Comparative Analysis of Classification Methods of KNN and *Naïve Bayes* to Determine Stress Level of Junior High School Students

Y C Tapidingan [*1], **D Paseru** [2]

[1,2]Program of Informatics Engineering, Faculty of Engineering, Universitas Katolik De La Salle Manado, Kairagi I Kombos, Manado

E-mail: juahkristoff@gmail.com [1], dpaseru@unikadelasalle.ac.id [2]

Submitted: 12 September 2019, revised: 31 December 2019, accepted: 1 January 2020

Abstrak. Stres pada umumnya didefinisikan sebagai keadaan dimana seseorang mengalami gangguan mental, ini merupakan respon terhadap kesulitan yang dialami. Siswa SMP cenderung tidak menyadari stres yang mereka hadapi. Penelitian ini bertujuan untuk membandingkan dua metode klasifikasi KNN dan Naïve Bayes untuk menentukan tingkat stres. Data penelitian ini dikumpulkan dari 254 responden dari SMP Katolik Don Bosco Bitung. Tes validasi *k-cross* dan pemisahan persentase dari data menunjukkan bahwa metode Naïve Bayes lebih baik dari metode KNN. Dengan k = 3, akurasi KNN mencapai 86,61% pada tertinggi dan Naïve Bayes mencapai 87,40%. Sementara itu, berdasarkan hasil uji persentase split, rata-rata akurasi Naïve Bayes lebih tinggi dari KNN dengan persentase 88,31%. Selain itu, untuk presisi dan daya ingat, Naïve Bayes lebih tinggi dari KNN dengan 88,30% dan 87,40% dilihat dari validasi *k-cross*.

Kata kunci: stress; Naïve Bayes; KNN; akurasi; perbandingan

Abstract. Stress is generally defined as a state where someone is mentally disturbed as the response to the adversity that he/she experiences. Junior High School students usually are not aware of the stress that they encounter. This research aims to compare two classification methods of KNN and Naïve Bayes to determine stress level. The data of this research were gathered from 254 respondents from Catholic Junior High School of Don Bosco Bitung. The tests of k-cross validation and percentage split from the data showed that Naïve Bayes method excelled KNN method. With k=3, KNN accuracy reached 86.61% at the highest and Naïve Bayes reached 87.40%. Meanwhile, based on percentage split test, the average of Naïve Bayes accuracy was higher than KNN with percentage of 88.31%. Moreover, for the precision and recall, Naïve Bayes was higher than KNN with 88.30% and 87.40% seen from the k-cross validation.

Keywords: stress; Naïve Bayes; KNN; accuracy; comparison

1. Introduction

Stress is thought or feeling that occurs as the response to adversity or threats, which are called as stressor [1]. [1] conveys that stress can positively motivate or trigger someone to reach some points, but on the other hand it can negatively cause health problems like indigestion and insomnia, especially for junior high school (JHS) students. Hence, determining students' stress level should be a concern so it can be directly handled. Unfortunately, the stress level that the students face is detected too late. Besides the lack of the personnel and consultation time at school, the other obstacle to introduce stress level is the students' reluctance to consult their problems.

Stress level can be determined based on classification method from data mining, which consists of several classification methods. The study conducted by [2] results that data mining classification method shows better performance compared to the tested system. There are several methods but K-Nearest Neighbors (KNN) and Naïve Bayes are the most frequently used methods [3]. Based on the procedure, these two methods are chosen because KNN can handle a lot of training data with abundant noise [4]. This notion is supported by [5] who claims that a lot of training data can be handled with KNN. Besides that, Naïve Bayes process is faster when applied to a lot of data set and it is easier to understand [3]. [6] adds that increasing number of Naïve Bayes data can increase the accuracy of method. These two methods return sufficient accuracy score [7].

This study aims to compare the accuracy of KNN method and Naïve Bayes method using WEKA to determine the stress level of JHS students. The data of this study were obtained using a questionnaire given to 254 respondents, which consists of respondents' identity and questions.

2. Theoretical Framework and Methodology

2.1. Definition of Stress

[8] explains that stress is a term that is commonly used to explain an unstable feeling condition, caused by anger, frustration, fatigue, or pressure. Furthermore, [8] notes that stress theoretically can be viewed as an effort to withstand physiological reaction when faced with suppressing condition or danger, which is called as *stressor*. According to the study of [9], stress can be classified into three categories, low stress, medium stress, and acute stress. Moreover, [10] explains that the stress level depends on how someone is exposed to the stressor, as the following:

a. Low stress. It is the early phase for someone to respond stressor which indicates a warning to make a resistance. This phase is followed with strong stimulation to physical symptoms, where someone shows strong feeling of anxiety and anger, fear, increase of heart rate and breathing rhythm, and sweat.
b. Medium stress. In this phase, the body slowly returns to its normal state characterized by reduced intensity and the recovery of the energy spent. Stimuli that arise are still high, but different from the previous level. At this stage, the visible stimuli are fatigue, getting offended easily, and anger.
c. Acute stress. This stage occurs when stressor exposes someone continuously. The intensity of heart rate and breathing decreases, but with the ongoing stress the energy will be drained. It is also possible to be followed with impaired heart and kidney function, allergies, and depression.

2.2. K-Nearest Neighbors

K-Nearest Neighbors (KNN) method is an algorithm used to estimate and predict, which is frequently used in classification process [11]. In their study, [11] explain that classification has similarity with estimation, but the target variable is in the form of categorical not numeric. In the description the classification works by:

a. Examining the data set that contains the target variables and predictions that are used as training data.
b. Sorting new data that are stored without including information to form the basis of training data, then a new classification for the data is determined.

KNN is categorized in instance-based learning where the testing to the new data to the training data already exists so the classification process of the new data is done by pairing the majority of similar training data during the testing [11]. This majority is drawn from the number of the nearest neighbors [12]. The distance function of KNN that is generally used is *Euclidean* distance with the formulation as the following:

$$d_{Euclidean} = \sqrt{\sum_i (x_i - y_i)^2} \tag{1}$$

Where:

d = *Euclidean distance*
x_i= test data
y_i= training data

2.3 Naïve Bayes

Naïve Bayes is derived from the Bayes's theorem assuming that all features are conditioned independently of each other against the target variable [13]. Bayes's theorem is formulated for the probability of an event using existing knowledge of the related conditions. The Bayes theorem is calculated from the following equation:

$$P(A \mid B) = \frac{P\ (B|A)\ P\ (A)}{P(B)} \tag{2}$$

Where A and B are events, P(A) is the probability of event A, and P(B) is the probability of event B. P(A | B) is the probability condition of event A for event B [13].

2.4 Methodology

This study is conducted in the following stages:

a. Literature study. It is done by analyzing the relevant theories and studies related to the topic. The literature includes books, electronic journals, and other reliable sources as the theoretical framework.
b. Data collecting. The data were gathered from students of Don Bosco Bitung Catholic JHS that consisted of 699 students based on School Monthly Report in October 2017. The data used for this study were obtained from 254 students. The source of the data was taken from the previous study conducted by [14]. The data were in the form of age, class, gender, number of children in the family, what number is the student in the family, and 20-question questionnaire,
c. Data Analysis. In this stage, the data were analyzed using Weka to determine the stress level using the two methods KNN and *Naïve Bayes*. The results of the two classification methods then were compared to calculate the accuracy of the two methods.

3. Results

This study used 254 data taken from the study of [14] that gathered the data from students of Don Bosco Bitung Catholic JHS. The result of the study explained that KNN could be applied to classify the stress level of JHS students, but it did not point out the accuracy obtained in the study [14]. The value of *k*=5 was used as parameter of neighborhood in this study [14]. Then, the KNN variable was numeric [11], so in the study [15] changed gender variable F and M into decimal to enable the data to be calculated using KNN, with F=80 and M=76. Then stress was classified into three, low stress, medium stress, and acute stress [9] where the distribution of JHS students stress can be seen in Table 1.

Table 1. Data Set

No	Gender	Age	Class	Children	Child number	Questions							Stress Classification
						1	2	3	4	5	...	20	
1	76	13	2	2	1	2	2	2	4	2	...	4	Low stress
2	80	12	2	2	1	4	4	3	3	3	...	5	Medium stress
3	80	14	2	2	3	3	2	5	5	2	...	4	Medium stress
4	80	13	3	3	3	3	1	1	2	1	...	4	Low stress
5	76	14	3	2	1	3	3	5	5	3	...	5	Acute Stress
...	...	...	...	...	...	...	...	...	...	...	...	...	...
254	80	14	2	5	2	5	2	3	4	3	...	4	Acute Stress

Table 2. Distribution of Stress Level

Low stress	Medium Stress	Acute Stress	Total
36	191	27	254

3.1 Testing on KNN

Out of 254 respondents, there were 36 students with low stress, 191 students with medium stress, and students with acute stress (Table 1). Then, testing on KNN with a value of k=1 to k=40 using Weka 3.9 was done, which is a k-cross validation test model with a value of folds=10 and percentage split. In the percentage split, the data was divided into 90, 80, 70 and 60 [16].

Table 3. Accuracy to Change on K-cross Validation and Percentage Split

k value	Accuracy Test				
	k-cross validation	Percentage split			
		90	80	70	60
1	84.64%	72.00%	78.43%	78.94%	81.37%
2	85.03%	84.00%	74.50%	73.68%	80.39%
3	86.61%	88.00%	82.35%	80.26%	87.35%
4	83.85%	80.00%	75.47%	76.31%	82.35%
5	85.43%	84.00%	82.35%	82.89%	88.23%
…	…	…	…	…	…
40	76.37%	80.00%	68.62%	71.05%	74.50%

Based on the data presented in Table 1, the highest accuracy value is taken to form the confusion matrix presented in Table 4 and Table 5. Then, Table 6 shows the precision and recall of the highest accuracy value.

Table 4. Confusion Matrix of K-cross Validation with Accuracy of 86.61%

a	b	c	Classified as
16	20	0	a = low stress
2	188	1	b = Medium stress

0	11	16	c = acute stress

Table 5. Confusion Matrix of Percentage Split of the Highest Accuracy

90			80			70			60			Classified as
a	**b**	**c**	**a**	**b**	**a**	**a**	**b**	**c**	**a**	**b**	**c**	
1	2	0	6	4	0	7	6	0	9	6	0	a = low stress
0	20	0	0	34	0	0	53	0	1	75	0	b = Medium stress
0	1	1	0	5	2	0	7	3	0	5	6	c = acute stress

Table 6. Precision and Recall KNN Based on the Highest Accuracy Values

Accuracy Test									
k-cross validation		**Percentage split**							
		90		**80**		**70**		**60**	
Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
87.20%	86.60%	89.60%	88.00%	86.00%	82.40%	86.30%	82.90%	89.00%	88.20%

3.2 Test on Naïve Bayes

Unlike the KNN, the test for Naïve Bayes with the k-cross validation requires only one test, as the KNN needs to determine the k value. The uncertainty of the k values used as classification reference makes KNN accuracy always change, but not for Naïve Bayes that only requires a one-time test with a k-cross validation accuracy obtained 87.40%, with confusion matrix that can be seen in Table 7 obtained from testing using WEKA.

Table 7. Confusion Matrix of K-cross Validation Test

a	b	c	Classified as
31	5	0	a = low stress
12	170	9	b = Medium stress
0	6	21	c = acute stress

The result of accuracy test with percentage split can be seen on Table 8 and confusion matrix of each test on Table 9H. Then, the precision and recall of the test can be seen in Table 9.

Table 8. Accuracy of Percentage Split of Naïve Bayes Test

Percentage Split Value	Accuracy
90	88.00%
80	90.19%
70	86.84%
60	88.23%

Table 9. Confusion Matrix of Percentage Split Test

90			80			70			60			Classified as
a	**b**	**c**	**a**	**b**	**c**	**a**	**b**	**c**	**a**	**b**	**c**	
2	1	0	9	1	0	10	3	0	13	2	0	a = low stress
2	18	0	2	31	1	3	48	2	5	68	3	b = Medium stress
0	0	2	0	1	6	0	2	8	0	2	9	c = acute stress

Table 10. Precision and Recall Naïve Bayes

Accuracy Test									
k-cross validation		**Percentage split**							
		90		**80**		**70**		**60**	
Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
88.30%	87.40%	89.80%	88.00%	90.40%	90.20%	86.80%	86.80%	89.10%	88.20%

3.3 Comparison of KNN and Naïve Bayes

In comparing the values of KNN and Naïve Bayes accuracy, the KNN accuracy is influenced with the number of the set nearest neighbors. Therefore, the values of KNN is taken from the highest accuracy values regardless the number of tested neighbors.

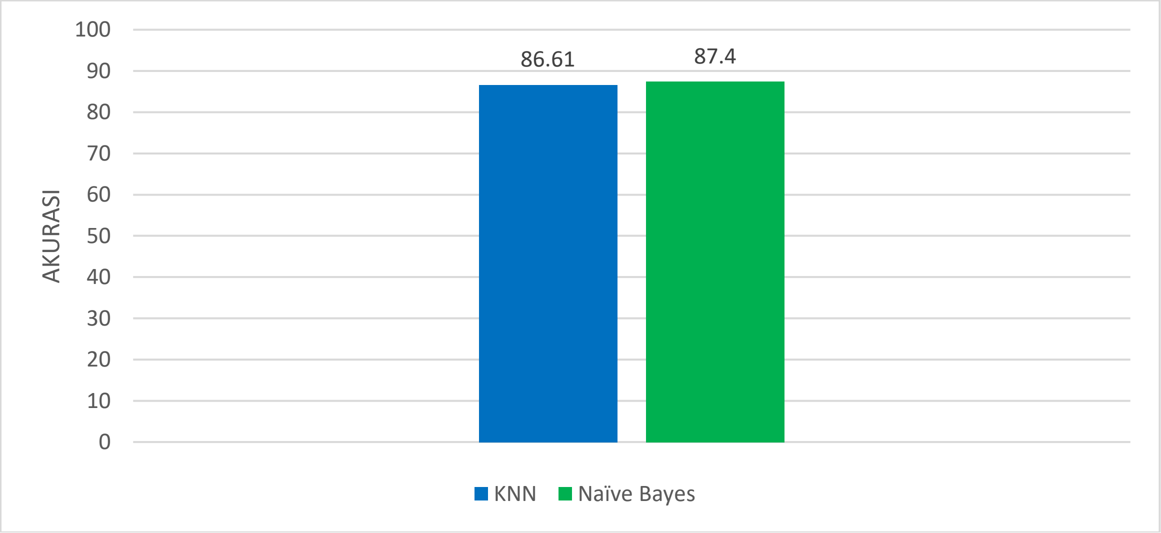

Figure 1. Comparison of accuracy of KNN and Naïve Bayes k-cross validation test

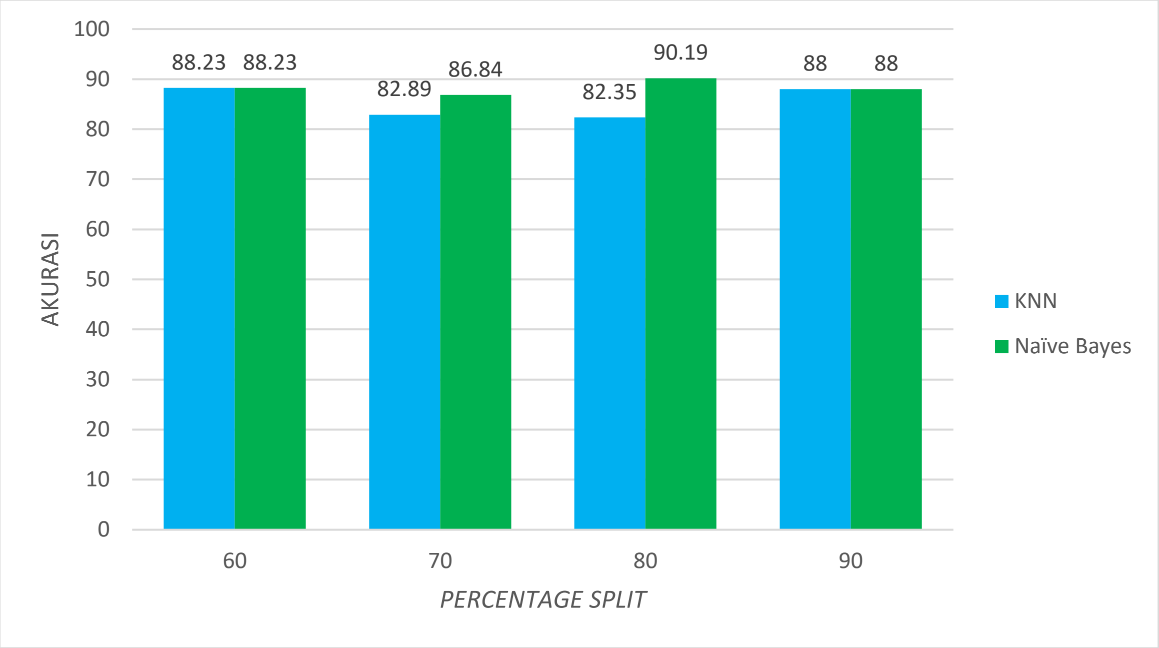

Figure 2. Comparison of accuracy from KNN and Naïve Bayes percentage split test

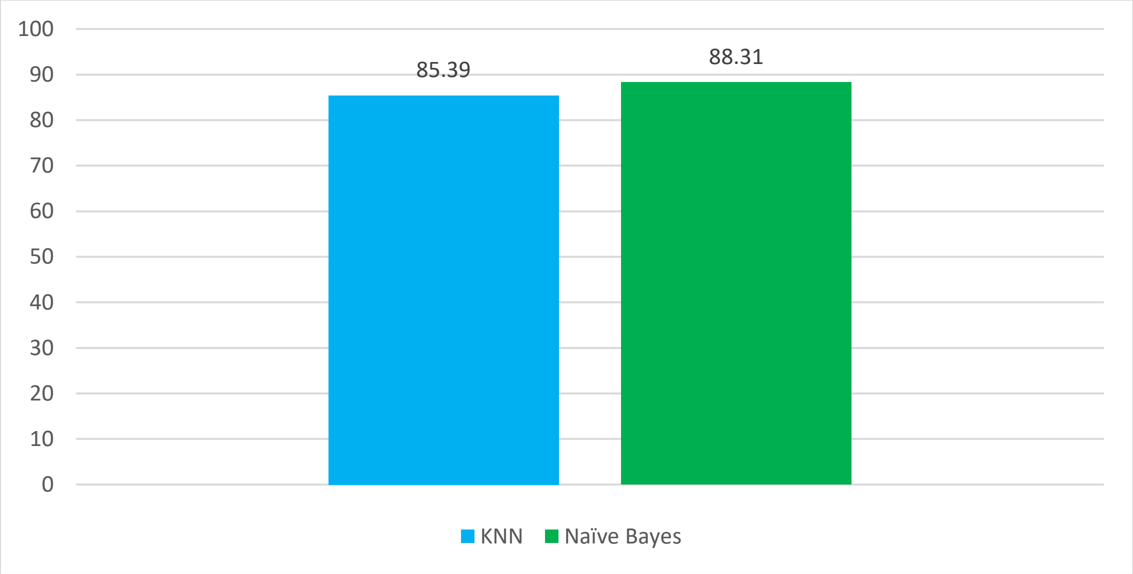

Figure 3. Comparison of average accuracy from KNN and Naïve Bayes percentage split Test

4. Discussion

The study of [17] explains that k=1 on KNN shows inflexible result because it only uses one nearest neighbor on the stored record. However, the use of big number of neighbors will blur the result as well, so the value of k=13 is the most optimum result since the accuracy reaches 75.14% from value of k=1 to *k*=49 [17]. This result is supported by [18] who uses k=13 and obtains accuracy of 97.28%, in other hand k=7 only obtain 54% for it accuracy [15]. On the other tests from 1 to 40, it is obtained k=3 with accuracy of 93% [19].

For the value of Naïve Bayes accuracy of 78.69%, it depends on the number of training data [6]. The similar thing is also concluded by [20] that the number of test data and training data can affect the accuracy of Naïve Bayes, in this case the accuracy is 80%. The value of 90.57% is obtained from the study conducted by [21] and this value is still higher compared to the implementation on heart disease risk prediction [22] where the accuracy is 78%.

The accuracy comparison of KNN and Naïve Bayes done by [7] shows the superiority of Naïve Bayes with the accuracy of 98.1% compared to KNN (the accuracy level is 95.3%). This is also supported with the study of [3] that the accuracy of Naïve Bayes is higher than KNN that is 72.5% compared to 57.5% in predicting the divorce case in Cimahi and the study of [23] on the classification of Indonesian articles with the Naïve Bayes accuracy of 70% compared to 40% of KNN accuracy. Not only compared to KNN, Naïve Bayes also seems to be superior to Support Vector Machine [24] and Neural Network [2], but [5] shows that KNN and Naïve Bayes give balance result. However, different opinion from [25] in determining the feasibility of planting teak tree says that KNN is superior compared to Naïve Bayes with accuracy of 96.66% compared to 82.63%. This opinion is also supported by [26] concerning the document text classification, where the KNN accuracy reaches 55.17%, surpassing Naïve Bayes with 39.01% accuracy.

It can be seen in Table 4 of the k-cross validation tests for KNN, the number of data that are successfully reclassified are 220 data and the false data are 34 data. In testing with percentage split as shown in Table 5, the number of data tested changed from 254 data to depending on the percentage split that is for the test data of 90% as many as 25 data, 80% test data as many as 51 data, 70% test data as many as 76, and 60% test data as many as 102. From those results, the data that were successfully reclassified correctly for 25 test data were 20 data and 5 incorrect data. Furthermore, out of 51 data there were 42 correct test data and 9 false data, out of 76 test data there were 63 correct data and 13 false data, and out of 102 test data, there were 90 correct data and 12 false data.

Table 7 shows a total of 222 correctly classified data and 32 false data from the k-cross validation test for Naïve Bayes method. In the results of a percentage split test in Table 9, the number of data tested from the 254 changed based on the value of percentage split of the test data by 90% as many as 25, 80% test data as many as 51, 70% test data as many as 76, and 60% test data as many as 102. From those results, the data that were successfully reclassified correctly for 25 test data were 22 data and the incorrect ones were 3, for 51 test data, 46 were correct and 5 were incorrect. Furthermore, for 76 test data, 66 were correct and 10 were incorrect, and for 102 test data the correct data were 90 and the incorrect ones were 12.

Based on the result obtained above, the comparison of KNN and Naïve Bayes in determining the stress level of 254 data shows that:

a. KNN and Naïve Bayes methods can be used to determine the stress level since they have accuracy values above 70%
b. Naïve Bayes method excels KNN in k-cross validation and percentage validation test, with the accuracy of Naïve Bayes as 87.40% and for the percentage split average as 88.31%.

Based on the k-cross validation test, the accuracy of Naïve Bayes is higher than KNN. However, for percentage split test, for 60% training data of KNN and Naïve Bayes has the same accuracy that is 88.23%, but for 70% and 80% percentage split Naïve Bayes excels KNN, and for 90% KNN and Naïve Bayes value is same.

5. Conclusion

Based on the discussion, it can be concluded if the change in the amount of data made affects accuracy, precision, and recall both through the k-cross validation and percentage split tests. The highest accuracy value of KNN of the k-cross validation test is at a k=3 value of 86.61%, a precision of 86.60% and a recall of 87.40%, but with the same value it produces different results for the percentage split test where the accuracy obtained reaches 88.00%, precision of 89.60% and recall of 88.00%. In the percentage split test of 80%, 70%, and 60%, the value of k=5 is obtained as the optimal.

However, in general accuracy, precision, and recall of Naïve Bayes are still higher than KNN. This can be seen from the accuracy of k-cross validation of Naïve Bayes and the accuracy average of percentage split test, with the highest *Naïve Bayes* of *percentage split* 80% for the accuracy of 90.19% with *precision* 90.40%, and *recall* 90.20%. This is also influenced by the number of data used in the test, so it is suggested that in the future the number of data is increased.

6. References

[1] P. Parks, *Teen And Stress*. San Diego: Reference Point Press, 2015.

[2] J. S. Kanchana, H. T. Fathima, R. Surya, and R. Sandhiya, “Stress Detection Using Classification Algorithm,” *Int. J. Eng. Res. Technol.*, vol. 7, no. 04, 2018.

[3] I. A. Dahlia, M. Irfan, and W. Uriawan, “Perbandingan Metode Naive Bayes dan K-Nearest Neighbor untuk Prediksi Perceraian (Studi Kasus : Pengadilan Agama Cimahi),” *Insight*, vol. 1, 2018.

[4] N. M. S. Iswari, W. Wella, and R. Ranny, “Perbandingan Algoritma kNN, C4.5, dan Naive Bayes dalam Pengklasifikasian Kesegaran Ikan Menggunakan Media Foto,” *J. Ultim.*, 2017.

[5] N. T. Romadloni, I. Santoso, and S. Budilaksono, “Perbandingan Metode Naïve Bayes, KNN dan Decision Tree Terhadap Analisis Sentimen Transportasi KRL Commuter Line,” *IKRA-ITH Inform.*, vol. 3, 2019.

[6] N. R. Indraswari and Y. I. Kurniawan, “APLIKASI PREDIKSI USIA KELAHIRAN DENGAN METODE NAIVE BAYES,” *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, 2018.

[7] D. Prajarini, “Perbandingan Alogoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Kulit,” *Informatics J.*, vol. 1, 2016.

[8] R. M. Spielman, *Psychology*. Houston: Openstax, 2017.

[9] Warno, T. Umar, and R. Arlizon, “Pengaruh Layanan Bimbingan Kelompok Terhadap Penurunan Tingkat Stres Siswa Kelas VIII SMP IT Al-Ikhsan Boarding School Riau,” *J. Online Mhs. Bid. Kegur. dan Ilmu Pendidik.*, vol. 2, 2015.

[10] J. S. Nevid, *Essential of Psychology : Concepts and Applications*, 5th ed. Boston: Cangage Learning, 2015.

[11] D. T. Larose and C. D. Larose, “Data Mining and Predictive Analytics,” *Wiley Ser. Methods Appl. Data Min.*, 2015.

[12] Y. L. Prasadad, *Big Data Analytics Made Easy*. Chennai: Notion Press, 2016.

[13] G. Hackeling, *Mastering Machine Learning With Scikit-learn*, 2nd ed. Birmingham: Packt Publishing, 2017.

[14] Y. C. Tapidingan, D. Paseru, and R. Turang, “Sistem Klasifikasi Penentuan Tingkat Stres Siswa Sekolah Menengah Pertama Menggunakan Metode K-Nearest Neighbors,” *J. disajikan dalam 3rd Int. Conf. Oper. Res. Univ. Sam Ratulangi, Manad. 20-21 Sept.*, 2018.

[15] T. D. Rahmawati and F. N. Adnan, “Penentuan Produk Asuransi BPJS Berdasarkan Profil Pelanggan Dengan Pendekatan K-Nearest Neighbor Manhattan Distance,” *J. Inf. Syst.*, vol. 2, 2016.

[16] D. Sartika and D. Indra, “Perbandingan Algoritma Klasifikasi Naive Bayes, Nearest Neighbour, dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian,” *J. Tek. Inform. Dan Sist. Inf.*, 2017.

[17] Indrayanti, D. Sugianti, and M. A. Al Karomi, “Optimasi Parameter K Pada Algoritma K-Nearest Neighbour Untuk Klasifikasi Penyakit Diabetes Mellitus,” *Pros. SNATIF Ke-4 2017*, 2017.

[18] S. W. Binabar and Ivandari, “Optimasi Parameter K pada Algoritma KNN untuk Deteksi Penyakit Kanker Payudara,” *IC-Tech*, vol. 13, 2018.

[19] I. N. Atthalla, A. Jovandy, and H. Habibie, “Klasifikasi Penyakit Kanker Payudara Menggunakan Metode K-Nearest Neighbor,” *Pros. Annu. Res. Semin.*, vol. 4, 2018.

[20] I. L. Qurnia, E. Prasetyo, and R. F. Zainal, “CLASSIFICATION OF DIABETES DISEASE USING NAIVE BAYES Case Study : SITI KHADIJAH HOSPITAL,” 2016.

[21] M. S. Islam, M. I. Fauzan, and M. T. Pratama, "Penggunaan Naïve Bayes Classifier Untuk Pengelompokan Pesan Pada Ruang Percakapan Maya Dalam Lingkungan Kemahasiswaan", *J. Computech & Bisnis*, vol. 4, no. 2017.

[22] M. Sabransyah, Y. N. Nasution, and D. Tisna, “Aplikasi Metode Naive Bayes dalam Prediksi Risiko Penyakit Jantung Naive Bayes Method for a Heart Risk Disease Prediction Application,” *J. EKSPONENSIAL*, 2017.

[23] R. N. Devita, H. W. Herwanto, and A. P. Wibawa, “Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa indonesia,” *J. Teknol. Inf. dan Ilmu Komput.*, 2018.

[24] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, O. Rusli, and Rudy, “News Article Text Classification in Indonesian Language,” in *Procedia Computer Science*, 2017.

[25] D. Srianto and E. Mulyanto, “Perbandingan K-Nearest Neighbor dan Naïve Bayes Untuk Klasifikasi Layak Tanam Pohon Jati,” *J. Teknol. Inf.*, vol. 15, 2016.

[26] Z. E. Rasjid and R. Setiawan, “Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques,” in *Procedia Computer Science*, 2017.