JOURNAL LA MULTIAPP

VOL. 03, ISSUE 02 (052-060), 2022
DOI: 10.37899/journallamultiapp.v3i2.614

Stunting Classification in Children's Measurement Data Using Machine Learning Models

Syahrial Syahrial[1], Rosmin Ilham[2], Zulaika F Asikin[2], St. Surya Indah Nurdin[2]

[1]*Computer Science Department, University of Muhammadiyah Gorontalo, Indonesia*
[2]*Midwifery Study Program, Muhammadiyah University of Gorontalo, Indonesia*

**Corresponding Author: Syahrial Syahrial*
Email: syahrial@umgo.ac.id

Article Info

Article history:
Received 3 March 2022
Received in revised form 25 March 2022
Accepted 30 March 2022

Keywords:
Classification
Stunting
Machine Learning

Abstract

The study conducted a stunting classification of measurement data for children under 5 years old. The dataset has attributes such as: gender, age, weight (BB), height (TB), weight / height (BBTB), weight / age (BBU), and height / age (TBU). The research uses the CRISP-DM methodology in processing the data. The data were tested on several classification models, namely: logistic regression (LR), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), k-nearest neighbor (KNN), classification and regression trees (CART), nave bayes (NB), support vector machine - linear kernel (SVM-Linear), support vector machine - rbf kernel (SVM-RBF), random forest classifier (RPC), adaboost (ADA), and neural network (MLPC). These models were tested on the dataset to find out the best model in accuracy. The test results show that SVM-RBF produces an accuracy of 78%. SVM-RBF has consistently been at the highest accuracy in several tests. Testing through k-fold cross validation with k=10.

Introduction

According to WHO stunting malnutrition is a growth and development disorder in children caused by malnutrition (malnutrition), repeated infections, and insufficient psychosocial stimulation. Stunting diagnosis by measuring height and comparing with growth curve based on WHO. The history includes a history of growth, risk factors, and a physical examination, especially to see the presence of body dysmorphic and disproportionate conditions, to support the diagnosis of stunting. The need for predicting infant nutrition levels is very high considering that the stunting rate in Indonesia since 2018 is still above 30%. Several classification methods have been used to measure nutrient levels. However, the method is applied to different data so that it can cause bias in its accuracy.

This study aims to find a good classification method with high accuracy. The application of the method is carried out by utilizing various methods for classifying the same data. Currently, there are stunting data from several villages within the Kab. Gorontalo. The data to be processed are 1731 records. It is expected to get a machine learning classification method that has the highest accuracy. This research collects references from various sources such as: journals, reports, proceedings. This is done to gain insight into the existing problems.

Stunting Malnutrition

Stunting is a chronic malnutrition problem caused by inadequate nutritional intake for a long time due to feeding that does not match nutritional needs. Stunting occurs when the fetus is still in the womb and only appears when the child is two years old. Malnutrition at an early age

increases infant and child mortality, causes sufferers to get sick easily and have poor posture as adults (Indonesia, 2014).

Stunting according to the WHO Child Growth Standard is based on the index of body length for age (PB/U) or height for age (TB/U) with a limit (z-score) <-2 SD (WHO, 2021). The TB/U indicator describes nutritional status that is chronic, meaning that it appears as a result of long-standing conditions such as poverty, inappropriate parenting behavior, often suffering from repeated illnesses due to poor hygiene and sanitation (DepKes, 2007).

Zscore=(MV-AV)/SS (1)

MV = Measured value

AV = Average value in the reference population

SD = Standard deviation of the reference population

The results of the z-score above then refer to the grouping table to determine the nutritional status.

Table 1. Malnourish Classification Reference

Index	Classification	Z-Score
TB/U	Normal	-2.0 < Z
	Moderately malnourished	-3.0 < Z < -2.0
	Severely malnourished	Z < -3.0 or edema

Supervised Learning on Machine Learning

In this paper, we use various classification techniques in the category of supervised learning. There are many algorithms available. Logistic regression (LR), this algorithm uses a single estimator in a multinomial logistic regression model on data with classes. Use LR to define boundaries between classes and establish class probabilities that depend on the distance from the boundary. The probability moves quickly to an extreme value (0 or 1) as the data set gets larger. Based on that, LR is more than just a method for classification. This algorithm can make more robust and detailed predictions and can be adjusted in different ways. LR is a predictive approach whose results can be dichotomous. LR is linear interpolation and is commonly used in applied statistics and discrete data analysis (Peng & Lee, 2021).

Linear discriminant analysis (LDA) is a statistical-based learning function. LDA provides the probability of each data in the class compared to just doing the classification. LDA is used in statistics and machine learning to find a linear combination of features that can best separate two or more classes (Kotsiantis et al., 2007). Quadratic discriminatory analysis (QDA) is generally used as a statistical tool for observing different multivariate in normal population (Ghosh et al., 2021). QDA based on graphical lasso can also be classified on the recognition of human activity data (Jinjia et al., 2019).

K-nearest neighbor (KNN) is a method of classifying data based on its distance or similarity from other data in the vicinity. KNN can use various methods in calculating the distance (Fachrie, 2020). Classification and regression trees (CART) is a classification model that involves the identification and construction of a binary decision tree using training data whose class is well known. The number of entities in the two subgroups defined in each binary split, corresponding to the two branches emerging from each intermediate node, becomes progressively smaller, so a sufficiently large training sample is required if good results are to be obtained. The decision tree starts with the root node t derived from any variable in the feature space that minimizes the size of the impurity of the two sibling vertices (McLachlan, 2005).

Support vector machine (SVM) performs data classification by looking for a maximized hyper plane that can be used to separate the 2 classes. This model is closely related to the classical

multilayer perceptron neural network. SVM revolves around the notion of margin‖—the two sides of a hyperplane that separates two classes of data are often also called linear SVM (SVM-Linear). Maximizing the margin and thereby creating the largest possible distance between the separating hyperplane and the instances on either side has been shown to reduce the upper bound on the expected generalization error (Kotsiantis et al., 2007). SVM with radial basis function kernel (SVM-RBF) which is a 3-layer feedback network. Each hidden layer applies a radial activation function whose output is derived from the sum. The data training process is generally divided into 2 stages. The first stage is determining the center and width of the hidden layer based on the clustering algorithm. The second stage of weights connecting the hidden layer with the output layer is determined by the Singular Value Decomposition (SVD) or Least Mean Squared (LMS) algorithm (Howlett & Jain, 2001).

The random forest classifier (RFC) is a collection of CART models trained on a data set of the same size as the training set, called a bootstrap, created from random re-sampling of the training set itself. After the tree is constructed, a bootstrap set, which does not include certain records from the original dataset in the form of an out-of-bag (OOB) sample, is used as the test set. The misclassification rate of all test sets is an OOB estimate of the generalization error (Breiman et al., 2017). Adaptive boosting (ADA) is a boosting method that works through weight adjustment without requiring any a priori knowledge about learning from the method (Schapire, 1990). Furthermore, improvements were made in the form of AdaBoost.M1, AdaBoost.M2, AdaBoost.R, AdaBoost.MO, AdaBosst.MH. AdaBoost has the advantages of speed, simple operation, and easy implementation in computer programs. The available parameters are only the number of iterations so that it is easy to combine with other methods (Freund et al., 1999).

Multilayer perceptron classifier (MLPC) is an artificial neural network model that belongs to the feedforward category. MLPC consists of at least 3 layers of nodes, namely: the input layer, the hidden layer, and the outer layer. The hidden layer and the outer layer use a nonlinear activation function. MLPC uses the backpropagation method in doing learning. MLPC uses backpropagation in studying data patterns. Backpropagation applies gradient descent in search of optimal learning. Parameter adjustments need to be made in order to produce good test results (Windeatt, 2008).

K-Fold Cross Validation

K-fold cross validation (K-FOLD) is a cross validation method to evaluate and validate models in the context of this study on the classification model. This method applies several combinations of the amount of training data and the amount of testing data from the dataset. The number of combinations is determined from the value of k. Suppose the value of k = 3, then there are 3 combinations, namely: 1/3 of the dataset becomes training data and two 2/3 becomes testing data, 1/2 dataset becomes training data and also testing data, and 2/3 of the dataset becomes training data and the remaining 1/3 becomes testing data (Stone, 1974).

Methods

This research refers to the Cross Industry Standard Process for Data Mining (CRISP-DM) in the process of data processing, modeling, and evaluation. This method is a data mining process with a life cycle which is divided into 6 phases which include business understanding, data understanding, data processing, modeling, evaluation, and deployment. For research purposes, we only use 5 phases.

The stages in the CRISP-DM (Cross Industry Standard Process for Data Mining) life cycle used in this study are; (1) Business Understanding At the business understanding stage, an understanding of research is carried out as the best decision support for classifying stunting nutritional status which is expected to be able to provide the best treatment for toddlers who experience stunting nutritional status, so that anticipatory steps can be taken in the form of early prevention of children under five with stunting nutritional status. This is done to prevent

and reduce the rate of toddlers experiencing stunting which results in disability so that the development of toddlers is hampered or can even result in death. Utilization of the dataset and the use of the proposed algorithm, this research will achieve the planned goals; (2) Data Understanding The second phase or phase carried out in the life cycle of the CRISP-DM method is data understanding. The dataset used in this study is a dataset on the nutritional status of stunting in toddlers in 2020 which was obtained from the Pandanaran Health Center Semarang. The dataset of the stunting nutritional status of toddlers totals 300 data records. In the dataset there are 5 attributes and 1 label with integer and binominal data types. From a total of 300 records in the stunting nutritional status data for toddlers, it is stated that 258 toddlers experience stunting nutritional status and 42 other toddlers experience normal nutritional status; (3) Data Processing (Data Preparation) This stage processes the dataset to simplify the data. The existing dataset has attributes, namely: gender (Gender), age in months (Age), weight kg (BB), height cm (TB), weight/height (BBTB), weight/age (BBU).) and height / age (TBU). The classification attribute is class; (4) Modeling This stage is the application of modeling techniques using classification algorithms that are included in supervised machine learning. At the modeling stage in this study is the classification of stunting nutritional status in toddlers by applying the following algorithms: logistic regression (LR), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), k-nearest neighbor (KNN), classification and regression trees (CART), nave bayes (NB), support vector machine - linear kernel (SVM-Linear), support vector machine - rbf kernel (SVM-RBF), random forest classifier (RPC), adaboost (ADA), and neural network (MLPC). Each algorithm applies its respective default parameters; (5)Evaluation This stage is to evaluate the previous stage, namely modeling. The purpose of the evaluation is to adjust the model obtained so that it is appropriate and in accordance with the targets to be achieved. At this stage, the models are tested in k-fold cross validation with k=10. The test results will be compared to see which model is the best in classifying stunting nutritional status based on its accuracy.

Results and Discussion

The study followed the CRISP-DM flow with the use of 5 phases.

Business Understanding

This phase is done to understand the attributes of the data. The data attributes in the dataset are gender (Gender), age in months (Age), weight kg (BB), height cm (TB), weight/height (BBTB), weight/age (BBU) and height / age (TBU). The number of data in the dataset is 1731 records.

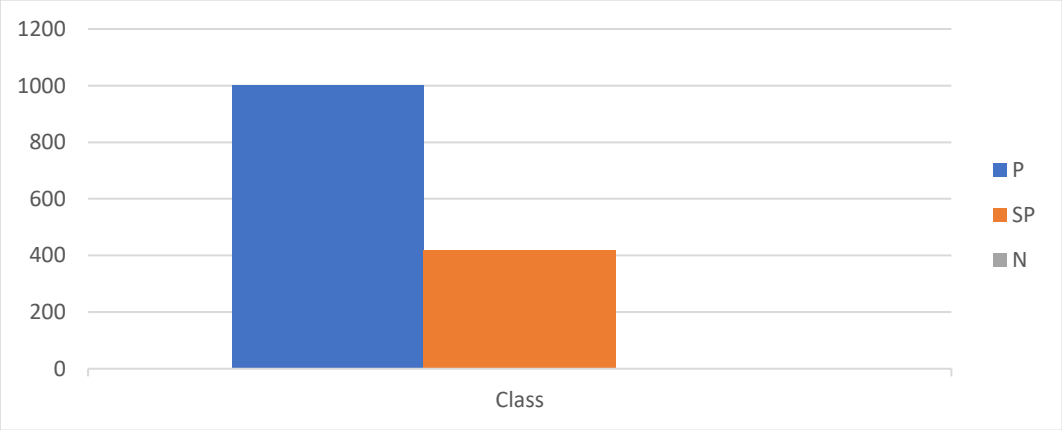


Figure 1. Amount of Records for Every Class

The number of recordings between classes looks unbalanced, because the data on the ground there are indeed many that fall into that category.

Data Understanding

The historgram distribution of data for each attribute is shown in Figure 2 below. The histogram shows that the age attribute is evenly distributed from 0 months to 40 months. Height shows that heights 60 to 70 have the highest amount of data

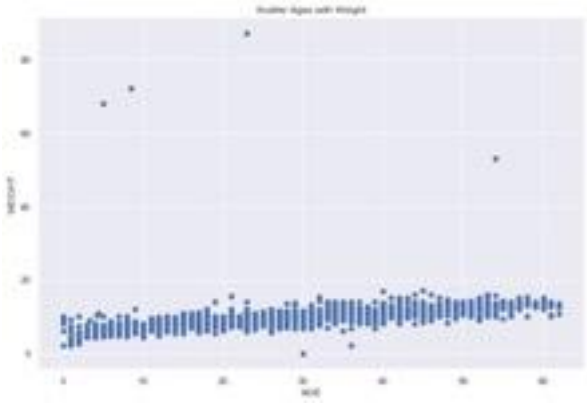


Figure 2. Scatter Data Based On Age With Weight

Figure 3 below shows that there is an outlier in height below from the existing data set. This happened due to a recording error at the time of data collection. This outlier data will be removed from the dataset at the data preparation stage. Outliers will greatly affect the performance of the classification model later.

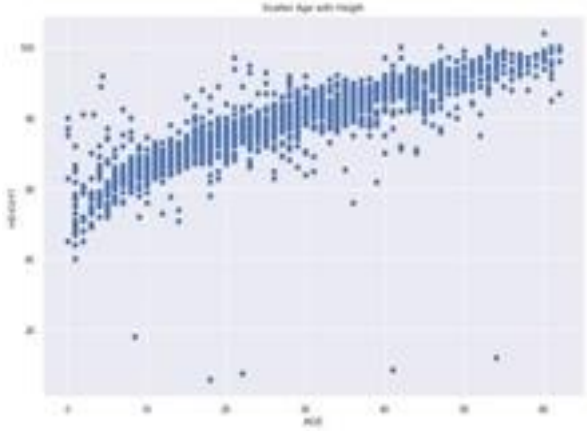


Figure 3. Scatter Data Based On Age With Height

On the distribution of data based on height and weight shows outliers from each data

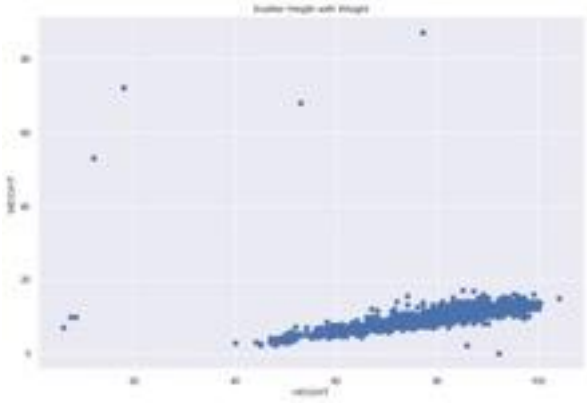


Figure 4. Scatter Data Based on Height With Weight

Figure 4 shows an outlier at a height below 20

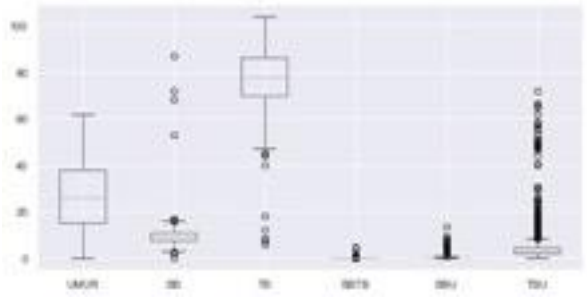


Figure 5. Boxplot Diagram of Dataset

The boxplot diagram also shows that there are outliers in height and weight. Figure 6 shows the correlation between the attributes in the dataset. There is a high correlation between height and age and weight and age

Figure 6. Correlations of Every Attribute on Dataset

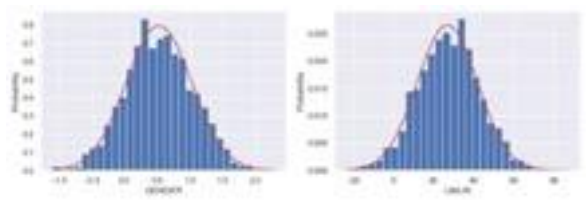
Data Preparation

This phase changes the target value from symbol to number: SP=0, P=1, N=2 and gender P=0, L=1 in gender attributes

	GENDER	UMUR	BB	TB	BBTB	BBU	TBU	TARGET	KELAS	JK
0	0.0	24.0	8.8	79.0	0.111392	0.366667	3.291667	1.0	P	P
1	0.0	26.0	10.0	84.0	0.119048	0.384615	3.230769	1.0	P	P
2	1.0	21.0	8.0	73.0	0.109589	0.380952	3.476190	1.0	P	L
3	0.0	17.0	7.0	71.0	0.098592	0.411765	4.176471	1.0	P	P
4	0.0	36.0	10.5	79.0	0.132911	0.291667	2.194444	1.0	P	P
...	...	...	...	...	...	...	...	...	...	...
1726	0.0	30.0	10.0	84.0	0.119048	0.333333	2.800000	1.0	P	P
1727	0.0	47.0	11.0	89.0	0.123596	0.234043	1.893617	0.0	SP	P
1728	1.0	1.0	6.9	66.0	0.104545	6.900000	66.000000	1.0	P	L
1729	0.0	0.0	6.0	63.0	0.095238	0.000000	0.000000	1.0	P	P
1730	1.0	47.0	12.8	95.0	0.134737	0.272340	2.021277	1.0	P	L

1731 rows × 10 columns

Histogram representation of each attribute



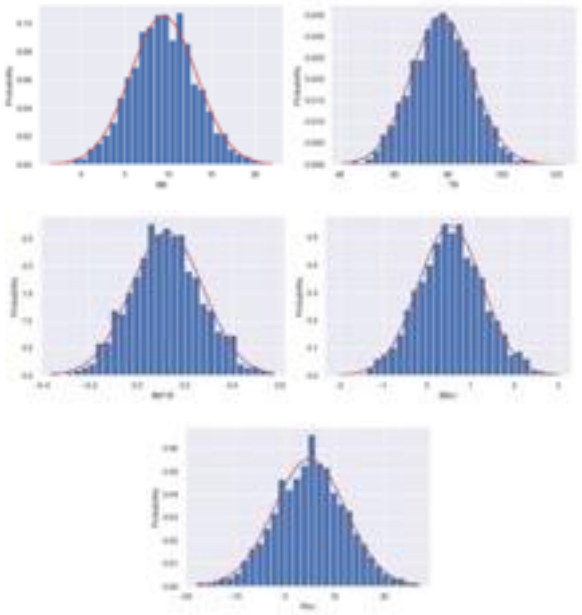


Figure 7. Histogram of Everi Attibute

Modelling

The classification modeling process is carried out in several stages, namely; (1) Divide the dataset into X and y; (2) Split X and y into the training and the test set; (3) Scaling the features; (4) Build the list of models to use; (5) Evaluate each model; (6) Compare algorithms.

This step is done to form a good model in doing classification. Data testing must be prepared by dividing a certain amount of data into training data and data testing.

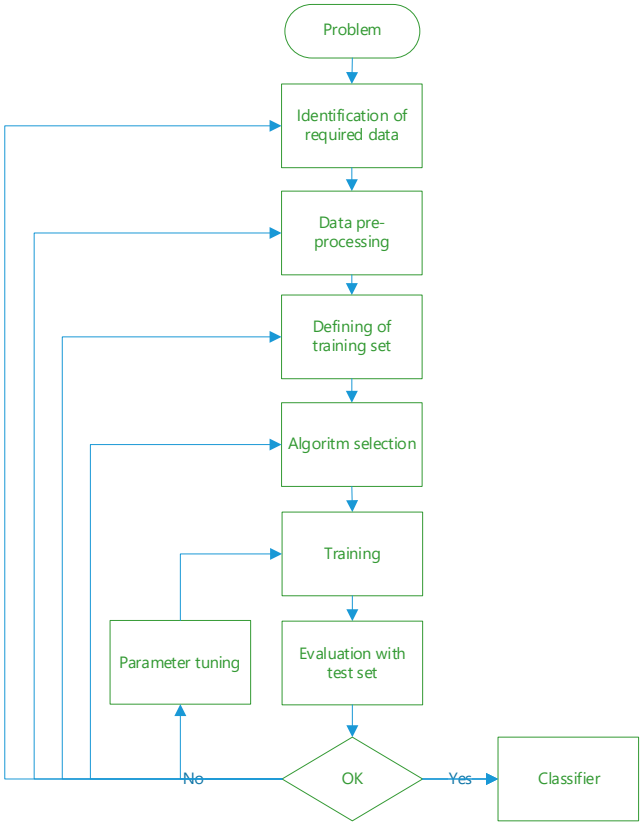


Figure 8. Processes of Supervised Machine Learning

The process of each algorithm is done modeling based on the process on figure 8. Algorithms applied

	LR	LDA	QDA	KNN	CART	NB	SVM-Linear	SVM-RBF	RFC	ADA	MLPC
0	0.719586	0.695006	0.581722	0.741320	0.671932	0.462423	0.635857	0.774533	0.721030	0.572214	0.619232
1	0.716787	0.695819	0.583047	0.748608	0.672771	0.457345	0.636550	0.778939	0.719711	0.581712	0.619216
2	0.723986	0.692915	0.565155	0.742733	0.676937	0.454536	0.635148	0.772333	0.713143	0.565707	0.619258
3	0.724700	0.697967	0.558539	0.755787	0.686461	0.463195	0.635168	0.776050	0.713205	0.587384	0.619268
4	0.718241	0.694391	0.577474	0.752878	0.691482	0.450193	0.635148	0.775279	0.716088	0.590246	0.619226
5	0.721828	0.693650	0.559332	0.744203	0.661172	0.449395	0.636576	0.774575	0.721061	0.568695	0.619211
6	0.721103	0.700162	0.566515	0.742811	0.684183	0.448733	0.635903	0.774575	0.708789	0.584611	0.619299
7	0.722485	0.695814	0.569398	0.756480	0.688573	0.455276	0.636524	0.771656	0.721093	0.587405	0.619195
8	0.725435	0.696518	0.549338	0.753566	0.676978	0.461089	0.637290	0.778110	0.710171	0.570123	0.619221
9	0.724596	0.697159	0.554775	0.747790	0.682729	0.463064	0.634334	0.773058	0.715989	0.591826	0.619164
10	0.721812	0.692185	0.551366	0.748530	0.673423	0.458253	0.637285	0.776009	0.731931	0.596142	0.619221
11	0.731196	0.697253	0.562856	0.752190	0.674820	0.452398	0.634381	0.781050	0.725430	0.572328	0.619206
12	0.723251	0.696528	0.556225	0.745678	0.666328	0.461506	0.637968	0.773934	0.717605	0.555620	0.619206
13	0.724002	0.693656	0.567855	0.748483	0.679142	0.452304	0.636612	0.778881	0.703707	0.587452	0.619278
14	0.719732	0.696580	0.582291	0.747195	0.673506	0.454410	0.637379	0.773188	0.716870	0.594057	0.619320
15	0.722532	0.696471	0.580211	0.750756	0.694385	0.455953	0.636508	0.781790	0.723287	0.591143	0.619159
16	0.720384	0.696523	0.583792	0.748561	0.685679	0.460062	0.635763	0.776035	0.713163	0.593092	0.619143
17	0.722537	0.699421	0.550735	0.758680	0.678495	0.455291	0.634376	0.777432	0.712454	0.583114	0.619185
18	0.727599	0.697300	0.555589	0.751428	0.664024	0.447899	0.637280	0.771682	0.706636	0.599046	0.619237
19	0.724622	0.696450	0.579403	0.749270	0.677755	0.458096	0.635799	0.776733	0.713158	0.568679	0.619206
20	0.720342	0.696507	0.554734	0.743457	0.690011	0.453774	0.635132	0.768783	0.719555	0.590324	0.619242
21	0.721890	0.688635	0.560625	0.739229	0.681410	0.458112	0.635158	0.770326	0.700918	0.604749	0.619237
22	0.728303	0.692884	0.562037	0.736331	0.666182	0.461615	0.634319	0.770269	0.711756	0.601872	0.619138
23	0.727667	0.698055	0.591065	0.744224	0.684266	0.450751	0.633714	0.776744	0.718966	0.561495	0.619237
24	0.721134	0.692957	0.599734	0.750005	0.684235	0.445871	0.639422	0.767318	0.716771	0.575795	0.619211

Figure 9. Result of every models via k-fold cross validation at k=10

Evaluation

K-fold cross validation is applied to each classification model using k=10. Testing is also performed on seed parameters from 1 to 25. Results show svm-RBF is consistently at the top accuracy compared to other models. The highest accuracy was obtained on SVM-RBF of 78.10% at seed=12

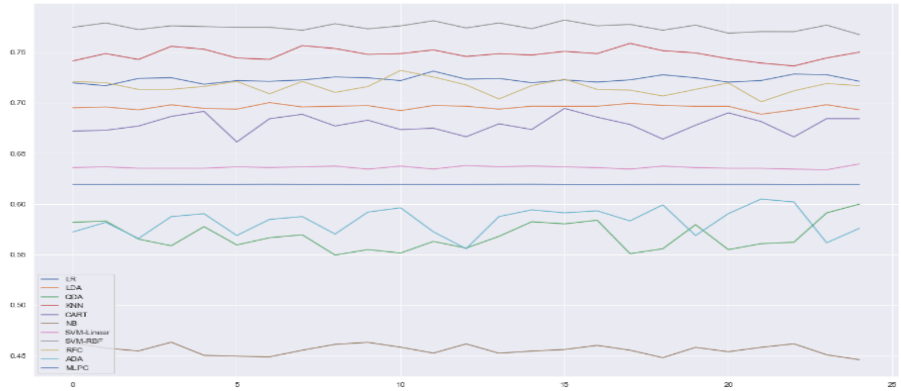


Figure 10. Models test on seed 1 to 25 on kfold cross validation k=10.

The parameters of each model are used by the default parameters only. This is done to find out the initial ability of each model in classifying datasets.

Conclusion

The results of testing the classification of measurement data of children under 5 years old using machine learning models can be done. The results showed the highest accuracy on the SVM model with kerner=RBF of 78.10%. The SVM-RBF consistently tops its accuracy compared to the other 10 models with default parameters. Further research needs to be done tuning parameters on each model to get the best parameters from each model for dataset classification.

References

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Routledge.

Fachrie, M. (2020). Machine Learning for Data Classification in Indonesia Regional Elections Based on Political Parties Support. *Jurnal Ilmu Komputer dan Informasi*, *13*(2), 89-96.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, *14*(771-780), 1612.

Ghosh, A., SahaRay, R., Chakrabarty, S., & Bhadra, S. (2021). Robust generalised quadratic discriminant analysis. *Pattern Recognition*, *117*, 107981.

Howlett, R. J., & Jain, L. C. (2001). *Radial basis function networks 2: new advances in design* (Vol. 2). Springer Science & Business Media.

Indonesia, M. C. A. (2014). Proyek Kesehatan dan Gizi Berbasis Masyarakat untuk Mengurangi Stunting. *Tersedia di http://www. mcaindonesia. go. id/assets/uploads/media/pdf/Factsheet_HN_ID. pdf (diakses 25 Oktober 2018)*.

Jinjia, W., Shaonan, J., & Yaqian, Z. (2019). Quadratic Discriminant Analysis Based on Graphical Lasso for Activity Recognition. In *2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP)* (pp. 70-74). IEEE.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3-24.

McLachlan, G. J. (2005). *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons.

Peng, C. Y. J., & Lee, K. L. 8c Ingersoll, GM (2021). *An introduction to logistic regression analysis and reporting. Thejournal ofEducational Research*, *96*(1), 3-14.

Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, *5*(2), 197-227.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, *36*(2), 111-133

WHO. (2021). “Weight-for-length/height,” 2021.

Windeatt, T. (2008). Ensemble MLP classifier design. In *Computational Intelligence Paradigms* (pp. 133-147). Springer, Berlin, Heidelberg.