JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 2, Maret 2024, hlm. 267 - 274 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i2.2973
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

DETECTION OF CHILDREN'S NUTRITIONAL STATUS USING MACHINE LEARNING WITH LOGISTIC REGRESSION ALGORITHM

Yuliana[1*], Paradise[2], Mudawil Qulub[3]
[1]Information Technology, Shanti Bhuana Institute
[2]Informatics Engineering, Telkom Institute of Technology Purwokerto
[3]Computer Science, Bumigora University
email: *yuliana@shantibhuana.ac.id

Abstract: Children's nutritional issues are an important concern for parents to pay attention to growth and development, especially health and well-being. According to the results of the Ministry of Health's Indonesian Nutrition Status Survey (SSGI), there are 4 nutritional problems for children in Indonesia, namely stunting, wasting, underweight and everweight. In this research, how to predict signs of symptoms of a decline in a child's nutritional status using a machine learning algorithm, a prediction model was designed using logistic regression in Python IDE to predict whether a child is indicated by a decline in nutrition or not. Dataset from Bengkayang Community Health Center data consisting of 657 pediatric patient data. The dataset is divided into 7 features (independent variables) and 1 predictor (dependent variable). Test results show perfect performance with precision, recall, F1-score, accuracy values of 100%. Then the visualization results on the ROC (Receiver Operating Characteristic) curve to depict the TP (True Positive) value on the Y axis against the FP (false Positive) value on the become overfit. It is recommended that in preparing the training dataset, measure the training data and reduce the features, after carrying out feature selection to increase the accuracy of the model.

Keywords: child nutritional status; growth and development logistic regression; machine learning

Abstract: Masalah Gizi anak menjadi perhatian penting bagi orangtua untuk memperhatikan tumbuh kembang, terutama kesehatan dan kejahteraan. Menurut hasil survei status Gizi Indonesia (SSGI) Kemenkes memperlihatkan 4 permasalahan gizi anak di Indonesia yaitu *stunting*, *wasting*, *underweight*, dan *everweight*. Dalam penelitian ini, bagaimana memprediksi tanda gejala penurunan status gizi anak menggunakan *algoritma machine learning* dirancang model prediksi menggunakan *logistic regression* pada Python IDE dengan memprediksi anak terindikasi penurunan gizi atau tidak. Dataset dari data Puskesmas Bengkayang yang terdiri 657 data pasien anak. Dataset dibagi menjadi 7 feature (variabel independen) dan 1 predictor (variabel dependen). Hasil Pengujian memperlihatkan kinerja yang sempurna dengan nilai presisi, recall, F1-score, akurasi, sebesar 100%. Kemudian hasil Visualisasi pada kurva ROC (*Receiver Operating Characteristic*) untuk menggambarkan nilai TP (True Positif) di sumbu Y terhadap nilai FP (false Positif) di sumbu X juga menunjukkan nilai yang sangat tinggi dan sudah mendekati angka 1 ini pertanda bahwa model ini menjadi overfit. Sebaiknya dalam persiapan training dataset diukur dengan data training dan mengurangi feature, setelah melakukan feature Selection untuk meningkatkan akurasi model.

Keywords: logistic regression; machine learning; status gizi anak; tumbuh kembang

INTRODUCTION

Children are the assets of the nation and future generations who will continue the future progress and development of the Nation [1]. So that the state's hopes and attention are focused on the growth and development of children. The 1945 Constitution Article 28B paragraph 2 reads "Every child has the right to survival, growth and development and the right to protection from violence and discrimination. In their growth, children must be prosperous, nurtured, protected and guided with love and affection, especially in their family environment [2].

In Indonesia, the problem of child nutrition is still quite high, which begins with weight loss. According to the results of the Indonesian Nutrition Status Survey (SSGI) of the Ministry of Health, there are 4 child nutrition problems in Indo nesia, namely stunting, wasting, underw eight, and everweight. Stunting is one of the nutritional problems that is of concern to the government and society because the rate of spread is still quite high, in 2022 it is estimated to reach 21.6%. Another nutritional problem is wasting or thinness. According to the SSGI 2022, the prevalence of wasting in Indonesia increased by 0.6 points from 7.1% to 7.7 last year [3].

The problem of child nutrition is certainly an important concern for parents to pay attention to child growth and development, especially the health and welfare of children as the successor of the family. Parents need to pay attention to healthy nutrition that is balanced and a healthy lifestyle and positive attitude. Nutritional problems for children can lead to various diseases such as growth failure, undernutrition and malnutrition, premature babies, low birth weight babies, cow's milk protein aller gies and congenital metabolic disorders. Another problem has a major impact on the number of stunting cases. One of the government's current focuses is stunting prevention.

This effort aims for Indonesian children to grow and develop healthily by having emotional, social, and physical readiness to learn and innovate and comp ete at the global level. This needs attention from stakeholders to monitor the improvement of toddler growth, posyandu, increase access to medical services for sick children under five, improve the quality of household drink ing water, correct defecation behavior, and provide BANSOS. In order to create healthy child development for the next generation of a healthy nation.

Referring to previous researchers on the topic of applying Artificial Intelligence-based information technol ogy in the field of Machine Learning by taking the problem of early detection of diabetes with experimental results show ing that hyperparameter tuning-based models that can improve performance to predict accuracy values of 82%, 81% precision, 79% recall and 80% F1-score [4].

Analyzing machine learning mod els for human activity recognition with research results using accuracy and preci sion efficiency approaches [5]. Applying Hybrid machine learning in heart disease prediction models results in an accuracy value of 84.48% increasing to 1.32% [6]. Comparing the upport Vector Machine (SVM) and Artificial Neural Network (ANN) methods on Toddler Nutrition Classification Case Study at Salissingan Health Center from the results of the analysis obtained the size of getting the best method [7].

Prediction of stunting in toddlers by applying the k-Nearest Neighbors classification algorithm based on the parameters used in the dataset. From the system built, it shows an accuracy of 97% based on data partitioning testing in the confusion matrix using data partitions 90% training data and 10% testing data with k = 5 nearest neighbors [8].

In this study, which is a differentiator from previous researchers who need to be followed up on how to predict the detection of signs of symptoms of decreased nutritional status of children using machine learning algorithms. In this study, a prediction model is designed using logistic regression in Python IDE for the detection of symptoms of nutritional status decline by providing predictions of children who are indicated to be malnourished or not malnourished based on the initial data provided. experiments were conducted using datasets from bengkayang puskesmas data using measurements of independent variables and dependent variables.[9][10][11]

METHOD

This research uses the logistic regression method. The dataset that has been obtained is divided according to the criteria, then testing is carried out by dividing the data, applying the logistic regression method for modeling, then measuring the performance of the model. The following is the flow of research conducted can be seen in image 1.

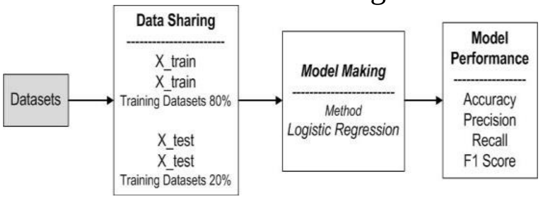


Image 1. Research flow

Logistic Regression

Logistic regression is used to predict binary categories (0 or 1), with only two possibilities as predictions 'yes or 'no, win/lose, happy/unhappy etc.'. These predictions are made based on one or more features (independent variables) that become predictors (dependent variables). Each feature will be given its own weight. [12][13][14][15]

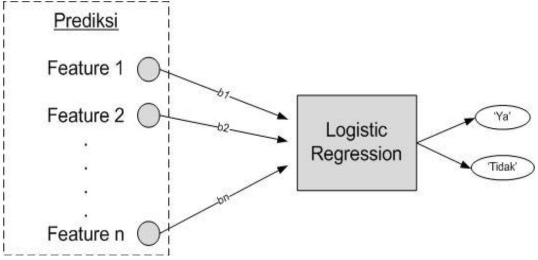


Image 2. Feature and weight simulation to generate predictions

Logistic regression is a method in the field of statistics used by machine learning to be processed in computers or we can call it a logistic function or Sigmoid function, which produces output in the form of a curve denoted by the letter S with a value between 0 and 1. The logistic regression model is used to model the probability (possibility) if it is greater than 0.5 then it can be considered that the input value inputted into the model will be categorized into class 0, otherwise if the probability number is lower than 0.5 then the input value will be categorized into class 1. Logistic regression requires one or more features as predictors. If there are as many as n features (representing the letter x) then to get an output will require n + 1 coefficients (represented by b) as follows: Output = bo + b1X1 + b2X2 + ... bnXn This method finds the best values for the coefficients bo, b1, b2, and bn based on calculations using the available training dataset.

Child Nutrition Status Dataset

The dataset for predicting the

decline in children's nutritional status used in this study is sourced from the Bengkayang Health Center. The dataset is divided into training data and test data. Then labeled with predictions of experiencing a decrease in nutritional status or not experiencing a decrease in nutritional status using a predetermined machine learning method.

Confusiion Matrix

Confusion Matrix is a test in knowing the performance of machine learning algorithms. This test is based on the confusion matrix table as follows:

Table 1. Confussion Matrix

	Prediction	
Actual	Decreased nutritional status	No decline in nutritional status
Decreased nutritional status	TP	FN
No decline in nutritional status	FP	TN

In table 1 TP is the number of positive data and positive prediction results. TN is the number of positive data and negative prediction results. FP is the amount of negative data and positive prediction results. And TN is the amount of negative data and negative prediction results. Confusion matrix consists of several calculations:

1. Precision is a calculation to determine the number of true positive data from all true positive prediction results. Precision can be done using the equation: Preci-sion = (TP)/(TP+FP).
2. Recall is a calculation to determine the amount of data that is predicted to be true positive from all true positive results. Recall uses the Equation: *Recall = TP / (TP + FN*).
3. F1 Score is a calculation to determine the average of precision and recall comparisons. By using the following equation*: (2 * Recall * Precision) / (Recall + Precision*).
4. Accuracy is a calculation to dete rmine the accuracy of the model in the correct classification. Accuracy can be done with Equation: (TP+TN) / (TP+FP+FN+TN).

RESULTS AND DISCUSSION

Logistic regression methodology was used to predict the detection of child malnutrition status with 657 data availa ble. The variables used in predicting are as follows:

Table 2. Description Variables used to label the dataset

Variables	***Variable Type and Measurement***
jk (Gen-der)	male=0 female=1
Age	Number
BB	Body Weight
TB	Height
BB/U	Children's nutritional status based on weight for age (BB/U): *Very underweight*: less than 3 standard deviation, *Underweight*: less than 3 to less than 2 standard deviation, *Normal weight*: less then 2 to more than 1 standard deviation, *Overweight* risk: more than 1 standard deviation. Normal 0; Less 1; Very Less 2
TB/U	Assessment of TB by age to determine if the child has Normal: 1, Short: 2, Very Short: 3
Status	Nutritional Status based on each index or combination. Good Nutrition: 3, Undernutrition: 6, Malnutrition: 7
Target	0 and 1 If information on whether there is a decrease in nutritional status

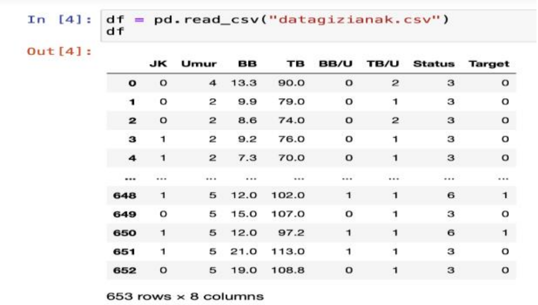
The following is general infor mation about the data set at this stage is the implementation of Machine Learning

DOI: http://dx.doi.org/10.33330/jurteksi.v10i2.2973
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

programming using the python language and several pandas libraries are provided. This stage seen in Image 4 is the result of data import. Data exploration is used to determine the dimensions of the dataset.

Training dataset preparation

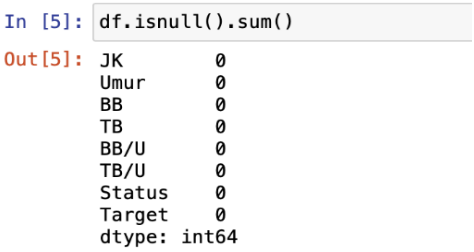
The Data Set is used to load the prediction model for child nutrition data. Which patients are at risk of deteriorating nutritional status. The data that will be used is by dividing the data; Dependent & Independent Data. Dependent Variable: Target; Independent Variables: jk, age, BW, TB, BW/U, TB/U and status.



```
In [4]:  df = pd.read_csv("datagizianak.csv")
         df

Out[4]:
              JK  Umur    BB     TB  BB/U  TB/U  Status  Target
         0     0     4  13.3   90.0     0     2       3       0
         1     0     2   9.9   79.0     0     1       3       0
         2     0     2   8.6   74.0     0     2       3       0
         3     1     2   9.2   76.0     0     1       3       0
         4     1     2   7.3   70.0     0     1       3       0
       ...   ...   ...   ...    ...   ...   ...     ...     ...
       648     1     5  12.0  102.0     1     1       6       1
       649     0     5  15.0  107.0     0     1       3       0
       650     1     5  12.0   97.2     1     1       6       1
       651     1     5  21.0  113.0     1     1       3       0
       652     0     5  19.0  108.8     0     1       3       0

       653 rows × 8 columns
```

Image 3. General dataset information

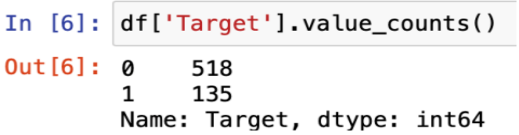
Check the data, whether there is still empty data or not. It can be seen that the image shows all 0 numbers, meaning that there is no null value data. This means that there is no data that contains blanks, making it easier for us to create an accurate model.



```
In [5]:  df.isnull().sum()

Out[5]:  JK        0
         Umur      0
         BB        0
         TB        0
         BB/U      0
         TB/U      0
         Status    0
         Target    0
         dtype: int64
```

Image 4. Check for null data

Next, let's try to see how many children have decreased nutritional status (column "target" =1): it turns out that the statistics show that there are 135 children who have a decrease in nutritional status.



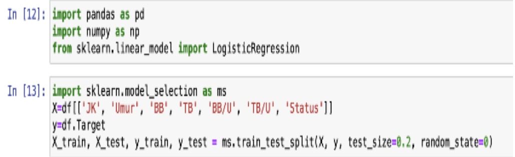
```
In [6]:  df['Target'].value_counts()

Out[6]:  0    518
         1    135
         Name: Target, dtype: int64
```

Image 5. Statistical data on decreased nutritional status

This amount of data is more than half of the dataset, it seems that the dataset is good enough, so there is no data that needs to be cleaned or changed.

Split the dataset into training and test

Furthermore, 80% of the dataset is used as training dataset and the remaining 20% as test dataset. This training dataset is used to build the model, then the test dataset is used to test the model and evaluate the model.



```
In [12]: import pandas as pd
         import numpy as np
         from sklearn.linear_model import LogisticRegression

In [13]: import sklearn.model_selection as ms
         X=df[['JK', 'Umur', 'BB', 'TB', 'BB/U', 'TB/U', 'Status']]
         y=df.Target
         X_train, X_test, y_train, y_test = ms.train_test_split(X, y, test_size=0.2, random_state=0)
```

Image 6. Training data sets

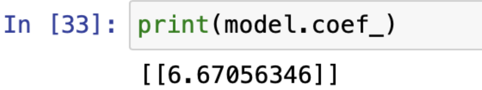
To use the letter X (uppercase) to represent all features, and the letter y (lowercase) to represent the target feature. In the training dataset, the features are stored in a DataFrame named X_train while the target is stored in y_train. Then in the test dataset, the feature will be stored in the DataFrame named X_test and the target is stored in y_test. For training purposes, we will use all existing features (jk, age, bb, tb, bb/u, tb/u and status. Training dataset and test dataset should be chosen randomly. Train_test_split() function to call by inputting the test dataset size such as 0.2 = 20%.

Modelling

At this stage we will use scikit-learn to create a Lo-gistic Regression mo

DOI: http://dx.doi.org/10.33330/jurteksi.v10i2.2973
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

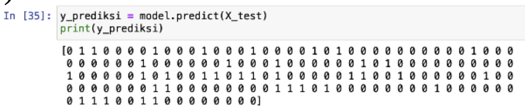
del. By importing the required Logistic Re-gression package in **sklearn.linear _model**. The task is to train the model, the fit() function is called by inputting the training dataset. In this section, the machi ne learning algorithm has successfully found the coefficients for the training da-taset.



```
In [33]: print(model.coef_)

         [[6.67056346]]
```

Image 7. Slope Value

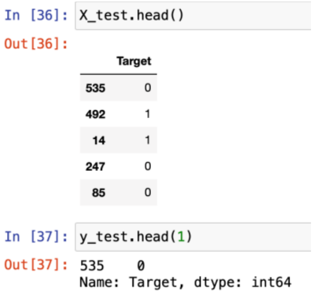
Before entering the model performance measurement section, researchers try to use this model to issue all prediction results on the test dataset, the X_test contains 40 rows of data, so the results will predict (in numbers 0 and 1).



```
In [35]: y_prediksi = model.predict(X_test)
         print(y_prediksi)

         [0 1 1 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0
          0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0
          1 0 0 0 0 0 1 0 1 0 0 1 1 0 1 1 0 1 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 1 0 0
          0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 1 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
          0 1 1 1 0 0 1 1 0 0 0 0 0 0 0 0]
```

Image 8. Datasets Test Measurement Model

The model has successfully made predictions whether or not to experience a decrease in nutritional status. The following shows the test dataset results:



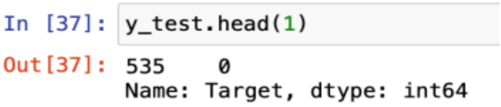
```
In [36]: X_test.head()

Out[36]:
```

	Target
535	0
492	1
14	1
247	0
85	0

```
In [37]: y_test.head(1)

Out[37]: 535    0
         Name: Target, dtype: int64
```

Image 9. test dataset results

The first row at ID 535 shows that there is no decrease in nutritional status. The model predicts no decrease in nutritional status = 0. Let's do a predic tion by comparing the original target in DataFrame y_test? Here are the results:

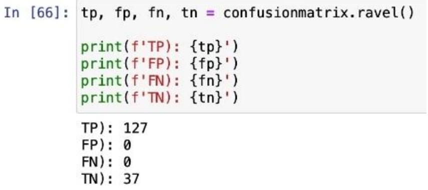


```
In [37]: y_test.head(1)

Out[37]: 535    0
         Name: Target, dtype: int64
```

Image 10. prediction results

Measuring Model Performance

At this stage is the testing stage in seeing the performance of the machine learning algorithm, this test is based on the **confusion matrix** table. Import the metrics package in scikit-learn and input the test dataset.



```
In [66]: tp, fp, fn, tn = confusionmatrix.ravel()

         print(f'TP): {tp}')
         print(f'FP): {fp}')
         print(f'FN): {fn}')
         print(f'TN): {tn}')

         TP): 127
         FP): 0
         FN): 0
         TN): 37
```

Image 11. testing stage

Confusion Matrix can also be visualized, here we do import matplotl ib.pyplot as plt

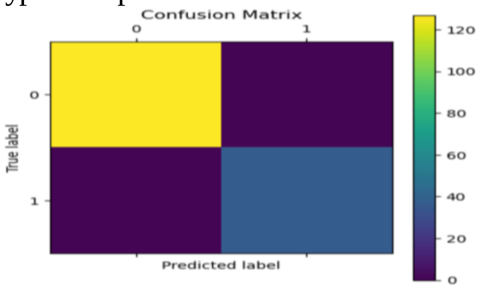
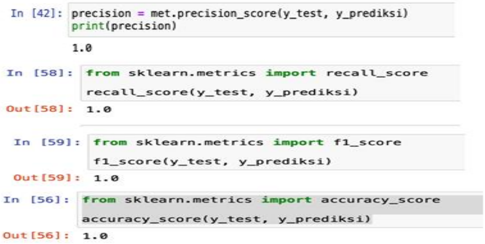


Image 12. Visual representation of Confusion Matrix

The X-axis represents the predicted label, while the Y-axis represents the True Label with values 0 and 1, respectively. Now take measurements by using the score() function to calculate; TP : True Positive, FP : False Positive, TN : True Negative, FN : False Negative
Precision= TP/(FP+TP)
Recall/sensitivity = TP / (FN + TP)
*F-1 Score = (2 * Recall * Precision)/ (Recall + Precision)*.
Accuracy : (TP+TN) / (TP+FP+FN+TN).



```
In [42]: precision = met.precision_score(y_test, y_prediksi)
         print(precision)

         1.0

In [58]: from sklearn.metrics import recall_score
         recall_score(y_test, y_prediksi)
Out[58]: 1.0

In [59]: from sklearn.metrics import f1_score
         f1_score(y_test, y_prediksi)
Out[59]: 1.0

In [56]: from sklearn.metrics import accuracy_score
         accuracy_score(y_test, y_prediksi)
Out[56]: 1.0
```

Image 13. calculation results

The measurement results show a value of 1.0 this value shows a very high value. Next we use the ROC curve (Receiver Operating Characteristic) this curve is tasked with describing the TP (True Positive) value on the Y axis against the FP (false Positive) value on the X axis. AUC (Area Under Curve) task is to show the area under the curve, which is used to indicate the good and bad size of a model. An AUC close to 1 indicates a near-perfect model, while 0.5 is a poor model.

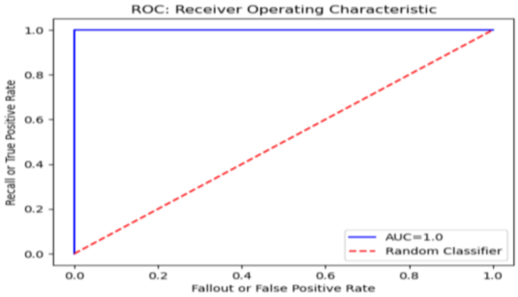
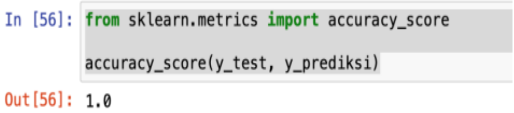


Image 14. ROC curve visualization

In this study shows a perfect size, it can be seen in the results of scikit-learn with the function roc_auc_score() to calculate AUC;



```
In [56]: from sklearn.metrics import accuracy_score

         accuracy_score(y_test, y_prediksi)

Out[56]: 1.0
```

Image 15. scikit-learn results

This means that the very high value is close to 1, which is a sign that this model is overfit.

CONCLUSION

This research implements the use of logistic regression algorithms by applying logistic functions to produce binary or zero and one as a classification determination. The design of the prediction model in the python IDE programming language in detecting the nutritional status of children whose output results are predictions. Then the results show a decrease in nutritional status or not based on the initial data given. Experiments were conducted using a dataset of bengkayang health center data consisting of 657 pediatric patient data, each of which has 7 features (independent variables) and 1 predictor (dependent variable). This test shows perfect performance with precision, recall, F1-score, accuracy, amounting to 100%. Visualization results on the ROC (Receiver Operating Characteristic) curve to describe the TP (True Positive) value on the Y-axis against the FP (false Positive) value on the X-axis show a very high value and is close to 1, this is a sign that this model is overfit.

We recommend that in the preparation of train-ing datasets be measured with training data only without test data and reduce features because not all features in the dataset can be useful for model building, after that you have to do feature selection to increase the accuracy of the model.

REFERENCE

[1] K. K. RI, “Peraturan Menteri Kesehatan RI tentang Penanggulangan Gizi Pada Anak akibat Penyakit,” 2019.

[2] B. Soediono, “INFO DATIN KEMENKES RI Kondisi Pencapaian Program Kesehatan Anak Indonesia,”

J. Chem. Inf. Model., vol. 53, p. 160, 2014.

[3] Badan Penelitian dan Pengembangan Kesehatan, “Survei Status Gizi 2007 - 2020,” no. September, pp. 15–17, 2021.

[4] Erlin, Yulvia Nora Marlim, Junadhi, Laili Suryati, and Nova Agustina, “Deteksi Dini Penyakit Diabetes Menggunakan Machine Learning dengan Algoritma Logistic Regression,” *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 11, no. 2, pp. 88–96, 2022, doi: 10.22146/jnteti.v11i2.3586.

[5] R. R. Pratama, “Analisis Model Machine Learning Terhadap Pengenalan Aktifitas Manusia,” *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 19, no. 2, pp. 302–311, 2020, doi: 10.30812/matrik.v19i2.688.

[6] A. S. Ghareb, A. A. Bakar, and A. R. Hamdan, “Hybrid feature selection based on enhanced genetic algorithm for text categorization,” *Expert Syst. Appl.*, vol. 49, pp. 31–47, 2016, doi: 10.1016/j.eswa.2015.12.004.

[7] H. Hananti and K. Sari, “Perbandingan Metode Support Vector Machine (SVM) dan Artificial Neural Network (ANN) pada Klasifikasi Gizi Balita,” *Semin. Nas. Off. Stat.*, vol. 2021, no. 1, pp. 1036–1043, 2021, doi: 10.34123/semnasoffstat.v2021i1.1014

[8] H. H. Sutarno, R. Latuconsina2, and A. Dinimaharawati3, “Prediksi Stunting Pada Balita Dengan Menggunakan Algoritma Klasifikasi K-Nearest Neighbors Stunting Prediction in Children Using K-Nearest Neighbors Classification Algorithm,” *e-Proceeding Eng.*, vol. 8, no. 5, pp. 6657–6661, 2021.

[9] A. Arman A/Wiwin K, “Sistem Pakar Deteksi Status Gizi dan Psikologi Anak Menggunakan Metode Dempster Shafer,” *Bimasakti*, pp. 1–7, 2014.

[10] R. A. Pamungkas and L. D. Farida, “Implementasi Dempster Shafer Untuk Deteksi Dini Gizi Buruk Pada Balita,” *Pseudocode*, vol. 10, no. 1, pp. 21–29, 2023, doi: 10.33369/pseudocode.10.1.21-28.

[11] N. F. Sahamony, T. Terttiaavini, and H. Rianto, “Analysis of Performance Comparison of Machine Learning Models for Predicting Stunting Risk in Children ’ s Growth,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. April, pp. 413–422, 2024.

[12] T. Rymarczyk, E. Kozłowski, G. Kłosowski, and K. Niderla, “Logistic regression for machine learning in process tomography,” *Sensors (Switzerland)*, vol. 19, no. 15, pp. 1–19, 2019, doi: 10.3390/s19153400.

[13] A. M. Widodo, Y. S. Anggraeni, N. Anwar, A. Ichwani, and B. A. Sekti, “Performansi K-NN, J48, Naive Bayes dan Regresi Logistik sebagai Algoritma Pengklasifikasi Diabetes,” *Pros. SISFOTEK*, vol. 5, no. 1, pp. 27–33, 2021.

[14] A. S. T. Nishadi, “International Journal of Advanced Research and Publications Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterlab,” *Int. J. Adv. Res. Publ.*, vol. 3, no. 8, pp. 69–74, 2019.

[15] Q. R. Cahyani *et al.*, “Prediksi Risiko Penyakit Diabetes menggunakan Algoritma Regresi Logistik Diabetes Risk Prediction using Logistic Regression Algorithm Article Info ABSTRAK,” *JOMLAI J. Mach. Learn. Artif. Intell.*, vol. 1, no. 2, pp. 2828–9099, 2022, doi: 10.55123/jomlai.v1i2.598.