 50%.

The creating objects model classifier. The Scikit-learn library for python programming is used to create the model classifier[25].

The training classifier. This training process is to fit the split training set and testing set into the model classifier.

Make a prediction. The result of this prediction is label data for the training set, then this label used to evaluate the process.

Evaluation. The evaluation process is to measure the quality of prediction. In this evaluation, three measurements are conducted, i.e., accuracy score, confusion matrix, and classification reports such as precision, recall, f1-score, and support.

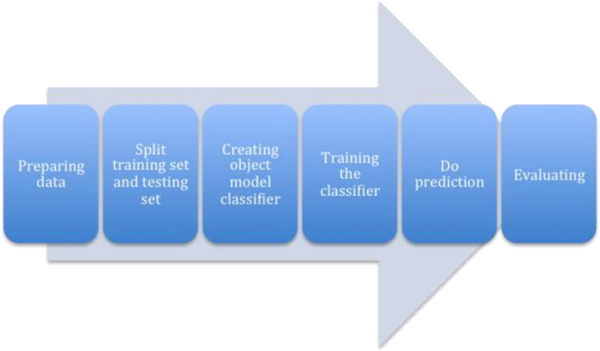

Preparing data
Split training set and testing set
Creating object model classifier
Training the classifier
Do prediction
Evaluating

Fig 1: Methodology to classify the nutritional status

3. RESULTS AND DISCUSSION

Based on the methodology research, several results found which start from the dataset, research finding and also an analysis of the finding as follows:

3.1 DATASET

The preprocessing results, as shown in Table 1 conducted by selecting and removing the data feature. The selection data is based on the required data to process the labelling. Furthermore, the data removing is the detection of missing value and null value in each row data. If the data contain missing value and null value, the data are removed in a row. The process after Table 1 is split the dataset into the three-labelled dataset. The last, there are three set data that contains all of the data features with one label. Figure 2, 3, and 4 show distribution of the label of class based on splitting the dataset process. Then after that, train every dataset to get the classification results.

In Figure 2, the nutritional status for weight per age dataset that labelled with four labels, i.e., malnutrition, nutrition-less, good nutrition, and over nutrition. The data contain majority distribution in the label "Good Nutrition" with 0.879% distribution. It caused most of the nutritional status of early childhood normally "Good Nutrition". Later, the label "Less Nutrition", "Over Nutrition" and "Malnutrition" are distributed with 0.077%, 0.033% and 0.01%, respectively.

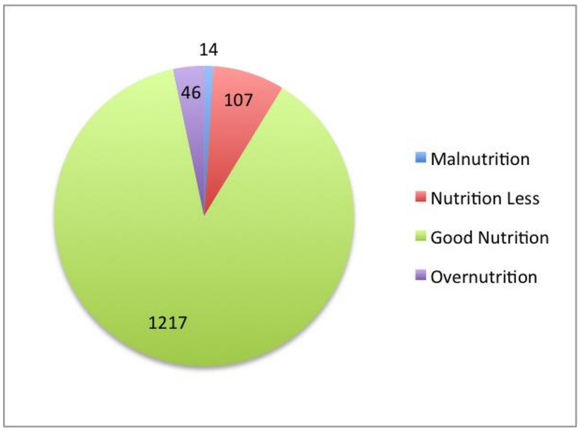


Fig. 2. Data distribution of nutritional status in body weight per age

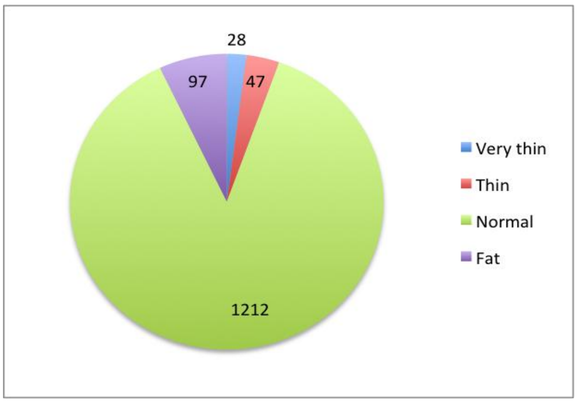


Fig. 3. Data distribution of nutritional status in body height per age

Then Figure 3 shows the nutritional status for height per age dataset that labelled with four labels, i.e., very thin, thin, normal, and fat. Label "Normal" is the majority label in the nutritional status data. It shows that most of the nutritional status of early childhood is "Normal". The label "Normal" achieves 0.875% of data distribution. Later, the label "Fat", "Thin" and "Very Thin" are distributed with 0.070%, 0.033% and 0.02%, respectively.

The last nutritional status for weight per height in Figure 4 also labelled with four labels, i.e., very short, short, normal, and high. Label "Normal" is the majority label in the nutritional status data, as shown in Figure 4. It also shows that most of the nutritional status of early childhood is "Normal". The label "Normal" achieves 0.844% of data distribution. Later, the label "Short", "Very Short" and "High" are distributed with 0.102%, 0.037% and 0.017%, respectively.

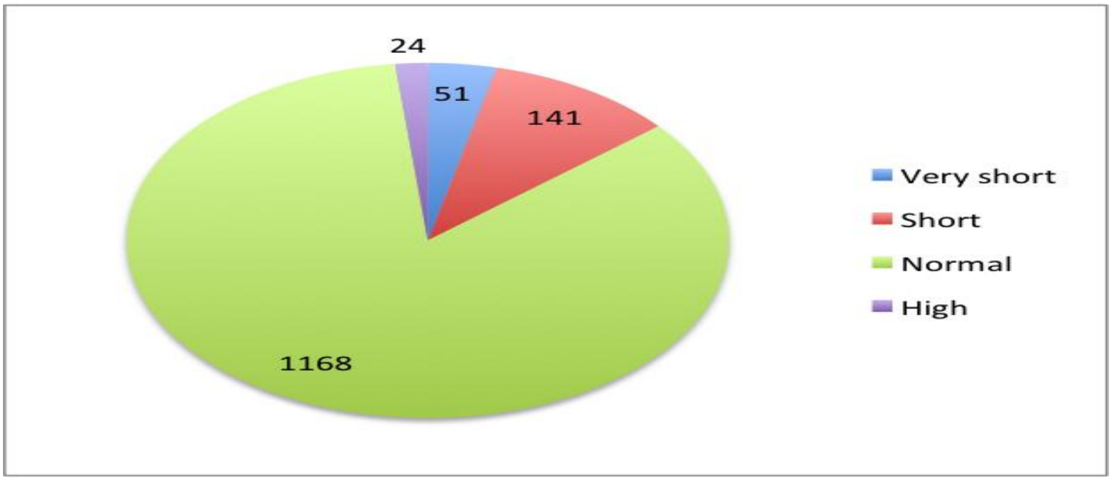


Fig. 4. Data distribution of nutritional status in body weight per body height

Table 1 Dataset of nutritional status

Id	Age (Years)	Body Weight / BW (Kg)	Height (Cm)	Nutritional Status		
				BW / Age	Height / Age	BW / Height
1	56	11.6	92	Malnutrition	Very Short	Normal
2	55	14.8	100.8	Good Nutrition	Normal	Normal
3	52	15.7	107	Good Nutrition	Normal	Normal
4	52	11.8	93.5	Nutrition Less	Short	Normal
5	51	17	100	Good Nutrition	Normal	Normal
…	…	…	…	…	…	…
1382	2	4.7	52.5	Good Nutrition	Short	Fat
1383	1	5.4	51	Good Nutrition	Normal	Fat
1384	2	5.9	55	Good Nutrition	Normal	Fat

3.2 EXPERIMENTAL RESULTS

The experiment was carried out with three iterations and every iteration conducted for one labelled nutritional status dataset. The first iteration is to measure the accuracy of nutritional status based on body weight and age. Furthermore, the second and third iteration is to measure for height and age, body weight and height, respectively. Every iteration is conducted by split training set and testing set in 10%, 20%, 30%, 40%, and 50%. Later, creating object model classifiers based on the classifier algorithms are examined. The classifier is trained to make a prediction based on the split testing data. The evaluation is the last process in every iteration.

Table 2, Table 3, and Table 4 are the experiment results of every iteration for each dataset. Table 2 shows that the decision tree algorithm is superior to other algorithms based on classification accuracy of nutritional status in body weight per age. It is shown based on the experimental result of five training set, and all of the accuracy results show decision tree can achieve the best performance. The smallest accuracy value is in training set 20% and the other accuracy score for every number of the training set can achieve more than 0.9. For each experiment, the average accuracy score for each training set percentage also different. This accuracy score is interest result that there is no fixed number of the training set. For the first experiment in Table 2 shows that the best average accuracy score is in training set 40% with 0.899 accuracy values.

The second experiment in Table 3 shows that the best accuracy score is in training set 10% with 0.935 accuracy values. The different model classification performance is shown in Table 3. There are no single superior model classification performances that can outperform to the examined classifier. Random Forest and Decision Tree are consecutively achieving best model classification performance in training set data. Random Forest can achieve the best performance in training set 20%, 40% and 50%. Meanwhile, Decision Tree can achieve the best performance in training set 10% and 30%. The decision tree achieves the average accuracy value of each model classifier with 0.920 accuracy values.

Then the last examination in Table 4 shows the best accuracy score is 40% (0.852). The experiment results shown that K-Nearest Neighbors achieve the best performance in each training set. Each training set can achieve more than 0.9. If compared to other experiments, as shown in Table 2 and Table 3, the K-Nearest Neighbors results are the smallest accuracy values. Based on Table 2, Table 3, and Table 4 indicate that there are no best model classifications for all experiment. Each model classifications have their own best performance in every experiment.

Table 2 Experimental results for classification accuracy of nutritional status in body weight per age

#	Training set (%)	Model Classification					Average
		SVM	KNN	RF	DT	NB	
1	10	0.871	0.885	0.899	**0.928**	0.878	0.892
2	20	0.848	0.863	0.877	**0.895**	0.859	0.869
3	30	0.868	0.875	0.894	**0.921**	0.875	0.887
4	40	0.881	0.888	0.912	**0.935**	0.877	**0.899**
5	50	0.873	0.879	0.905	**0.909**	0.876	0.888
	Average	0.868	0.878	0.897	**0.918**	0.873	0.887

Table 3 Experimental results for classification accuracy of nutritional status in body height per age

#	Training set (%)	Model Classification					Average
		SVM	KNN	RF	DT	NB	
1	10	0.914	0.928	0.950	**0.957**	0.928	**0.935**
2	20	0.863	0.877	**0.921**	0.917	0.870	0.890
3	30	0.868	0.877	0.913	**0.925**	0.870	0.891
4	40	0.875	0.886	**0.917**	0.913	0.877	0.894
5	50	0.867	0.882	**0.895**	0.886	0.870	0.880
	Average	0.877	0.890	0.919	**0.920**	0.883	0.898

Table 4 Experimental results for classification accuracy of nutritional status in body weight per body height

#	Training set (%)	Model Classification					Average
		SVM	KNN	RF	DT	NB	
1	10	0.863	**0.899**	0.878	0.871	0.712	0.845
2	20	0.852	**0.895**	0.841	0.881	0.708	0.835
3	30	0.853	**0.892**	0.841	0.877	0.726	0.838
4	40	0.870	**0.899**	0.865	0.870	0.756	**0.852**
5	50	0.866	**0.886**	0.857	0.844	0.737	0.838
	Average	0.861	**0.894**	0.856	0.869	0.728	0.842

Finally, the Experimental results resume all of the classification accuracies, as shown in Figure 5 to compare the average accuracy score of every experiment. The Support Vector Machine and Naïve Bayes are the most underperformance accuracy for this nutritional dataset, as shown in Figure 5. So, SVM and NB are not recommending for use both algorithms in nutritional status data.

The other measurements are conducted to do an evaluation of classification results and to confirm the accuracy score of classification, i.e., confusion matrix, and classification reports such as precision, recall, f1-score, and support. Every experiments and training process has own these measurements. The classification measurement is reported based on the best accuracy of every experiment.

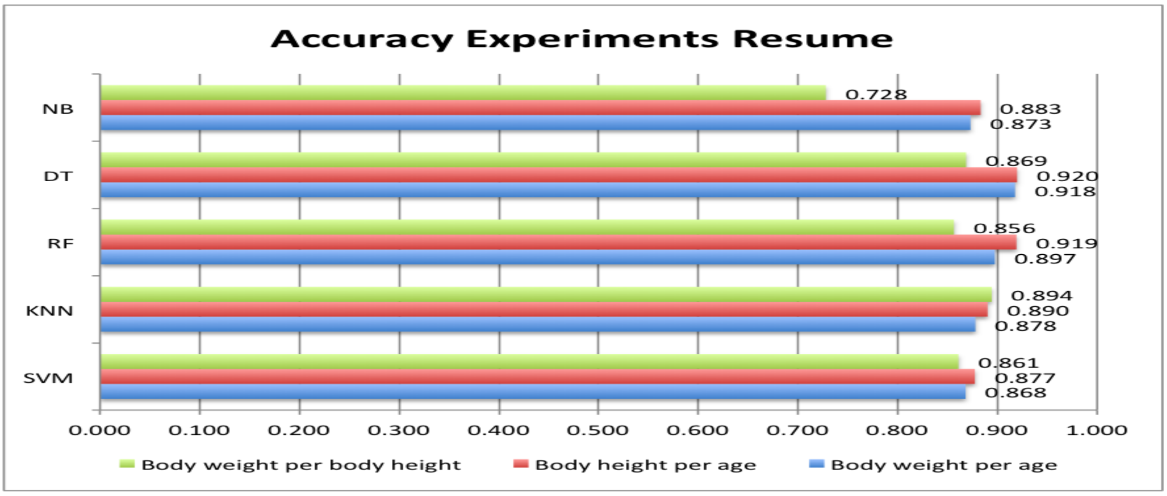
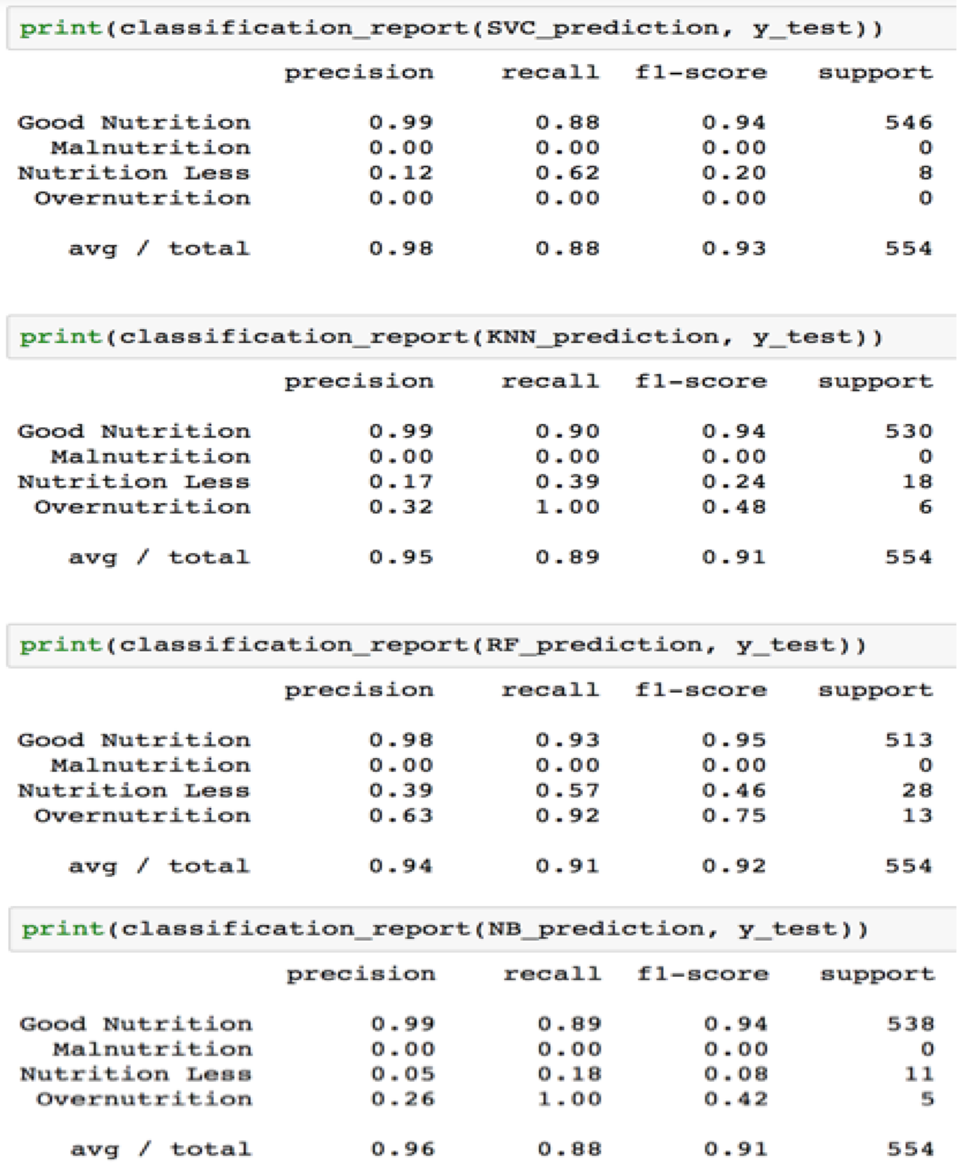


Fig. 5. Average of experiments accuracy

Figure 6, Figure 7 and Figure 8 are classification report for each experiment. Based on the classification report in Figure 6 shows that label malnutrition and has zero value for all algorithms and label ‘over nutrition’ for the SVM algorithm.



```
print(classification_report(SVC_prediction, y_test))

                  precision    recall  f1-score   support

  Good Nutrition       0.99      0.88      0.94       546
    Malnutrition       0.00      0.00      0.00         0
 Nutrition Less       0.12      0.62      0.20         8
   Overnutrition       0.00      0.00      0.00         0

     avg / total       0.98      0.88      0.93       554
```

```
print(classification_report(KNN_prediction, y_test))

                  precision    recall  f1-score   support

  Good Nutrition       0.99      0.90      0.94       530
    Malnutrition       0.00      0.00      0.00         0
 Nutrition Less       0.17      0.39      0.24        18
   Overnutrition       0.32      1.00      0.48         6

     avg / total       0.95      0.89      0.91       554
```

```
print(classification_report(RF_prediction, y_test))

                  precision    recall  f1-score   support

  Good Nutrition       0.98      0.93      0.95       513
    Malnutrition       0.00      0.00      0.00         0
 Nutrition Less       0.39      0.57      0.46        28
   Overnutrition       0.63      0.92      0.75        13

     avg / total       0.94      0.91      0.92       554
```

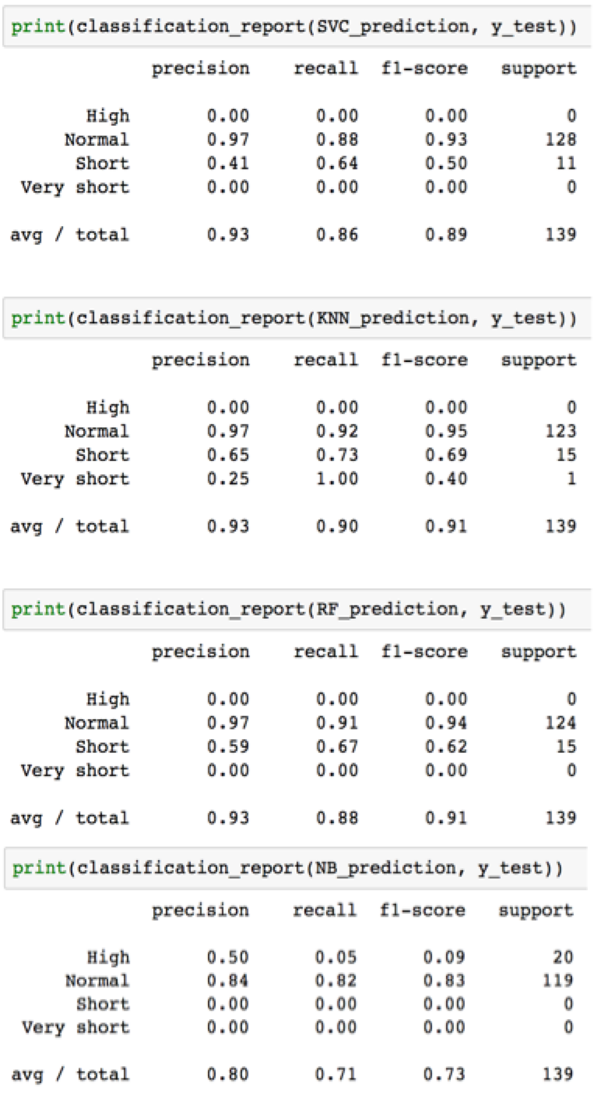
```
print(classification_report(NB_prediction, y_test))

                  precision    recall  f1-score   support

  Good Nutrition       0.99      0.89      0.94       538
    Malnutrition       0.00      0.00      0.00         0
 Nutrition Less       0.05      0.18      0.08        11
   Overnutrition       0.26      1.00      0.42         5

     avg / total       0.96      0.88      0.91       554
```

Fig. 6. Classification report for nutritional status in body height per age



```
print(classification_report(SVC_prediction, y_test))

             precision    recall  f1-score   support

       High       0.00      0.00      0.00         0
     Normal       0.97      0.88      0.93       128
      Short       0.41      0.64      0.50        11
 Very short       0.00      0.00      0.00         0

avg / total       0.93      0.86      0.89       139
```

```
print(classification_report(KNN_prediction, y_test))

             precision    recall  f1-score   support

       High       0.00      0.00      0.00         0
     Normal       0.97      0.92      0.95       123
      Short       0.65      0.73      0.69        15
 Very short       0.25      1.00      0.40         1

avg / total       0.93      0.90      0.91       139
```

```
print(classification_report(RF_prediction, y_test))

             precision    recall  f1-score   support

       High       0.00      0.00      0.00         0
     Normal       0.97      0.91      0.94       124
      Short       0.59      0.67      0.62        15
 Very short       0.00      0.00      0.00         0

avg / total       0.93      0.88      0.91       139
```

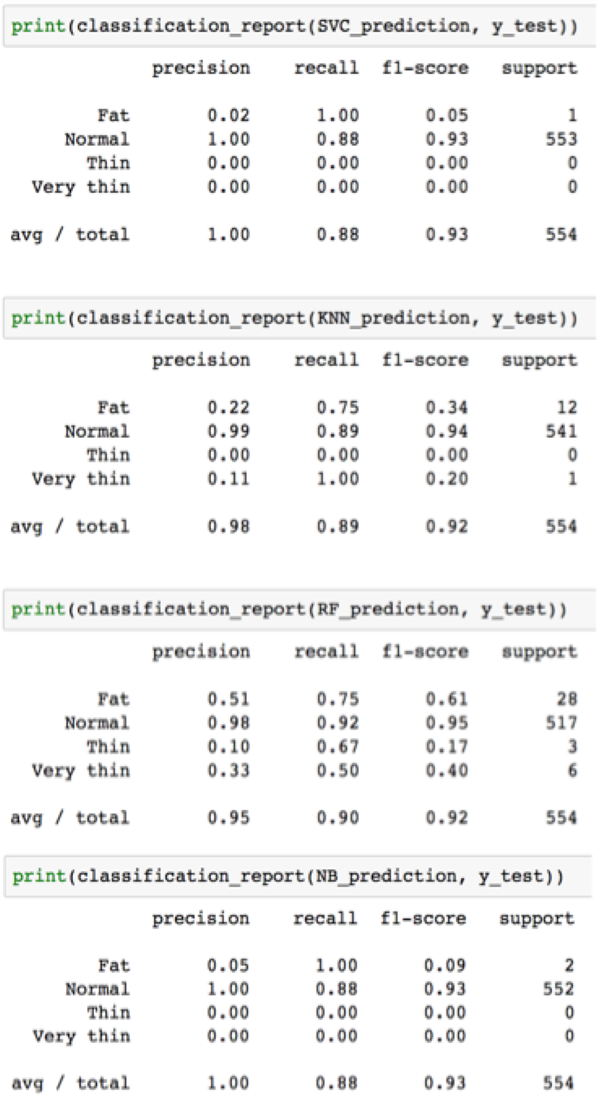
```
print(classification_report(NB_prediction, y_test))

             precision    recall  f1-score   support

       High       0.50      0.05      0.09        20
     Normal       0.84      0.82      0.83       119
      Short       0.00      0.00      0.00         0
 Very short       0.00      0.00      0.00         0

avg / total       0.80      0.71      0.73       139
```

Figure 7 Classification report for nutritional status in body height per age



```
print(classification_report(SVC_prediction, y_test))

             precision    recall  f1-score   support

        Fat       0.02      1.00      0.05         1
     Normal       1.00      0.88      0.93       553
       Thin       0.00      0.00      0.00         0
  Very thin       0.00      0.00      0.00         0

avg / total       1.00      0.88      0.93       554
```

```
print(classification_report(KNN_prediction, y_test))

             precision    recall  f1-score   support

        Fat       0.22      0.75      0.34        12
     Normal       0.99      0.89      0.94       541
       Thin       0.00      0.00      0.00         0
  Very thin       0.11      1.00      0.20         1

avg / total       0.98      0.89      0.92       554
```

```
print(classification_report(RF_prediction, y_test))

             precision    recall  f1-score   support

        Fat       0.51      0.75      0.61        28
     Normal       0.98      0.92      0.95       517
       Thin       0.10      0.67      0.17         3
  Very thin       0.33      0.50      0.40         6

avg / total       0.95      0.90      0.92       554
```

```
print(classification_report(NB_prediction, y_test))

             precision    recall  f1-score   support

        Fat       0.05      1.00      0.09         2
     Normal       1.00      0.88      0.93       552
       Thin       0.00      0.00      0.00         0
  Very thin       0.00      0.00      0.00         0

avg / total       1.00      0.88      0.93       554
```

Figure 8 Classification report for nutritional status in body weight per height

Figure 7 shows that label ‘high’ has zero value for SVM, KNN, Random Forest algorithm and label ‘very short’ has zero value for SVM, Random Forest and Naïve Bayes algorithms. Then for the ‘short’ label in Figure 7 also has zero value for Naïve Bayes algorithms. Figure 8 shows that label ‘thin’ and ‘very thin’ has zero value for SVM and Naïve Bayes algorithm. Then label ‘thin’ has zero value for the KNN algorithm. These all phenomena means that no F1-score calculation for this label.

In general, precision value for all experiment and algorithms can achieve more than 0.93, except Naïve Bayes algorithms in test two. This result shows that all algorithms have steady precision for nutritional datasets.

4. CONCLUSIONS

Based on results and discussions, we can conclude that Decision Tree and K-Nearest Neighbors are best algorithm choice for the dataset characteristic. The classification algorithms should combine or hybrid mode if we want to continue in software development. The experiment result shows that the values of accuracy and precision have influential significant. For future work, we want to explore the classification algorithms with multi-class. It is because the dataset has three feature labels.

5. ACKNOWLEDGMENT

Ministry Research, Technology and High Education, Republic of Indonesia support this work in Community Service Scheme Program with contract number: L1/001/C.6/III/2019 on 16 August 2019.

6. REFERENCES

[1] D. Mclaren and W. C. Read, “Classification of Nutritional Status in Early Chilhood,” Lancet, vol. 300, no. 7769, pp. 146–148, Jul. 1972.

[2] S. Winiarti, H. Yuliansyah, and A. A. Purnama, “Identification of Toddlers’ nutritional status using data mining approach,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 1, 2018.

[3] S. Winiarti, S. Kusumadewi, I. Muhimmah, and H. Yuliansyah, “Determining the nutrition of patient based on food packaging product using fuzzy C means algorithm,” in 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2017, vol. 2017-Decem, pp. 1–6.

[4] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. USA: Morgan Kaufmann, 2012.

[5] H. Yuliansyah and L. Zahrotun, "Designing web-based data mining applications to analyze the association rules tracer study at university using a FOLD-growth method," Int. J. Adv. Comput. Res., vol. 6, no. 27, pp. 215–221, Oct. 2016.

[6] H. Yuliansyah, Hafsah, I. Arfiani, and R. Umar, "Discovering Meaningful Pattern of Undergraduate Students Data using Association Rules Mining," in 2019 Ahmad Dahlan International Conference Series on Engineering and Science (ADICS-ES 2019), 2019, pp. 13–17.

[7] H. Yuliansyah, D. P. Niranda, and I. Arfiani, "Recommender system for high school selection based on apriori method," Int. J. Sci. Technol. Res., vol. 9, no. 2, pp. 2360–2364, 2020.

[8] K. Sya'iyah, H. Yuliansyah, and I. Arfiani, "Clustering Student Data Based On K-Means Algorithms," Int. J. Sci. Technol. Res., vol. 8, no. 8, pp. 1014–1018, 2019.

[9] I. Riadi, S. Winiarti, and H. Yuliansyah, "Development and evaluation of android based notification system to determine patient's medicine for pharmaceutical clinic," in 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2017, no. September, pp. 1–5.

[10] Y. Sun et al., "Image classification base on PCA of multi-view deep representation," J. Vis. Commun. Image Represent., vol. 62, pp. 253–258, Jul. 2019.

[11] U. Barman and R. D. Choudhury, "Soil texture classification using multi class support vector machine," Inf. Process. Agric., Aug. 2019.

[12] Y. Fang, J. Zhao, L. Hu, X. Ying, Y. Pan, and X. Wang, "Image classification toward breast cancer using deeply-learned quality features," J. Vis. Commun. Image Represent., vol. 64, p. 102609, Oct. 2019.

[13] P. Limsripraphan and P. Kumpan, "Algorithm for Mango Classification Using Image Processing and Naive Bayes Classifier," Ind. Technol. Lampang Rajabhat Univ. J., vol. 12, no. 1, pp. 112–125, 2019.

[14] A. Azhari and L. Hernandez, "Brainwaves feature classification by applying K-Means clustering using single-sensor EEG," Int. J. Adv. Intell. Informatics, vol. 2, no. 3, pp. 167–173, 2016.

[15] S. R. Sreeja and D. Samanta, "Classification of multiclass motor imagery EEG signal using sparsity approach," Neurocomputing, Aug. 2019.

[16] S. Alexander et al., "Classification of treatment-related mortality in children with cancer: a systematic assessment," Lancet Oncol., vol. 16, no. 16, pp. e604–e610, Dec. 2015.

[17] E. A. Roxas, R. R. P. Vicerra, L. A. G. Lim, E. P. Dadios, and A. A. Bandala, "SVM Compound Kernel Functions for Vehicle Target Classification," J. Adv. Comput. Intell. Intell. Informatics, vol. 22, no. 5, pp. 654–659, Sep. 2018.

[18] Y. Takama, X. Xu, C.-C. Yu, Y.-S. Chen, and L.-H. Chen, "Classification of Street Lighting Conditions for a Community-Centric System," J. Adv. Comput. Intell. Intell. Informatics, vol. 20, no. 6, pp. 875–881, Nov. 2016.

[19] P. Ambavamata and P. Keeratiwintakorn, "Radar Based Traffic Incident Detection using Support Vector Classification for Road Safety," Inf. Technol. J., vol. 10, no. 2, 2014.

[20] X. Zhou, J. An, X. Zhao, and Y. Dong, "Using Data Mining on Students' Learning Features: A Clustering Approach for Student Classification," J. Adv. Comput. Intell. Intell. Informatics, vol. 20, no. 7, pp. 1141–1146, Dec. 2016.

[21] V. Mittal, D. Singh, and L. M. Saini, "Critical analysis of classification techniques for polarimetric synthetic aperture radar data," Int. J. Adv. Intell. Informatics, vol. 2, no. 1, p. 7, Apr. 2016.

[22] A. Noviyanto and W. H. Abdulla, "Honey botanical origin classification using hyperspectral imaging and machine learning," J. Food Eng., vol. 265, p. 109684, Jan. 2020.

[23] M. Latah and L. Toker, "A novel intelligent approach for detecting DoS flooding attacks in software-defined networks," Int. J. Adv. Intell. Informatics, vol. 4, no. 1, p. 11, Mar. 2018.

[24] Z. P. Agusta and A. Adiwijaya, "Modified balanced random forest for improving imbalanced data prediction," Int. J. Adv. Intell. Informatics, vol. 5, no. 1, p. 58, Dec. 2018.

[25] F. Pedregosa et al., "Scikit-learn: Machine Learning in {P}ython," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.