Comparative Analysis of Machine Learning Algorithms for classification about Stunting Genesis

*Agus Byna
{agusbyna@unism.ac.id}

Universitas Sari Mulia,info@unism.ac.id

Abstract. Background The use of machine learning is very much needed for health experts as data and information processing to make it easier to analyze automatically. To produce accuracy in solving problems. Application of machine learning with comparative three algorithms to solve stunting problems. Because toddlers in Indonesia are still high, especially at age 2 -3 years. Seen from many factors that are at risk of causing stunting. The instrument is needed in Machine Learning. The goal (1). In addition to providing knowledge in the field of Informatics. It's also useful for health experts in managing data in making decisions, as to facilitate analysis automatically. (2). Can reduce the impact on the incidence of stunting. Methods Comparison of three algorithms in the classification of the results. That was compared yielded an accuracy of 86% AUC 0.85 for the Decision Tree algorithm with a diagnosis level of Good classification, Algorithm KNN with an accuracy of 58.7% AUC 0.57 fail classification, Algorithm Naïve Bayes with 55% AUC accuracy 0.51, using 13 stunting data variables.

Keywords: Genesis Stunting, Decision Tree, KNN, Naïve Bayes, Machine Learning.

1 Introduction

Where the health sector is currently making changes in the collection of data and information about patients individually so that the volume of data produced becomes very large so that health experts will be difficult to analyze. Having machine learning can provide a solution to the problem patterns found. With an easy way to analyze automatically.

Stunting events are chronic nutritional problems that adversely affect the physical growth characterized by a decrease in growth speed so that it requires a medical handler [2].

Based on data WHO in 2016, in southeast Asia the prevalence of stunting toddlers reaches 33.8%. In the year 2011, Indonesia ranked five out of 81 countries with the largest number of stunting children in the world reaching 7,547,000 children. Indonesia has been reported to have an impact on stunting events with more stunting numbers than some countries in the African continent, such as Ethiopia, Kenya, Uganda, and Sudan. During the year 2007-2011, Indonesian reported having children with moderate weight, low body weight, and excess weight, each reaching 13%, 18%, and 14% respectively. In the year 2012, the mortality rate of children under five years in Indonesia reached 152,000 [5].

The prevalence of stunting infants in Indonesia is still volatile since 2007-2017. The prevalence of stunting infants in Indonesia in 2007 is 36.8%, 2010%, 35.6%, 2013%, and 37.2%.

NS-UNISM 2019, November 23, Banjarmasin, Indonesia

DOI 10.4108/eai.23-11-2019.2298349

2.5 According to WHO, the prevalence of short infants becomes a matter of public health if the PR significance is 20% or more. Because the percentage of stunting events in Indonesia is still very high and the health problems of children and toddlers should be addressed. In several Southeast Asian countries, is also highest compared to Myanmar (35%), Vietnam (23%), Malaysia (17%), Thailand (16%), and Singapore (4%) [6].

Many researchers are using the health Machine Learning algorithm to analyze the solutions in the amount of data and information available.

The accuracy result obtained using C 4.5 decision tree algorithm, RF with CHAID using pruning = 3 results in better accuracy that is in Ngaka 64% and 62.67%, while the use of pre pruning yields more accuracy Low [3].

Subsequent this research conducted about hepatitis patients with the C 4.5 algorithm that resulted in 77.29% accuracy and AUC 0.846 value and Naive Bayes algorithm resulted in 83.71% accuracy and AUC value of 0.812 resulted in a comparison of both of these algorithms accurately in the classification of hepatitis disease, with the highest value of its accuracy using Naïve Bayes algorithm (Septiani,2017). The use of the algorithm Naive Bayes in predicting the results of the accuracy value Naïve Bayes 89.08% [9].

Support Vector Machine is one of the research Machine Learning algorithms that Agus added a selection of the accurate variable feature by using the Backward Elimination which resulted in 81.62% and AUC 0921 in predicting the incident Stunting.[1].

So, the authors do a comparison with some Machine Learning algorithms in the classification of stunting events.

2 Research Methods

Stunting is defined as an indicator of the nutritional status of TB/U equal to or less than minus two standard deviations (-2 SD) below the average of the standard. Stunting is a short, very short body condition that exceeds the deficit-2 elementary school under the median length or height of the body. [16].

The Machine Learning algorithm in comparison is K-Nearest Neighbour is a method to classify the object based on data training using the closest distance or resemblance to the object, the same features are counted for the test data (whose classification is not known). The distance from this new vector to learning data vectors is calculated and is taken a number of the closest K. The new point of classification is predicted to include the most classifications of the point.[10].

In Figure 1 explained about Knn algortima which is a primitive form of machine learning. usually often called 'lazy learning' due to one induction occurring during the process [36] [36]. Figure 1 illustrates a simple example of classification. Using the KNN algorithm, with the k value set to 3. So that the 3 closest training data will point to the new points q1 and q2. This value will determine the class of some of these points. the majority of votes in this example q1 is made by a group with red points and q2 together with the blues.. Thus the KNN method may

be separated into two stages; first, for attribute or dimension r (the variable, in our case acceleration in g) the Euclidean distance, d, between new data point xi and training data point xj is calculated by the formula given in Mitchel [11].

$$\mathrm{d}(x_i,x_j)=\sqrt{\sum_{r=1}^{n}\left(a_r(x_i)-a_r(x_j)\right)^2} \quad (1)$$

By choosing the number of k values ??with the smallest Euclidean distance. The use of Euclidean distances is a convention with KNN. Other distance metrics can be used [12]. If the closest k value is from two classes a and b, class a will be selected if the number of points owned by class a is greater than class b, or na> nb. The KNN algorithm is present in the python package and provides the prob output value, which has the proportion of the closest value k in the training set included in the winning class. (2) where nwc shows the number of points in the winning class.

$$prob=\frac{n_{wc}}{k} \quad (2)$$

In increasing accuracy, threshold filters can be applied to prob values. To produce a minimum threshold classification. Then made by KNN which does not exceed this threshold is discarded. and will make the classification not exceed this threshold discarded. The field of machine learning algorithms is often evaluated through the construction of a confusion matrix [11].

The Decision Tree algorithm is one algorithm for classifying data. A decision tree model is a tree consisting of a root node, an internal node, and a terminal node. While the node's root and internal nodes are variables/features, the terminal node is the class label. Classifying, a data query will trace the root node and internal nodes until it reaches the terminal node. Labeling data class queries based on labels in internal nodes [10].

The classification that performs recursive partitions for the sample space within the decision tree algorithm consists of some typical internal nodes and leaf nodes. Each internal node is called a decision node representing a test on an attribute or a subset of attributes, and each edge is labeled with a specific value or range of value of the input attributes. Internal nodes are related to edges then divide the instance space into two or more. So partitioning each leaf node as a terminal node tree with class labels.

The illustration in Figure 1 shows the basic decision-making tree. In the circle as a decision node then squared as a leaf node. this example has three separate gender attributes of 3 criteria along with 2 class labels, i.e., YES and NO. Each path will be a root node to a leaf node forms a classification rule[10].

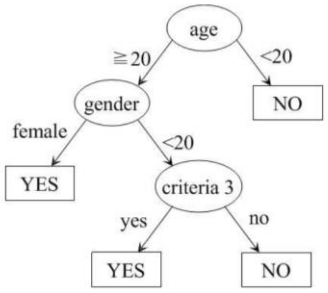

Fig1. Illustration of Decision Tree

To build a decision tree is to give a set of training data. Then apply the measurement function in all attributes to find the best cleat attribute. After that, the attribute occurs with a predefined separation. Using the instance space has been partitions into sections. Each partition, if the example of all the training data belongs to one class then the process ends. Otherwise, the separation process will be recursive. This process is performed until the entire partition is assigned to the same class. After completion, the decision tree is constructed. Quickly and easily generate classification rules. So, it can be used for classification of new instances with unknown class labels [10].

Naive Bayes classifier (NBC) is one of the algorithms with a simple probabilistic-based predictive technique. Through the application of the Bayes theorems (or rules of Bayes) resulted in a strong (naïve) assumption. Simple use of a probabilistic classifier that calculates a series of probabilities with frequencies. Then combined with the value in the specified set of data. By utilizing the Bayes theorem, it generates assumptions to all attributes to be independent given the value of class variables. This assumption is rare in real-world applications. Thus, the characterization of the s is naïve and the algorithm tends to perform well and learn quickly in a variety of supervised classification issues [13].

Naïve Bayesian classifier is based on Bayes' theorem and the theorem of total probability. The probability that a document d with vector x = < x1,...,xn> belongs to hypothesis his Here,

$$P(h1|xi) = \frac{P(xi|h1)P(h1)}{P(xi|h1)P(h1) + P(xi|h2)P(h2)} \quad (1)$$

P(h1|xi) is the posterior probability, while P(h1) is the prior probability associated with hypothesis h1. For m different hypotheses, we have

$$P(xi) = \sum_{j=1}^{n} P(xi|hj)P(hj) \quad (2)$$

Thus, we have

$$P(h1|xi) = \frac{P(xi|h1)P(h1)}{P(xi)} \quad (3)$$

A confusion matrix illustrates the accuracy of the solution to a classification problem. Given n classes, a confusion matrix is a m x n matrix, where Ci,j indicates the number of tuples from D that were assign to class Ci,j but where the correct class is Ci. Obviously the best solution will have only zero values outside the diagonal [14].

A confusion matrix contains information about actual and predicted classifications done by classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier. The entries in the confusion matrix have the following meaning in the context of our study [15]:

1. a is the number of correct predictions that an instance is negative,
2. b is the number of incorrect predictions that an instance is positive,
3. c is the number of incorrect of predictions that an instance negative, and
4. d is the number of correct predictions that an instances positive [16].

Some standards and terms:

1.True positive (TP): If the outcome from a prediction is p and the actual value is also p, then it is called a true positive.

2.False positive (FP): However if the actual value is n then it is said to be a false positive.

3.Precision and recall: Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved.

Both precision and recall are therefore based on an understanding and measure of relevance. Precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. The recall is nothing but the true positive rate for the class.

Of the three algorithms, researchers do a comparison which is a better level of accuracy. Data used that about Stunting (infant/child short-body), with the amount of sample data taken is 457 data. From that data shows 43% are patients experiencing Stunting and 57% are patients who are not experiencing the stunting event. Data management that has relevant attributes according to use is 14 attributes/variable is gender, infant age (month), Weight loss (gram), height (cm), breastfeeding, mother's age, mother's education, mother's height, mother's income (months), Dad's age, Father's education, father's height, father's income, and as a label are Stunting.

	jk	usia	bb_bayi	tb_bayi	...	pd_ayah	tb_ayah	pdp	stunting
0	1	36	2700	95.0	...	2	160	1200000	0
1	2	22	2800	80.0	...	1	170	1500000	1
2	2	16	3000	80.5	...	3	170	1000000	0
3	1	21	3000	90.0	...	3	160	800000	0
4	1	16	2900	70.0	...	3	160	1500000	1

Fig. 2. Genesis stunting of the dataset.

3. Results and Discussion

The results of the trial and machine learning algorithm of KNN, Naïve Bayes, Decision Tree The research uses Python programming language 3.7 with Scikit-Learn that works for the management of the Machine Learning dataset. Displays a plot about stunting events

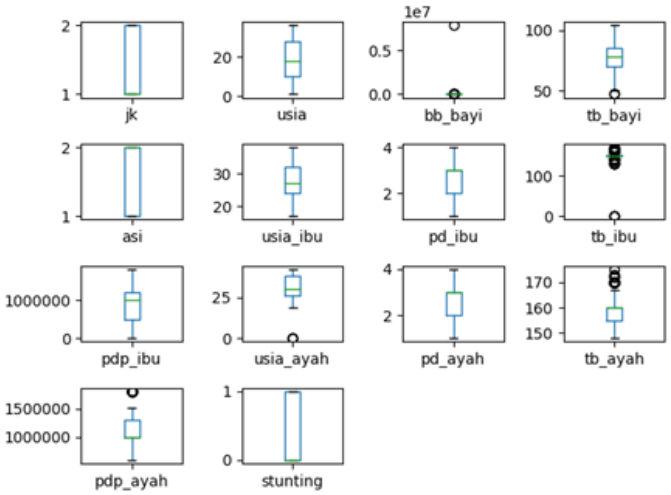

Fig. 3. Event Stunting Plot

This research explained that there is a meaningful relationship, birth weight, maternal education level, and family income level with stunting events. The maternal education level has the most dominant relationship with the stunting event [7].

Test result using tenfold validation. KNN algorithm for the value of n neighbors = 2, then the accuracy that is getting is 58.7%. With its AUC value, 0.57 for Roc can be seen in the image below.

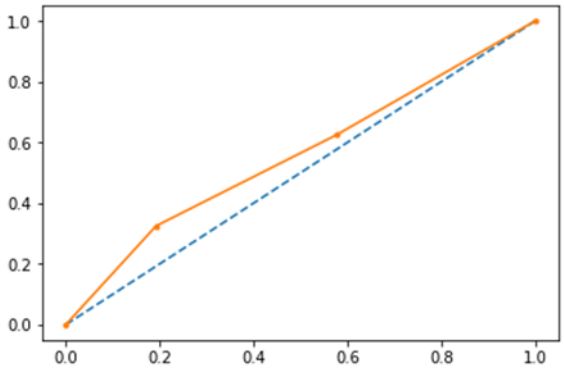

Fig. 4. Roc KNN on Stunting event subsequent test

The results are Naïve Bayes algorithm acquired accuracy is 55%. With its AUC value 0.51.

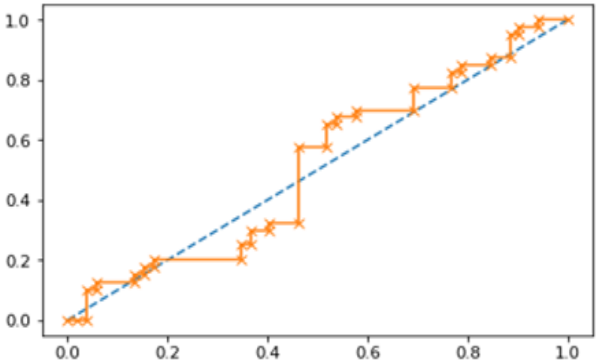

Fig. 5. Roc NB on Stunting events

The last test result with the Decision Tree algorithm gained accuracy is 86%. With its AUC value 0.85

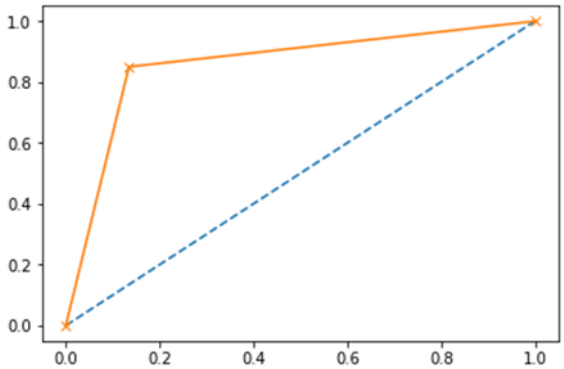

Fig 6. Roc Decision Tree on Stunting events

Discussion

The result of the test obtained the difference in accuracy between the KNN algorithm with Naïve Bayes is 3.7% with AUC values is 0.06, as well as the Decision Tree algorithm with Naïve Bayes, is 31% AUC 0.34 compared with KNN of 26.3% AUC 0.28.

Table 1. Table Comparison 3 algorithm

Algorithm	Accuracy	AUC
KNN	58,7%	0.57
Naïve Bayes	55%	0.51
Decision Tree	86%	0.85

4 Conclusion

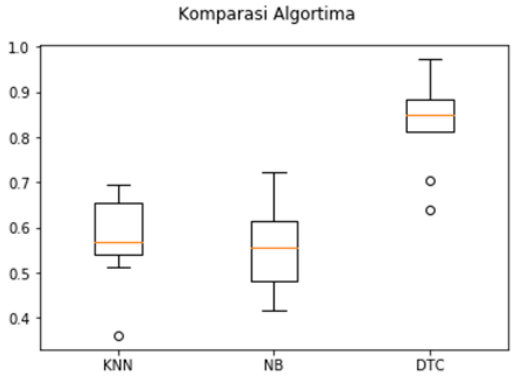

Figure 7. Comparison of 3 Algoritma

The results of the research in comparison to 3 KNN algorithms, Naïve Bayes, and Decision Tree were obtained for an accuracy value of 86%. And evaluated using the ROC curve, the AUC value based on the ROC curve for the Decision Tree algorithm is worth 0.85 with the diagnostic rate of Good classification, while the KNN algorithm is worth 0.57 with the diagnostic rate of file classification, then Naïve Bayes ' algorithm with the diagnostic rate of file classification, among the three algorithms the Decision Tree has a good classification, and the other two have classification files that have a very far difference.

5 Acknowledgments

The author wants to acknowledge the rector of the University of Sari Mulia and head of research institutes to support this study through the University of Sari Mulia Faculty of Beginners Research under the research scheme of Applied Technology and hospital Sari Mulia for cooperation and authority.

6 References

[1] Byna, Agus, and Fadhiyah Noor Anisa. "Backward Elimination Untuk Meningkatkan Akurasi Kejadian Stunting Dengan Analisis Algortima Support Vector Machine." DINAMIKA KESEHATAN JURNAL KEBIDANAN DAN KEPERAWATAN 9.2 (2018): 217-225.

[2] Destiadi, Alfian, Triska Susila Nindya, and Sri Sumarmi. "Frekuensi Kunjungan Posyandu dan Riwayat Kenaikan Berat Badan sebagai Faktor Risiko Kejadian Stunting pada Anak Usia 3–5 Tahun." Media Gizi Indonesia 10.1 (2016): 71-75.

[3] Mambang, Mambang, and Agus Byna. "ANALISIS PERBANDINGAN ALGORITMA C. 45, RANDOM FOREST DENGAN CHAID DECISION TREE UNTUK KLASIFIKASI TINGKAT KECEMASAN IBU HAMIL." SEMNASTEKNOMEDIA ONLINE 5.1 (2017): 2-1.

[4] Nofriansyah, Dicky, S. Kom, and M. Kom. Konsep Data Mining Vs Sistem Pendukung Keputusan. Deepublish, 2015.

[5] Ohyver, Margaretha, et al. "Logistic Regression and Growth Charts to Determine Children Nutritional and Stunting Status: A Review." Procedia computer science 116 (2017): 232-241.

[6] RI, Kementerian Kesehatan. "Hasil pemantauan status gizi (PSG) 2017." Jakarta: Kementerian Kesehatan RI (2018).

[7] Setiawan, Eko, Rizanda Machmud, and Masrul Masrul. "Faktor-Faktor yang Berhubungan dengan Kejadian Stunting pada Anak Usia 24-59 Bulan di Wilayah Kerja Puskesmas Andalas Kecamatan Padang Timur Kota Padang Tahun 2018." Jurnal Kesehatan Andalas 7.2 (2018): 275-284.

[8] Septiani, Wisti Dwi. "Komparasi Metode Klasifikasi Data Mining Algoritma C4. 5 Dan Naive Bayes Untuk Prediksi Penyakit Hepatitis." Jurnal Pilar Nusa Mandiri 13.1 (2017): 76-84.

[9] Sulaehani, Ruhmi. "Prediksi Keputusan Klien Telemarketing untuk Deposito Pada Bank Menggunakan Algoritma Naive Bayes Berbasis Backward Elimination." ILKOM Jurnal Ilmiah 8.3 (2016): 182-189.

[10] Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, (2016).

[11] Stehman, Stephen V. "Selecting and interpreting measures of thematic classification accuracy." Remote sensing of Environment 62.1 (1997): 77-89.

[12] Short, R., and Keinosuke Fukunaga. "The optimal distance measure for nearest neighbor classification." IEEE transactions on Information Theory 27.5 (1981): 622-627.

[13] Dimitoglou, George, James A. Adams, and Carol M. Jim. "Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability." arXiv preprint arXiv:1206.1121 (2012).

[14] Dunham, Margaret H. Data mining: Introductory and advanced topics. Pearson Education India,(2006).

[15] Anshul Goyal, Rajni Mehta, "Performance Comparison of Naïve Bayes and J48 Classification Algorithms", IJAER, Vol. 7, No. 11, (2012), pp.

[16] Manary, M. J., and N. W. Solomons. "Gizi Kesehatan Masyarakat, Gizi dan Perkembangan Anak." Jakarta: Buku Kedokteran ECG (2009).

[16]Xiang yang Li, Nong Ye, "A Supervised Clustering and Classification Algorithm for Mining Data With Mixed Variables", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, Vol. 36, No. 2, (2006), pp. 396-406