Naive Bayes Classifier (NBC) Application on the Nutritional Status of Adolescents in Medan

Tyas Permatasari[1*] ,Yatty Destani Sandy[2], Caca Pratiwi[3], Kanaya Yori Damanik[4] Agnes Irene Silitonga[5]

{tyaspermata@unimed.ac.id[1], yattysandy@unimed.ac.id[2], cacapratiwi@unimed.ac.id[3], kanayayori@unimed.ac.id[4], agnesirenesilitonga@unimed.ac.id[5]}

[1,2,3,4]Nutrition Study Program, Department of Family Welfare Education, Faculty of Engineering, Universitas Negeri Medan, Medan, Indonesia

[5]Digital Business Program, Faculty of Economy, Universitas Negeri, Medan, Indonesia

Abstract. Adolescent nutrition problems in Indonesia are currently faced with three nutritional burdens (the triple burden), namely stunting, obesity, and micronutrient deficiencies. An unhealthy diet and a sedentary lifestyle are factors that cause nutritional problems in adolescents. Information technology is developing very rapidly and significantly. In the midst of the current COVID-19 pandemic situation, various digital technology innovations have emerged, especially in the health sector. One technology that can make it easier to solve health problems is artificial intelligence (AI), in the form of machine learning. The purpose of this research in general is to see the classification of nutritional status of adolescents using machine learning methods with a classification approach (nave Bayes). The population in this study was teenagers in the city of Medan. The research will be conducted in junior high, high school, and university schools, and its implementation will start from April to August 2022. The sample was taken purposefully, with a total sample of 150 respondents. The research method used is cross-sectional in primary data collection for nutritional status. Furthermore, primary data has been collected through direct measurements and interviews using questionnaires. The data set will be classified and analysed using the naive Bayes method. The tools used in this study are Rapid Manner version 9.0.2. The tools will act to classify the cleaned data and assess the accuracy of the existing data. The results showed that the questionnaire or instrument had been validated by experts and that the respondents resembled the characteristics of the respondents. The average nutritional status in The percentage of adolescents who had an underweight nutritional status was 5.5%, 18.6% were overweight, and 22.8% were obese.

Keywords: adolescent, nutritional status, naive bayes, machine learning

1 Introduction

Adolescents, according to the World Health Organization, are those aged 10 to 19 years, which number up to 1.2 billion in the world and constitute 16% of the world's population [1]. The adolescent population in Indonesia is as high as 45 million, or about 18% or almost one fifth of the total population [2]. Adolescence is a period of rapid growth and development that requires an increase in nutrients [3]. In addition, adolescence is the second window of opportunity after an early childhood that will have an impact on cognitive growth and

ACEIVE 2022, October 20, Medan, Indonesia

DOI 10.4108/eai.20-10-2022.2328884

development and shape future habits [4]. Teenagers are the nation's assets and are part of the acceleration of the SUN Movement (1000 HPK). The fulfilment of nutritional intake will be the foundation for adolescents' achieving ideal nutritional status and optimal productivity.

Nutritional problems faced by adolescents in Indonesia today are stunting, wasting, and obesity [5]. Rickshas data shows that 26.3% of adolescents aged 13–18 years are stunted or short, 9% are wasting or thin, 16% are obese, and a quarter of adolescent girls are anaemic [6]. An unhealthy diet and a sedentary lifestyle are factors that cause nutritional problems in adolescents. Research shows that food consumption that is less diverse and in small quantities will cause a lack of energy intake and can cause adolescents to experience wasting and stunting [7]. Meanwhile, on the other hand, more nutrition in adolescents occurs due to the habit of eating fast food that contains high fat, foods high in sugar and a lack of physical activity carried out [8].

Nutritional status and food consumption have a role in nutritional problems, which will be related to increased susceptibility to diseases, especially the risk of non-communicable diseases such as diabetes, heart disease, low quality of life, and other diseases that can cause death and cause disruption of learning concentration [9] [10]. Therefore, adolescents need to know their nutritional status to be able to regulate their diet so that they achieve an ideal nutritional status.

Information technology is developing very rapidly and significantly. During the current COVID-19 pandemic situation, various digital technology innovations have emerged, especially in the health sector.[11]. One of the technologies that can make it easier to solve problems with artificial intelligence methods is artificial intelligence (AI), one of which is machine learning.[12]. The development of AI in the world of health needs to be done. Some research in the health sector mostly uses logistic regression methods, while machine learning methods with a Naive Bayes approach are still not widely used.[13]. This study was conducted to determine the classification of nutritional status in adolescents quickly, precisely and accurately using Naive Bayes so that interventions can be carried out following their nutritional status to accelerate the reduction of stunting and obesity rates in Indonesia.

2 Research Method

2.1 Data Source

The research started from collecting primary data through cross sectional method. One hundred forty five participants were categorized as beginner adolescent (Junior high school students) with 51 participants, middle adolescent (Senior high school students) with 49 participants, and latest adolescent aged (University students) with 45 participants. Data were collected by measuring weight and height and interviewing using an instrument that had been made and previously validated. Furthermore, the data are classified with Naïve Bayes.

2.2 Procedures

The analysis mechanism used to obtain the desired objectives in this study is as follows:

1) First, collect the data, which is then entered into Microsoft Excel.
2) Conducting the pre processing of data to clean up incomplete information from the data Afterwards, the data can be processed for further use.
3) Identification of data with descriptive analysis

4) Data obtained from pre-processing is entered into the RStudio program database.
5) After that, Classification the data with the Naïve Bayes method, through the following steps:
   a. Determine the value of K, set K=10
   b. The dataset was divided into a number of K partitions randomly, in this case divided into 10 partitions.
   c. Divide the data partition into testing data and training data alternately a number of K, which is 10 times.
   d. Classification the data with Naïve Bayes Classifier
   e. Iterations were carried out a number of K-experiments, where each experiment used the K-the partition data as testing data. This means that 10 experiments were carried out and each partition was used as testing data consecutively.
   f. Data that is not used as testing data is used as training data.

6) Predictions for new data using the previous model that has been determined as the best model.
7) Last, Interpretation and Conclusion from the

3 Results and Discussion

3.1 Pre processing Data

The data used for research is data that is not ready to be processed. The data still has unused variables, missing values, noise, and data types that are not in accordance with the research carried out, so it is necessary to process the data. The variables that will be used are the variables of sex, height, weight, Z-Score result category, and adolescent age, which will be calculated with the measurement date based on the date of birth of the adolescent. The data is then processed based on the adjustments needed through data cleaning and data transformation. The following is a layout of the data after processing, which can be seen in Table 1;

Table 1. Layout Data Preprocessing

Age (month)	Sex	Weight (kg)	Height (cm)	Nutritional Status
14	Male	53.7	165	Normal
15	Male	49.2	165	Normal
15	Female	53.4	143.1	Overweight
14	Male	60.5	163.5	Overweight
15	Female	50.3	154	Normal
15	Female	47.1	164	Normal
15	Male	43.3	164	Normal
14	Female	64.3	156	Overweight
…	…	…	…	….
15	Female	39.2	145	Normal

3.2 Descriptive Analysis

Based on Table 2, it can be seen in the age column that adolescent age spreads on average from 12 years to 24 years, which means that it is spread in the early, mid, and late adolescence age groups. The size of the spread of wart data is also spread between the first

quantile of 13 years and the third quantile of 22 years. Variable weight and height vary widely from lowest to highest having a fairly long range.

Table 2 Descriptive Analysis of Continuous Data

	Sex (Year)	Weight (kg)	Height (cm)
Minimum	12	32.6	140
1st Qu	13	44.3	151.4
Median	15	51.2	156
Mean	16	54	155.6
3rd Qu.	22	73.3	164.4
Maximum	24	100	180

The results showed that the age of adolescents in this study already represented the average category of early adolescence (11–14 years), middle adolescence (15–17 years), and late adolescence (18–21 years) [14]. This can represent all categories of teenagers in Medan City. The average body weight of adolescents in the study was 54; this is in accordance with the average age in the Recommended Dietary Allowances (RDA), which is 36–60 kg in men and 38–52 kg in women [15]. The results of the study showed an average TB of 155 when compared to the 2019 RDA in adolescents, which is an average of 145-168 cm in men and 147-159 cm in women [15].

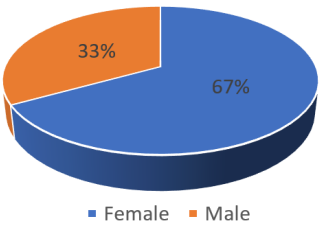

Figure 1. Adolescent sex distribution

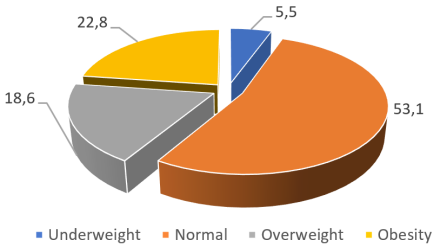

Figure 2. Nutritional Status in Adolescent in Medan City

Based on Figure 1, it can be seen that the tendency of adolescents is greater than that of men. Toddlers with male sex were 605 out of the 1037 population, or 58.34%, while toddlers with female sex amounted to 432 out of the 1037 population, or 41.66%. Based on figure 2, it can be seen that overall, the majority of adolescents' nutritional status falls into the normal category, but there are adolescents who are underweight, overweight, or obese. The percentage of adolescents with an underweight nutritional status was 5.5% of the total sample, 18.6% were overweight, and 22.8% were obese. This is in line with several studies in adolescents in Kendal that showed similar results with the distribution of nutritional status in underweight, normal, overweight, and obesity [16].

Adolescent health is built on a foundation of adequate nutrition, healthy eating, and activity routines. [17]. Good or optimal nutritional status occurs when the body obtains enough nutrients that are used efficiently, thus allowing physical growth, brain development, and creating health. [18]. Malnutrition status in young women will increase the risk of diseases, especially infectious diseases, and inhibit the growth and development of the body,

which will determine future health conditions. Overweight and obesity in adolescents can lead to health problems in adulthood, such as impaired respiratory function, the risk of degenerative diseases, and cardiovascular diseases. A window of opportunity exists during adolescence to ensure a successful transition to maturity. The nutritional status and eating habits that are developed during this stage of life have significant effects on both the health and wellbeing of the adolescent and future generations [14][19].

3.3 Classification

After cleaning the data, the data that already has complete information for each attribute is selected. This selection is done to group attributes according to the required information. The technique of taking the subject is done by purposive sampling. How to take adolescent subjects using the sample mean estimation formula (Equation 1):

$$n = \frac{Z_{\frac{\alpha}{2}}^2 x\sigma^2}{d^2} \quad (1)$$

$$n = \frac{1.96^2\ x0.6^2}{0.1^2}$$

$$n = 139$$

where:

σ = standard deviation of assumptions 0.6
d = desired precision of 0.1 (set)
$Z_{a/2}$ = 95% significance level (set)

The level of significance that the researcher wants is p 0.05, and a precision of 0.1 can represent the research subject. Based on the above formula, there are 138 adolescents. Subjects in the study were added 10% of the number of subjects to anticipate dropping out, so that the number of subjects became 150 children. Furthermore, the sample is divided into 3 age categories, namely, early teens (50 children), mid-teens (50 children), and late teens (50 children), so that it can represent the entire group of adolescent age levels.

3.4 Classification Performance

Evaluation was carried out for the selection of the best dataset distribution method and classification method as seen through classification performance measures. The classification performance measure used in this study takes into account the confusion matrix. The confusion matrix is a useful tool for analyzing how well or how accurately the classification method can recognize objects of observation from different classes. Table 3 is a confusion matrix for binary classification.

Table 3 Confusion Matrix

Confusion matrix		Actual class		Total
		Yes	No	
Prediction class	Yes	TP	FP	P
	No	FN	TN	N
Total		P	N	

Some of the classification performance measures that can be obtained from the confusion matrix are equations 1, 2,3, 4.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (2)$$

$$Sensitivy = \frac{TP}{TP+FN} \quad (3)$$

$$Specifity = \frac{TN}{TN+FP} \quad (4)$$

$$G - mean = Sensitivy\ x\ Specifity \quad (5)$$

Accuracy is the most commonly used method to assess classification performance. If the level of accuracy is high, but the sensitivity or specificity is low, then the classification can be said to be not good. In addition, the case of imbalanced data can be overcome by using the geometric mean (G-mean) and receiver operating characteristic (ROC) curves because they do not depend on the distribution of observations between classes. The single value that can be used to measure classification performance on the ROC curve is the area under the ROC curve (AUC).

The stages of data analysis are: data sampling with purposive sampling, data division into training data (20 out of 150) and testing data (10 out of 20), classification modeling using machine learning methods, namely naive Bayes, and classification performance evaluation based on the machine learning method used. Data processing is carried out using Rapid Manner 9.0.2 tools. Naïve Bayes classification which refers to Bayes' theorem can be seen in equation (6)

$$P(X) = \frac{p(C_i)p(C_i)}{p(x)}$$

where:
X = data with unknown class
C = a specific class of data
P ($C|X$) = conditional probability of class based on condition (posterior probability)
P (C) = class probability (prior probability)
P ($X|C$) = the conditional probability of based on the condition of class , referred to as likelihood
P (X) = probability data
i : class

3.5 Prediction

From the classification that has been carried out using the Naïve Bayes Classifier model, it is found that the best accuracy model is in the 10th iteration, so predictions will be made with training data using the 10th iteration model as a prediction model. The data to be predicted is data that has never been used before. The following is the prediction results can be seen in table 4.

Table 4. Prediction Data Layout

ID	Body Weight	Body Height	BMI	Category
1	53.7	165	0.2	Normal
2	75.8	158.9	2.34	Obesity
3	41.9	166.4	-2.93	Underweight
4	71.6	154	2.35	Obesity
5	55.4	151	1.08	Overweight
6	91.1	164	2.96	Obesity

The Naïve Bayes Classifier model can be used to make predictions with the results in table 4 and can also be used to make other predictions using the same variables.

4. Conclusions

In this study, the questionnaire used in primary data collection has been validated by an expert validator, namely a nutritionist. In addition, the instrument was also validated by respondents with the same characteristics. The results obtained from the analysis using Microsoft Excel are on average valid, and there are two invalid questions. Coding has also been carried out on several variables that have been obtained.

Based on cleaning data from a total of 145 respondents, the distribution of adolescents includes 51 respondents belonging to early adolescents, 49 respondents belonging to middle adolescents, and 45 respondents belonging to late adolescents. The percentage of adolescents with an underweight nutritional status was 5.5% of the total sample, 18.6% were overweight, and 22.8% were obese.

Acknowledgements

This study was funded by a research grant from Universitas Negeri Medan. Special thanks go to all respondents as well as enumerators from the bachelor's degree in applied nutrition at Universitas Negeri Medan.

References

[1] World Health Organization (WHO). (2014). *Health for the World's Adolescents: A Second Chance in the Second Decade*. Geneva, World Health Organization Department of Noncommunicable disease surveillance.
[2] Badan Pusat Statistik (BPS) Indonesia. (2016). *Statistik Kesejahteraan Rakyat.*
[3] Sandy, Tamtomo dan Indarto. (2020). Hubungan Berat Badan Dengan Kejadian Anemia Remaja Putri di Kabupaten Boyolali. *Jurnal Dunia Gizi*, Vol. 3, No. 2, Hal 94-98.
[4] UNICEF. (2018). *Programme Guidance for the Second Decade: Programming with and for adolescents.* Programme Division: New York.
[5] Kementerian Kesehatan Indonesia. (2020). *Gizi Saat Remaja Tentukan Kualitas Keturunan*. https://www.kemkes.go.id. Ministry of Health Indonesia.
[6] Riset Kesehatan Dasar Indonesia. (2018*). Badan Penelitian dan Pengembangan Kesehatan Kementerian RI tahun 2018.*
[7] Insani Hurry. (2019). Analisis Konsumsi Pangan Remaja dalam Sudut Pandang Sosiologi. *Jurnal Pendidikan Sosiologi*. Vol 9, No. 2, Hal 739-75.

[8] Swamilaksita, Sa'pang. (2017). Keragaman Konsumsi Pangan Dan Densitas Gizi Pada Remaja Obesitas Dan Non Obesitas. *Jurnal Nutrire Diaita*, Vol 9, No. 2, hal 44-50.
[9] Center for Disease Control and Prevention (CDC). (2015). *The Health Effects of Overweight and Obesity.*
[10] Marine Denov, Adiningsih. (2015). Perbedaan Pola Konsumsi Dan Status Gizi Antara Remaja Dengan Orang Tua Diabetes Melitus (DM) Dan Non DM. *Jurnal Media Gizi Indonesia*. Vol. 10, No. 2, hlm. 179–183.
[11] Astuti Fitri. (2021). Pemanfaatan Teknologi Artificial Intelligence untuk Penguatan Kesehatan dan Pemulihan Ekonomi Nasional. *Jurnal Sistem Cerdas*, Vol 04, No. 01:Hal 25-34.
[12] Zubair, Muksin. (2018). Penerapan Metode Naive Bayes Untuk Klasifikasi Status Gizi (Studi Kasus Di Klinik Bromo Malang). *Seminar Nasional Sistem Informasi*: Fakultas Teknologi Informasi – UNMER Malang.
[13] Yuliati, Sihombing. (2021). Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia. *Jurnal Manajemen, Teknik Informatika, dan Rekayasa*, Vol. 20, No. 2:Hal 417-426
[14] J.E. Brown. (2017). *Nutrition Through the Life Cycle, 6e (Cengage Learning,* Buston.
[15] AKG (2019). *Angka Kecukupan Gizi Yang Dianjurkan Untuk Masyarakat Indonesia. Peraturan Kementrian Kesehatan Republik Indonesia Nomor 28 Tahun 2019.*
[16] Putri MP, Dary, Manglik G. (2022). Asupan Protein, Zat Besi Dan Status Gizi Pada Remaja Putri. *Journal of Nutrition College*, 11(1).
[17] Begum A, Sharmin KN, Hossain MA, Yeasmin N, Ahmed T. Nutritional status of adolescent girls in a rural area of Bangladesh: A Cross Sectional Study. *Bangladesh J. Sci. Ind. Res.* 2017; 52(3), 221-228. https://doi.org/10.3329/bjsir.v52i3.34158
[18] Asmare B, Taddele M, Berihun S, Wagnew F.(2018). *Nutritional status and correlation with academic performance amongprimary school children,* Northwest Ethiopia. BMC Res Notes. 2018; 11(1): 1-6.
[19] Rah, Jee & Chalasani, Satvika & Oddo, Vanessa & Sethi, Vani. (2017). *Adolescent Health and Nutrition*. 10.1007/978-3-319-43739-2_25.