Comparison of k-Nearest Neighbor and Naive Bayes Methods for SNP Data Classification

Denny Indrajaya[1], **Adi Setiawan**[2], **Bambang Susanto**[3]
Universitas Kristen Satya Wacana, Salatiga, Indonesia

Article Info

Article history:

Received March 15, 2022
Revised June 10, 2022
Accepted October 04, 2022

Keywords:

Classification
k-Nearest Neighbor
Naive Bayes
Single Nucleotide Polymorphism

ABSTRACT

In an accident, sometimes the identity of a person who has an accident is hard to know, so it is necessary to use biological data such as Single Nucleotide Polymorphism (SNP) data to identify the person's origin. This research aims to compare the accuracy and the F1 score of the k-Nearest Neighbor method and the Naive Bayes method in classifying SNP data from 120 people who divide into groups, namely European (CEU) and Yoruba (YRI). Determination of the best method based on the average value of accuracy and the average value of F1 score from 1000 iterations with various percentage distributions of training datasets and testing datasets. In this research, the selection of SNP locations for the classification process was carried out by correlation analysis. The average accuracy obtained for the k-Nearest Neighbor method with the value of k=31 is 98.38% where the average F1 score is 98.39% while the Naive Bayes method obtained the average accuracy of 96.74% and the average F1 score of 96.63%. In this case, the k-Nearest Neighbor method is better than the Naive Bayes method in classifying SNP data to determine the origin of a person's ancestor tends to be from CEU or YRI.

Corresponding Author:

Denny Indrajaya,
Department of Mathematics, Faculty of Science and Mathematics,
Universitas Kristen Satya Wacana, Salatiga, Indonesia,
Email: 662018003@student.uksw.edu

How to Cite:

1. INTRODUCTION

DNA is a constituent of genes that play a role in determining hereditary traits and passing on biological information from one individual to their offspring [1]. In people's lives, it is not uncommon to find DNA tests to find out the similarity between 2 persons such as parents and children, which only require DNA data from 2 persons. In addition, if one day someone whose identity is unknown has an accident, which biological data could be used to find out the origin of the person, in the identification process is not enough if we only use biological data from 2 persons as the example mentioned earlier, biological data from several people will need because each person can come from several different breeds. Single Nucleotide Polymorphism (SNP) data can use to match a person's ancestral origin. SNP is a variation of genetic material indicated by the presence of single nucleotides (adenine, thymine, guanine, cytosine) in the DNA genome, where different nucleotides occur in various individuals of a population [2]. DNA sequence variation in living things occurs when there are particular nucleotides in a pair of homologous chromosomes, there are 99.9% DNA similarities then the rest are differences called SNP [3]. Therefore, it can be said that although the genetic differences which exist between 2 persons are only a small part, they can have various impacts on the differences that exist in each individual, one of which is physical differences. According to research [4], changes and differences in the structure of SNPs can have a role in the existence of diversity among individuals. The study [5] also states that SNP is effective in detecting genetic differences.

The use of SNP data in research has been carried out widely, some of which are the identification of mutations in genes [6], drug development and knowing a person's response to a drug [7], and the susceptibility of a person to certain diseases [8]. Research related to the analysis of SNP data using the Chi-Square method has also been carried out previously in research [9]. Based on those studies, which is known that genotype markers affect their phenotype.

The application of machine learning to SNP data has varied, as has been done in the study [10] about knowing a person is at risk of asthma with a machine learning approach based on SNP. That study used Random Forest and Recursive Feature Elimination methods to identify SNPs that influence asthma and used k-Nearest Neighbor (k-NN) and Support Vector Machine methods for classification. Research related to the categorization of SNPs using the Radom Forest method has also been used in other studies, that is the classification of farm animals (cattle) divided into six breeds [11]. However, in that study, attribute reduction was carried out using statistical methods (Delta, Fst, and PCA). The use of machine learning in SNP data related to mental disorders and cancer has also been carried out in research [12] using various classification methods, such as k-NN, Support Vector Machine, and Artificial Gene Making (AGM/Alpha). Attribute selection was also carried out, which namely ReliefF, Feature Selection Based on Distance Discriminant (FSDD), Feature Selection Based on R-value (RFS), and Algorithm Based on Feature Clearness (CBFS). Identification of positive SNPs and negative SNPs in the soybean genome has also been carried out in research [13] using the C5.0 algorithm. In this study, attributes that have a good influence on the results of classification which also carried out. As in previous studies, this study was carried out using SNP data. All attributes will not be used because the SNP data reached thousands. The number of those attributes will affect the time to computation and classification accuracy, so the number of attributes needs to be reduced while still paying attention to the accuracy obtained. The reduction of those attributes used for classification in this study was carried out based on correlation analysis, where attribute reduction technique with correlation analysis has not been carried out in the previously mentioned studies.

Based on the description presented, this study will show the use of correlation analysis in reducing the number of attributes used in the application of k-NN and Naive Bayes methods for the classification of SNP data. The purpose of applying these two methods to the SNP data is to find the tendency of a person's ancestor origin, especially in identifying the identity of an unknown person. The data used in this study is SNP data which divide into two groups, namely people from Europe (CEU) and people from Yoruba (YRI). In addition, an analysis will also be carried out regarding better methods between k-NN and Naive Bayes methods in the classification of SNP data.

The preparation of this paper will divide into four main parts. The first part is an introduction that explains the background of the research with the state of the art research. The second part is the research method, which explains about methods used to obtain the research results shown in the third part. In the third part, in addition to explaining the results of the study, the analysis carried out related to the application of k-NN and Naive Bayes methods to the SNP data was also presented, while the fourth part contained conclusions and suggestions for further research.

2. RESEARCH METHOD

This study focuses on comparing the methods of k-NN and Naive Bayes in classifying SNP data to determine the tendency of a person's ancestor origin from CEU or YRI. The method used for conducting the analysis is quantitative (sorting data process by correlation analysis and comparing methods through the help of a confusion matrix). The analysis process in this study is also directly carried out every time data processing because the result of the analysis will be needed for further data processing. So the adjustment of the data used and the classification process are carried out several times based on the analysis that will be done. The stages of the research carried out are depicted in the flow chart shown in Figure 1.

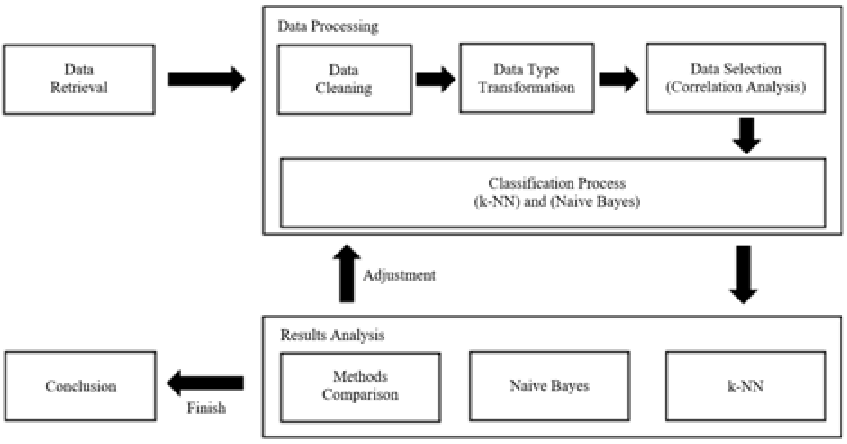


Figure 1. Research flow chart

2.1. Data Retrieval

The data used in this study were obtained from the HapMap project, which is 9305 SNP locations from 120 people divided into two different populations, namely 60 people from the European population (CEU) and 60 people from the Yoruba population (YRI). The data file name is SNPassoc_2.0-2.tar.gz, the file downloaded through the //cran.r-project.org/src/contrib/Archive/SNPassoc/ page, which is a place to download packages and data for R software. To open the successfully downloaded SNP data file, we can extract the file first, then a file with the name HapMap.rda will be obtained that contains the SNP data from the HapMap project. The file can open using R software. Table 1 shows a sample of data used in this study.

Table 1. Sample data from a HapMap project

id	Group	rs10399749	rs11260616	rs4648633	rs6659552	rs7550396	rs6688969	rs10753357	rs1495243
NA06985	CEU	CC	AA	TT	GG	GG	CC	AC	GG
NA06993	CEU	CC	AT	CT	CG	GG	CT	AA	AG
NA06994	CEU	CC	AA	TT	CG	GG	CT	AA	AA
NA07000	CEU	CC	AT	TT	GG	GG	CC	AC	GG
NA19222	YRI	CC	AT	TT	GG	GG	CC	CC	AG
NA19223	YRI	CC	AA	TT	GG	GG	CT	CC	AG
NA19238	YRI	CC	AA	TT	GG	GG	CC	CC	AG
NA19239	YRI	CC	AA	TT	GG	GG	CC	CC	AG

2.2. Data Retrieval

In this study, a classification process was carried out using k-NN and Naive Bayes methods, so there were two types of data processing. However, before applying those methods to SNP data, the data should be cleaned first, and the value of the correlation coefficient (r) between the SNP location and the class used was calculated as well. The value of r is interpreted and used for determining the location of the SNP for classification. To find the value of r which represents the correlation between the location of the SNP and the class used, it is necessary to transform the SNP data type first, from a character type to a numeric type. The value of the correlation coefficient (r) expresses the relationship between two variables, which is Formula (3.3.) states the Pearson correlation coefficient formula [14].

$$r_{xy} = \frac{\Sigma x_i y_i}{\sqrt{\Sigma x_i^2 \Sigma y_i^2}} \tag{1}$$

In correlation theory, there are positive correlation and negative correlation, where the value of the correlation coefficient r=0 means no linear relationship and $r = \pm 1$ indicates a perfect linear relationship between 2 variables [15]. There is a degree of relationship that can be seen from the value of the correlation coefficient shown in Table 2 [16].

Table 2. Interpretation of the correlation coefficient r

Interval Correlation Coefficient	Relationship Level
(-0.2,0) ∪ (0,02)	Very weak
(-0.4, -0.2] ∪ [0.2, 0.4)	Weak
(-0.6, -0.4] ∪ [0.4, 0.6)	Moderate
(-0.8, -0.6] ∪ [0.6, 0.8)	Strong
(-1, -0.8] ∪ [0.8, 1)	Very strong

The k-NN classifier method is a method used to predict the class of an object based on information from the objects closest to that object [17]. This method can only be used if the data used is numeric because it is used to calculate the distance between objects in the testing dataset and objects in the training dataset [17]. The distance calculation used in this method is Euclidean distance indicated by Formula (2).

$$d(x,y) = \parallel x - y \parallel = \sqrt{\Sigma_{i=1}^{n}(x_i - y_i)^2} \tag{2}$$

The classification steps using the k-NN method on the SNP data are as follows.

1. Data division is carried out into 2 parts, for example, 80% training dataset and 20% testing dataset.
2. One of the objects from the testing dataset is taken, then the distance between the object and all objects in the training dataset is calculated.
3. Sort the distance between the object of the testing dataset and each object in the training dataset from the closest to the farthest distance.
4. Determine the value of k to determine the result of the classification.
5. Take k objects in the training dataset that has the closest distance to the object from the testing dataset.
6. Calculate the number of YRI and CEU classes of the objects that have been taken in step 5. If the number of YRI is more than CEU, the classification result is YRI, but if the number of CEU is more than YRI, the classification result is CEU.

Based on [18] and [19] the Naive Bayes classifier method uses probability to determine the classification of an object to a given class, which is it takes p attributes with $x_i = (x_i1, x_i2, \ldots, x_ip)$, where $i = 1, 2, , n$ which means there are n predictors, and each sample is assumed to have one class $y \epsilon \{y_1, y_2, \ldots, y_j\}$ with j classes. The Naive Bayes method is formulated by Formula (3).

$$P(y_j|x_i) = \frac{P(x_i|y_j)P(y_j)}{P(x_i)} \tag{3}$$

where $P(y_j|x_i)$= probability y_j based on x_i condition (posterior probability), $P(x_i|y_j)$= probability of x_i based on y_j condition, $P(y_j)$= probability of y_j based on data in the training dataset, and $P(x_i)$= probability of the object x_i from the data in the testing dataset based on the data in the training dataset.

Based on Formula (3), if p attributes are used, then the class resulting from classifying by the Naive Bayes method is formulated by Formula (4).

$$\hat{y} = argmax_{yj} P(y_j) \Pi_{k=1}^{p} P(x_k|y_j) \tag{4}$$

where $\hat{y}$ states the classification result and the argmax is an operation that gives a class result based on the value of maximum a posteriori (MAP) obtained [19]. The classification steps using the Naive Bayes method on the SNP data are as follows.

1. Data division is carried out into 2 parts, for example, 80% training dataset and 20% testing dataset.
2. Calculate the required probability values based on the training dataset.
3. One of the objects from the testing dataset is taken to find the result of the classification.
4. Calculate the posterior probability for each class based on the data of the object from the testing dataset and the probability values obtained from step 2.
5. Based on step 4, the largest posterior probability is taken as a determinant of the classification result. If the largest posterior probability is YRI then the classification result is YRI, but if the largest posterior probability is CEU then the classification result is CEU.

The evaluation of k-NN and Naive Bayes methods is carried out based on the number of objects in the test dataset classified correctly and incorrectly, which are tabulated in a matrix called the confusion matrix [20]. Table 3 is an example of a confusion matrix, the values that can be obtained from the confusion matrix are True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN). Such values are used to obtain values that indicate the performance of the classifier method indicating in Table 4. In Table 4, there is a formula F1 score which is an alternative to measuring the performance (accuracy) of the classification method in addition to using the accuracy formula shown in Table 4 [17].

Table 3. Confusion matrix

		Prediction	
		Class 1	Class 2
Actual	Class 1	*TP*	*FN*
	Class 2	*FP*	*TN*

Table 4. The values that can be obtained from the confusion matrix

No	Name	Formula	Information
1	Accuracy	$\frac{TP+TN}{TP+FN+FP+TN}$	The percentage of models predicting correctly.
2	Recall	$\frac{TP}{TP+FN}$	The percentage of positive data is predicted to be positive.
3	Precision	$\frac{TP}{TP+FP}$	The percentage of data predictions as positives is correct.
4	F1	$2 \times \frac{Precision \times Recall}{Precision + Recall}$	The harmonic mean of precision and recall.

2.3. Analysis of Results

In obtaining the results that are the purpose of this study, it is necessary to carry out several analyzes as follows:

1. SNP Location
   SNP location analysis needs to be carried out because the classification process in this research does not use all SNP locations, but only a few. Location determination is implemented based on the correlation between the location of the SNP and the class used. An interpretation of the obtained r value will be carried out in conducting the correlation analysis. In addition, a correlation analysis was performed to see the influence of the location of the SNP on the classification results with k-NN and Naive Bayes methods shown by the graph.
2. Classification results using the k-Nearest Neighbor and Naive Bayes methods
   In conducting a classification analysis with the k-NN method, the value of k that can produce the best accuracy will be determined. An accuracy calculation trial will implement with several values of k whose results are presented in a graph. In conducting a classification analysis with the Naive Bayes method, a classification process will be shown to determine the influence of data on the probabilities obtained. Next will show a graph presenting the average accuracy, average recall, average

precision, and average F1 score of the application of k-NN and Naive Bayes methods with a wide variety of SNP locations and the number of locations. Those four average values are obtained from the calculation of the values of accuracy, recall, precision, and F1 score several times. In each calculation, the distribution of the training dataset and the testing dataset was randomized before being calculated. In addition, the average accuracy and the average F1 score are used to look at the performance of k-NN and Naive Bayes methods in classifying SNP data to determine whether a person's ancestor origin tends to be from Europe (CEU) or Yoruba (YRI).

3. Comparison of k-Nearest Neighbor and Naive Bayes methods
In comparing the k-NN and Naive Bayes methods in classifying SNP data, the average accuracy and average F1 score will be used. In addition, the average value of the average accuracy and the average F1 score for each method is also calculated. The calculation results obtained are used to compare the k-NN and Naive Bayes methods if it is found the k-NN method is better than the Naive Bayes method when viewed from the average accuracy and the Naive Bayes method is better than the k-NN method when viewed from the average F1 score, or vice versa, the k-NN method is better than Naive Bayes when viewed from the average F1 score and the Naive Bayes method is better than the k-NN method when viewed from the average accuracy. Based on those values, a better method was obtained in classifying SNP data to find out the tendency of a person's ancestor's origin.

3. RESULT AND ANALYSIS

3.1. SNP Location Selection

In choosing a location on SNP data, a correlation analysis is carried out first. Numerical type data is needed to perform correlation analysis. Knowing SNP data is data with a character type, it needs to be changed to numeric first. In changing the data type from character to numeric in SNP data, it is necessary to pay attention to the number of genotype marker data that may appear at each SNP location. For each SNP location, only a maximum of 3 data differences will appear, for example, AA, AT, and TT, namely Adenine-Adenine, Adenine-Thymine, and Thymine-Thymine. If we look at the difference, it must be that AA and AT have the same level of difference as AT and TT, while AA and TT must have a greater degree of a difference than AA and AT. In addition, by applying dominant, codominant, and recessive traits such as those in genes can be assumed that AA is dominant and TT recessive then AT is codominant. However, if an SNP location is found with only 2 genotype markers, it is concluded that if at that location there is a codominant genotype marker, then the other genotype marker is dominant. For this reason, assigning values to SNP data for locations that have 3 kinds of genotype markers can be done as shown in Table 5(a) and SNP locations that only have 2 kinds of genotype markers can be done as shown in Table 5(b).

Table 5. (a) Data type transformation on SNP data for 3 kinds of genotype markers, (b) Data type transformation on SNP data for 2 kinds of genotype markers

(a)				(b)				
AA	AT	TT	AA	TT	AA	AT	AT	TT
0	0.5	1	0	1	1	0.5	0.5	1
AA	AC	CC	AA	CC	AA	AC	AC	CC
0	0.5	1	0	1	1	0.5	0.5	1
AA	AG	GG	AA	GG	AA	AG	AG	GG
0	0.5	1	0	1	1	0.5	0.5	1
CC	CT	TT	CC	TT	CC	CT	CT	TT
0	0.5	1	0	1	1	0.5	0.5	1
CC	CG	GG	CC	GG	CC	CG	CG	GG
0	0.5	1	0	1	1	0.5	0.5	1
GG	GT	TT	GG	TT	GG	GT	GT	TT
0	0.5	1	0	1	1	0.5	0.5	1

In this study, the use of SNP data in classification was carried out to find the tendency of a person's ancestor origin by using the k-NN and the Naive Bayes methods. Before the classification process, it is necessary to find the value of the correlation coefficient between the location of the SNP and the class used. Based on the results of the calculation of the value of the correlation coefficient in the SNP data that has been carried out, it is found the number of locations has a very weak to very strong correlation as shown in Table 6.

Table 6. The multiplicity of SNP locations for each correlation interpretation

No	Interpretation	Number of SNP location
1	Very weak correlation	1141
2	Weak correlation	1270
3	Moderate correlation	726
4	Strong correlation	232
5	Very strong correlation	23

3.2. Results of SNP data classification using the k-NN method

Table 7. SNP locations for determining the value of k in the k-NN method

Reference SNP ID number lokasi SNP	Correlation Interpretation
rs3887675	Very weak
rs1784176	Weak
rs4041435	Moderate
rs3136687	Strong
rs10868791	Very strong

In carrying out classification with the k-NN method, the values of accuracy and F1 score obtained are different for each k value used, so it is necessary to find the k value with the best accuracy when using the k-NN method. Using the 5 SNP locations indicated in Table 7, the value of k that can maximize accuracy when classifying SNP data using the k-NN method was calculated. In finding the k value based on average accuracy, iterations 1000 times for each k value with a data division of 80% training dataset and 20% testing dataset, and the Euclidean distance were used. The value of k used is an odd number from the interval 1 to 96, The value of 96 is the number of people whose SNP data is used in the training dataset. Odd numbers were chosen because the SNP data in this study only had 2 classes, namely CEU and YRI. By using the odd numbers at the value of k, the same number of voting results will not be obtained during the classification process using the k-NN method. The graph of the average accuracy with some values of k is shown in Figure 2. Based on Figure 2 the highest average accuracy is obtained when the value of k is equal to 31. For this reason, the value of k used for further analysis related to classification by the k-NN method in the SNP data is 31.

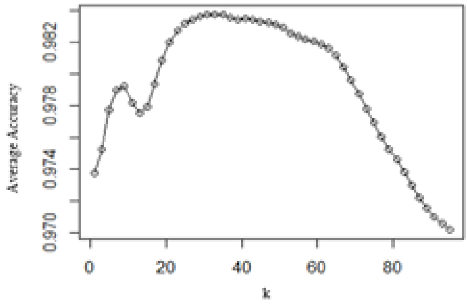
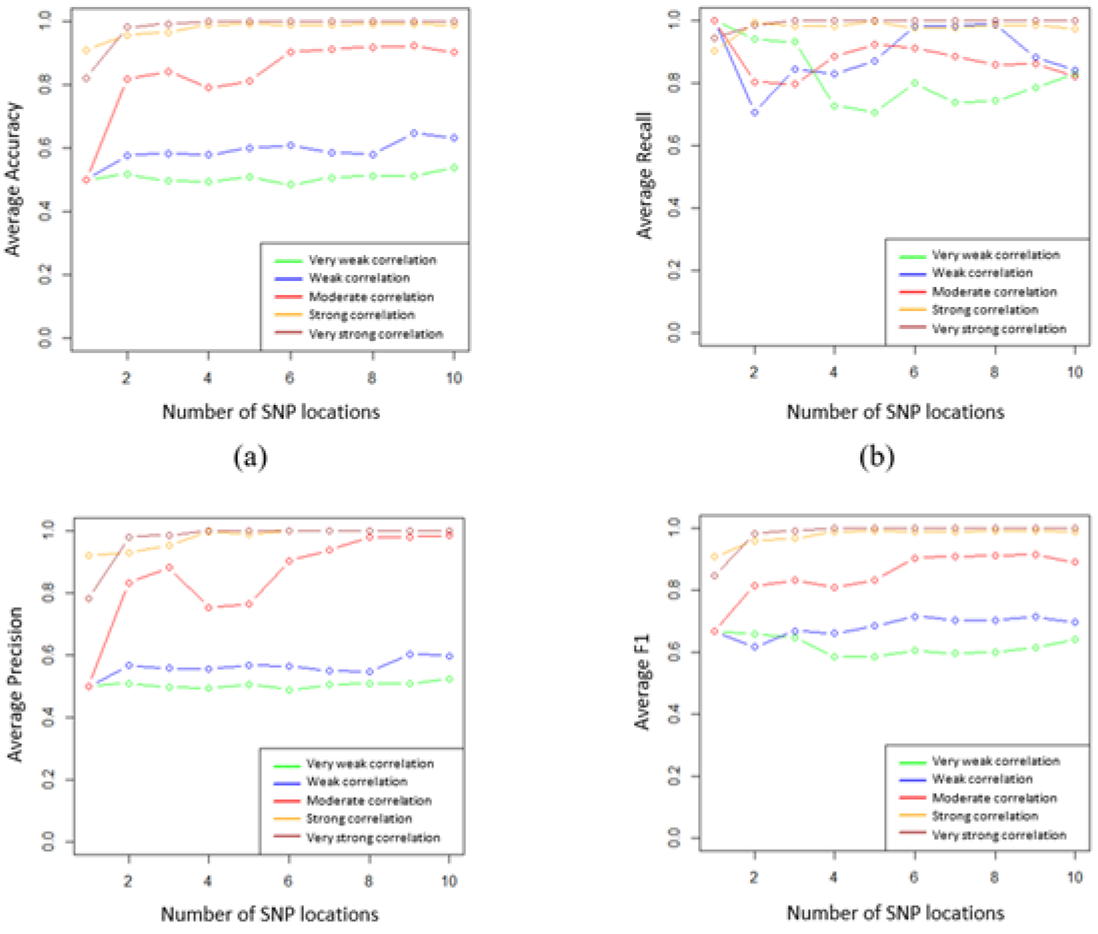


Figure 2. Graph of k values and average accuracy on the k-NN method

Based on experiments that have been carried out using the k-NN method with the number of various SNP locations, namely from 1 to 10, the results obtained in the four graph forms of average accuracy, average recall, average precision, and average F1 score as shown in Figure 3. The results in Figure 3 are obtained using the SNP locations shown in Table 8, the data division of 80% training dataset and 20% testing dataset, Euclidean distance, k value is equal to 31, and 1000 iterations. In Figure 3, the locations used for each classification process have the same interpretation of correlation, which is the interpretation on the graph indicated by the color of the lines. The classification is carried out using SNP locations that have correlations with very weak, weak, moderate, strong, and very strong interpretations to see the effect of using the SNP location based on the correlation coefficient value on the accuracy, recall, precision, and F1 scores obtained from the use of the k-NN method on SNP data. Figure 3 shows that the correlation between the location of the SNP and the class used influences the accuracy, recall, precision, and F1 score obtained using the k-NN classifier method.

Table 8. SNP locations used in Figure 3

Reference SNP ID number SNP location selected based on interpretation of the correlation between SNP location and the class used				
Very weak	Weak	Moderate	Strong	Very Strong
rs11260616	rs6688969	rs6659552	rs12745075	rs3767067
rs10753357	rs1495243	rs6577401	rs3124625	rs6670842
rs6681520	rs12093433	rs12119523	rs4649089	rs9436924
rs12136845	rs4846094	rs4661627	rs12727605	rs2003154
rs7419115	rs3795746	rs12076353	rs1342412	rs619228
rs4846056	rs7517964	rs4653095	rs10782748	rs6716734
rs6681992	rs16853611	rs1325241	rs912073	rs6814827
rs12405926	rs12093675	rs2359108	rs11165510	rs1485768
rs3026792	rs885795	rs12140004	rs3753841	rs809039
rs298459	rs16862254	rs784627	rs1778820	rs7752055



(a)

(b)

(c)

(d)

Figure 3. The results of the application of the k-NN method (a) Graph of average accuracy with the number of SNP locations, (b) Graph of average recall with the number of SNP locations, (c) Graph of average precision with the number of SNP locations, (d) Graph of average F1 score with the number of SNP locations

The application of the k-NN method using 5, 10, and 15 SNP locations with very weak, weak, moderate, strong, and very strong correlation was also carried out. The locations of the SNP used are indicated in Table 9. In the classification process, the data division used is 80% training dataset and 20% testing dataset, the distance used is Euclidean distance, and the k value used is 31. The accuracy value obtained from 100 iterations by randomizing the distribution of the training dataset and the testing dataset first in each iteration is shown in Figure 4.

Table 9. (a) The locations of the SNP used in Figure 4a, (b) the Locations of the SNP used in Figure 4b, (c) The locations of the SNP used in Figure 4c

(a)

Reference SNP ID number location SNP	**Correlation Interpretation**
rs3887675	Very weak
rs1784176	Weak
rs4041435	Moderate
rs3136687	Strong
rs10868791	Very strong

(b)

Reference SNP ID number location SNP	**Correlation Interpretation**
rs3887675	Very weak
rs2064437	Very weak
rs1784176	Weak
rs470039	Weak
rs4041435	Moderate
rs1979710	Moderate
rs3136687	Strong
rs1435085	Strong
rs10868791	Very strong
rs6814827	Very strong

(c)

Reference SNP ID number location SNP	**Correlation Interpretation**
rs11260616	Very weak
rs3887675	Very weak
rs2064437	Very weak
rs6688969	Weak
rs1784176	Weak
rs470039	Weak
rs6659552	Moderate
rs4041435	Moderate
rs1979710	Moderate
rs12745075	Strong
rs3136687	Strong
rs1435085	Strong
rs3767067	Very strong
rs10868791	Very strong
rs6814827	Very strong

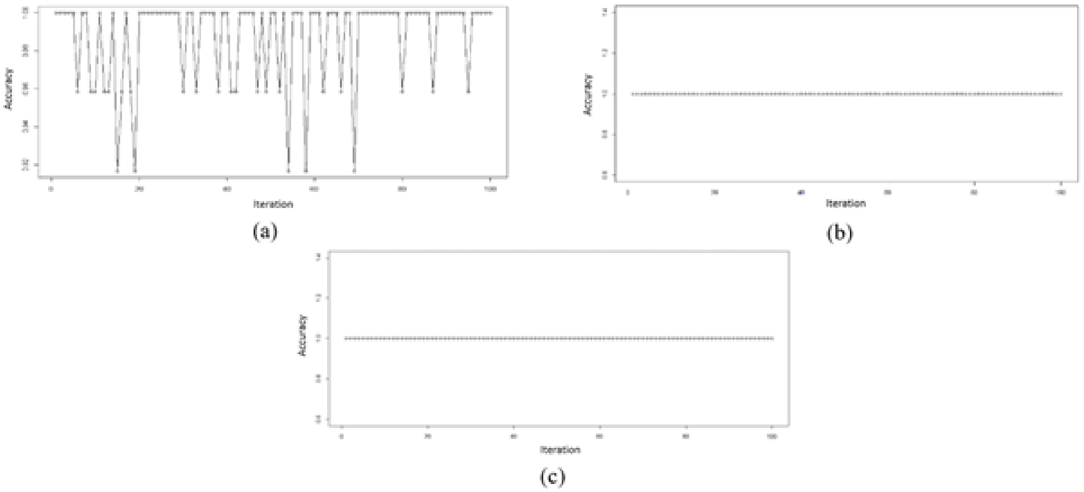


Figure 4. (a) Accuracy of the k-NN method using 5 SNP locations, (b) Accuracy of the k-NN method using 10 SNP locations, (c) Accuracy of the k-NN method using 10 SNP locations

3.3. Results of SNP data classification using the Naive Bayes method

Table 10. Example of fictitious data for the use of the Naive Bayes method with 1 SNP location

Respondent	Location of SNP 1	Class
1	AA	CEU
2	AA	CEU
3	AA	CEU
4	AG	YRI
5	GG	YRI
6	GG	YRI

In using the Naive Bayes method, it is necessary to pay attention to the location of the SNP used because in this method a conditional probability is used, which allows obtaining a probability value of 0 will make the result of the Naive Bayes classification undetermined. For more details, look at the example of fictitious data in Table 10. If data such as Table 10 is owned, several probability values will be obtained as follows.

$$P(CEU) = \frac{3}{6} = \frac{1}{2}, P(YRI) = \frac{3}{6} = \frac{1}{2}$$

In addition, probabilities were also obtained based on the data on the location of SNP 1 as follows.

$$P(AA \mid CEU) = \frac{3}{3} = 1, P(AG \mid CEU) = \frac{0}{3} = 0, P(GG \mid CEU) = \frac{0}{3} = 0, P$$

$$(AA \mid YRI) = \frac{0}{3} = 0, P(AG \mid YRI) = \frac{1}{3}, P(GG \mid YRI) = \frac{2}{3}$$

If the new object with AA data is located at the SNP 1 location, then to find the result of the classification of the object, the posterior probability values of CEU and YRI can be calculated first.

$$\text{Posterior probability of CEU} = P(CEU) \times P(AA \mid CEU) = \frac{1}{2} \times 1 = \frac{1}{2}$$

$$\text{Posterior probability of YRI} = P(YRI) \times P(AA \mid YRI) = \frac{1}{2} \times 0 = 0$$

Obtained posterior probability CEU is greater than posterior probability YRI, so the classification result obtained is CEU.

Table 11. Example of fictitious data for the use of the Naive Bayes method with 2 SNP locations

Respondent	Location of SNP 1	Location of SNP 2	Class
1	AA	TT	CEU
2	AA	TT	CEU
3	AA	TT	CEU
4	AG	TT	YRI
5	GG	AT	YRI
6	GG	AT	YRI

Look the example using Table 10 above shows that in the posterior probability of YRI, there is a multiplication by 0 which is the probability of AA with the condition YRI. it shows that the YRI class does not have an AA genotype marker at the SNP 1 location in the training dataset, whether the location will be used as a characteristic possessed by the CEU. But the Naive Bayes method will give poor results in some cases of classification if it uses several locations that have such characteristics, for example, consider the fictitious data in Table 11. Based on Table 11, the data classification will carry out on new objects with AA data at SNP 1 and AT data at SNP 2 locations. As said before, it would be necessary to look for the probability of the class used and the probability that may appear at each SNP location, so that is obtained:

$$P(CEU) = \frac{3}{6} = \frac{1}{2}, P(YRI) = \frac{3}{6} = \frac{1}{2}$$

The probabilities obtained at the location of SNP 1:

$$P(AA \mid CEU) = \frac{3}{3}1, P(AG \mid CEU) = \frac{0}{3} = 0, P(GG \mid CEU) = \frac{0}{3} = 0, P$$
$$(AA \mid YRI) = \frac{0}{3} = 0, P(AG \mid YRI) = \frac{1}{3}, P(GG \mid YRI) = \frac{2}{3}$$

Obtained

$$\text{Posterior probability of CEU } = P(CEU) \times P(AA \mid CEU)P(AT \mid CEU) = \frac{1}{2} \times 1 \times 0 = 0$$

$$\text{Posterior probability of YRI } = P(YRI) \times P(AA \mid YRI)P(AT \mid YRI) = \frac{1}{2} \times 0 \times \frac{2}{3} = 0$$

It showed that the values of the posterior probability of CEU and YRI are 0 so the classification result cannot be determined. Therefore, in carrying out classification using the Naive Bayes method on SNP data, it is necessary to sort the data first, especially in determining the training dataset to avoid obtaining a posterior probability value of 0 in both classes.

An experiment was carried out to find the average accuracy, average recall, average precision, and average F1 score of 1000 iterations using 1 to 10 SNP locations that have the same correlation interpretation by paying attention to the location of the SNP used, as well as the data division of 80% training dataset and 20% testing dataset. This experiment was carried out for each interpretation of correlations made in 1 graph to see the comparison. The SNP locations used in this experiment are the same as SNP locations used in the k-NN method shown in Table 9 and the results of this experiment are shown in Figure 5. In Figure 5, it can be seen that the correlation between the location of the SNP and the class used influences the classification results on the Naive Bayes classification method.

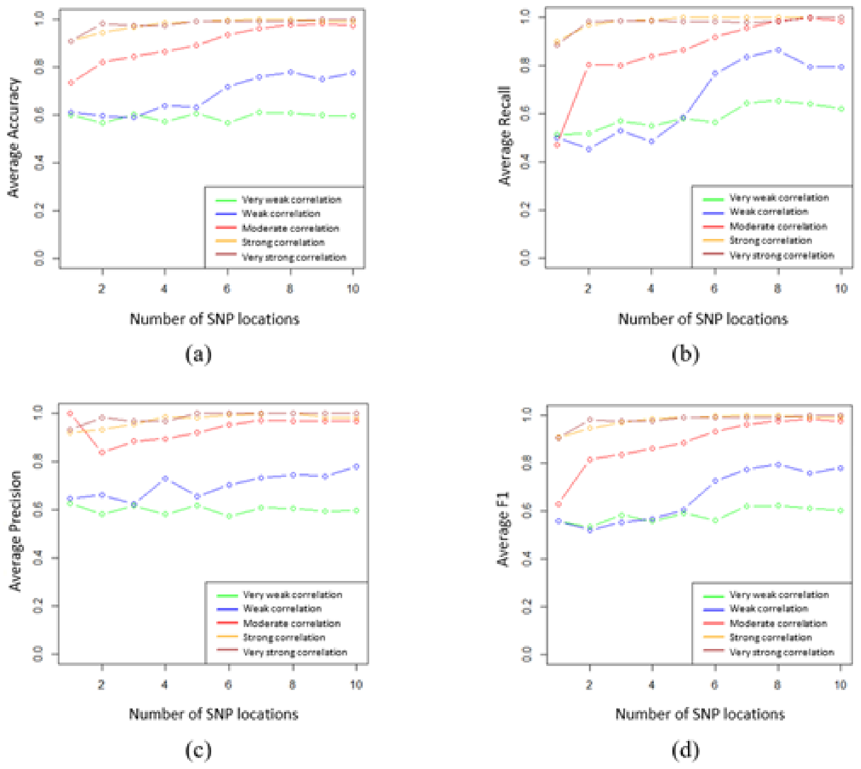


Figure 5. The results of the application of the Naive Bayes method (a) Graph of average accuracy with the number of SNP locations, (b) Graph of average recall with the number of SNP locations, (c) Graph of average precision with the number of SNP locations, (d) Average graph of F1 score with the number of SNP locations

The use of the Naive Bayes method was also carried out using 5, 10, and 15 SNP locations that finished in k-NN. The data used is also the same as the k-NN methods (shown in Table 8). The use of data division is 80% training dataset and 20% testing dataset. The accuracy value obtained from the classification process carried out 100 times is shown in Figure 6.

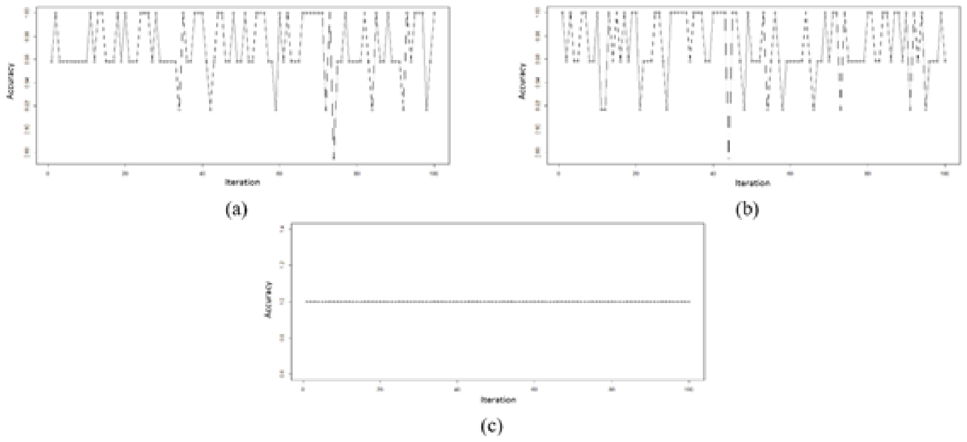


Figure 6. Accuracy of the Naive Bayes method using 5 SNP locations, (b) Accuracy of the Naive Bayes method using 10 SNP locations, (c) Accuracy of the Naive Bayes method using 15 SNP locations

3.4. Comparison of k-NN and Naive Bayes methods

Based on the description related to the use of k-NN and Naive Bayes methods that have been shown and the experiments carried out, one of the factors that influence the accuracy obtained is the locations of SNPs used based on correlation. So in comparing k-NN and Naive Bayes methods on SNP data to find out the tendency of a person's ancestor origin, the average accuracy and average F1 score are used, as well as the average value of the average accuracy and average F1 score with the following conditions:

1. The amount of SNP data used is 5 locations consisting of SNP locations with very weak, weak, moderate, strong, and very strong correlations, each of which has 1 location. The data used are shown in Table 12.
2. The training dataset used in the process classification using the Naive Bayes method should not create a posterior probability in both classes worth 0. Therefore, if a posterior probability of 0 is obtained in both classes, it is necessary to re-randomize the data in the training dataset and the testing dataset. Re-randomization of data is also performed if the posterior probability values obtained in both classes are the same.
3. The value of k used in the k-NN method is an odd number to avoid obtaining the same voting result. The value of k used is 31.
4. Data is divided into 4 distributions, namely 60% training dataset and 40% testing dataset, 70% training dataset and 30% testing dataset, 80% training dataset and 20% testing dataset, also 90% training dataset and 10% testing dataset.
5. The number of iterations is 1000 times by randomizing data in the distribution of the training dataset and the testing dataset in each iteration.

Table 12. Pembagian data untuk Training dan Testing

Reference SNP ID number location SNP	Correlation Interpretation
rs3887675	Very weak
rs1784176	Weak
rs4041435	Moderate
rs3136687	Strong
rs10868791	Very strong

Based on the process computational that has been carried out with k-NN and Naive Bayes methods, Figure 7 shows the first 100 values of accuracy and F1 score from the application of the two methods for 1000 iterations with data division of 80% training datasets and 20% testing datasets. Figure 7 shows there are many values of accuracy and the F1 score of the k-NN method is higher than the Naive Bayes method. If the values of accuracy and F1 score obtained from the iteration of 1000 times are averaged, results are obtained as shown in Table 13 and Table 14. The two tables also show the results of average accuracy and average F1 score from the data divisions of 60% training dataset and 40% testing dataset, 70% training dataset and 30% testing dataset, as well as 90% training dataset and 10% testing dataset.

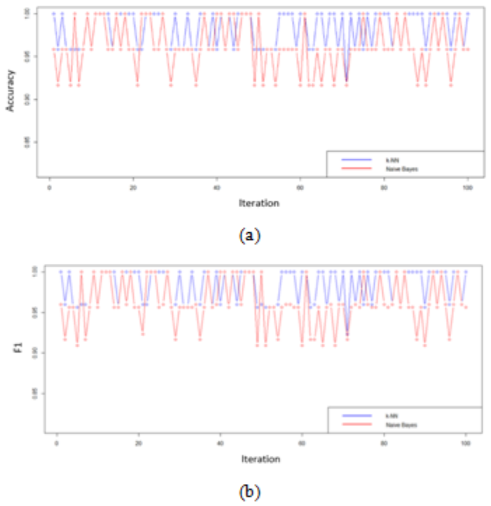


Figure 7. (a) Accuracy graph of k-NN and Naive Bayes methods, (b) F1 score graph of k-NN and Naive Bayes methods

Table 13. Average accuracy and average F1 score for each data division on the k-NN method

Data split		Average accuracy	Average F1 score
Training dataset (%)	Testing dataset (%)		
60	40	0.982708	0.983063
70	30	0.983889	0.983982
80	20	0.982917	0.982970
90	10	0.985833	0.985408

Table 14. Average accuracy and average F1 score for each data division on the Naive Bayes method

Data split		Average accuracy	Average F1 score
Training dataset (%)	Testing dataset (%)		
60	40	0.973333	0.972777
70	30	0.968611	0.968104
80	20	0.962463	0.961716
90	10	0.965049	0.962490

If the values of average accuracy and the values of average F1 score of 1000 iterations from all data divisions are re-averaged, the average accuracy obtained from the k-NN method is 0.9838 or 98.38% and the average F1 score is 0.9839 or 98.39%, while for the Naive Bayes method the average accuracy of 0.9674 or 96.74% and an average F1 of 0.9663 or 96.63% is obtained. Next, the average value of the average accuracy and the average F1 score is calculated to be one value in both methods. The average value of the average accuracy and average F1 score for the k-NN method is 0.98385 or 98.385% and for the Naive Bayes method is 0.96685 or 96.685%. It is noticed that the average accuracy and average F1 score on the k-NN method are greater than that of Naive Bayes, so the determination of the best method can be carried out without using the average value of average accuracy and average F1 score. Based on the values obtained, the k-NN classifier method is better than the Naive Bayes classifier method in classifying SNP data to find out the tendency of a person's ancestor origin.

Research related to the comparison between k-NN and Naive Bayes methods has been carried out previously with various kinds of data. In this study, a method comparison carries out by using the average accuracy and average F1 score obtained from several divisions of training datasets and testing datasets, as well as 1000 iterations. The k-NN method is better than Naive Bayes in classifying SNP data have two classes. The study [21] comparing k-NN and Naive Bayes methods for the classification of soil suitable for planting teak trees also obtained the same, the k-NN method is better than the Naive Bayes method. Research [22] also gains the accuracy value of the k-NN method is greater than the Naive Bayes method in text classification. Another thing with the research [23] obtained the Naive Bayes method is better than the k-NN method in classifying Indonesian articles. The study [24] which compared the methods of Naive Bayes, Decision Tree, and k-NN in finding alternative designs for energy simulation tools also obtained the accuracy value of the Naive Bayes method was better than the k-NN method. Likewise, the study [25] compared the three methods using data in the form of the word that found that the accuracy of the Naive Bayes method was better than the k-NN method with a little difference, namely 100% for the Naive Bayes method and 98.25% for the k-NN. The difference in the results of the comparison by those methods is certainly due to several factors. When viewed from previous studies and this study, it can be said that some of those factors are the characteristics of the data used, the data cleaning process carried out, and the quantity of data used.

4. CONCLUSION

Based on the evaluations that have been carried out on both methods of classifying SNP data that have two classes (CEU and YRI) with 1000 iterations the average accuracy value of 98.38% with the k value of 31 and the average F1 score of 98.39% obtained for the k-NN method, while the Naive Bayes method obtained the average accuracy of 96.74% and the average F1 score of 96.63%. However, the k-NN method is better than the Naive Bayes method of conducting SNP data classification to find the tendency of a person's ancestor origin.

Further research related to the classification of SNP data to determine the tendency of a person's ancestor origin can be done using other classification algorithms, such as Random Forest, Logistic Regression, and Neural Network. It can also be used the Naive Bayes method combined with the k-NN [26]. In the application of the k-NN method for SNP data classification, other distance formulas can use, such as the distance of Manhattan and the distance of Minkowski, then determining the location of SNP with correlation can also be used Spearman and Kendall correlations. The SNP data used in the classification to determine the tendency of a person's ancestor origin can be added with new data from other classes. In this case, other classes besides CEU and YRI, so the classification results obtained will be better if used to find out the origin of a person's ancestor origin.

5. ACKNOWLEDGEMENTS

The researchers would like to thank the Mathematics Study Program, Faculty of Science and Mathematics, Universitas Kristen Satya Wacana for providing facilities to conduct this research.

6. DECLARATIONS

AUTHOR CONTIBUTION

Denny Indrajaya carried out the fieldwork, wrote, and revised the manuscript, Adi Setiawan provided research idea, reviewed, and revised the manuscript, Bambang Susanto provided reference files and reviewed the parameters in research.

FUNDING STATEMENT

This research was funded by Universitas Kristen Satya Wacana.

COMPETING INTEREST

The authors declare that there are no conflicts of interest regarding the publication of this paper.

REFERENCES

[1] M. Tamimi, "Tes DNA dalam Menetapkan Hubungan Nasab," *Al-Istinbath: Jurnal Hukum Islam*, vol. 13, no. 1, pp. 83–98, 2014.

[2] R. Mathur, B. S. Rana, and A. K. Jha, "Single Nucleotide Polymorphism (SNP)," in *Encyclopedia of Animal Cognition and Behavior*. United States of America: Springer Cham, 2018.

[3] X. Ding and X. Guo, "A Survey of SNP Data Analysis," *Journal of Big Data Mining and Analytics*, vol. 1, no. 3, pp. 173–190, 2018.

[4] A. A. Komar, *Methods Single Nucleotide Polymorphisms - Methods and Protocols*, 2nd ed. United States of America: Humana Press, 2009.

[5] J. Ren, D. Sun, L. Chen, F. M. You, J. Wang, Y. Peng, E. Nevo, D. Sun, M. C. Luo, and J. Peng, "Genetic Diversity Revealed by Single Nucleotide Polymorphism Markers in a Worldwide Germplasm Collection of Durum Wheat," *International Journal of Molecular Sciences*, vol. 14, pp. 7061–7088, 2013.

[6] E. Salwati, S. Handayani, and R. P. Jekti, "Identifikasi Single Nucleotide Polymorphism (SNP) Gen pvmdr1 pada Penderita Malaria Vivaks di Minahasa Tenggara (Sulawesi Utara)," *Jurnal Biotek Medisiana Indonesia*, vol. 3.2, pp. 49–57, 2014.

[7] A. Putri and S. Wathon, "Aplikasi Single Nucleotide Polymorphism (SNP) dalam Studi Farmakogenomik untuk Pengembangan Obat," *Jurnal BioTrends*, vol. 9, no. 2, pp. 69–74, 2018.

[8] Triwani and I. Saleh, "Single Nucleotide Polymorphism Promoter -765g /C Gen Cox-2 sebagai Faktor Risiko Terjadinya Karsinoma Kolorektal," *Biomedical Journal of Indonesia*, vol. 1, no. 1, pp. 2–10, 2015.

[9] V. D. M. Butarbutar, A. Setiawan, and T. Mahatma, "Analisis Data SNP (Single Nucleotide Polymorphism) dengan Metode Chi-Square," in *Prosiding Seminar Nasional Matematika dan Pendidikan Matematika (Sendika) 2020*, vol. 6, no. 1, 2020, pp. 97–103.

[10] J. Gaudillo, J. Joseph Russell Rodriguez, A. Nazareno, L. Rigi Baltazar, J. Vilela, R. Bulalacao, M. Domingo, and J. Albia, "Machine Learning Approach to Single Nucleotide Polymorphism-Based Asthma Prediction," *Journal of PLOS ONE*, vol. 14, no. 12, 2019.

[11] F. Bertolini, G. Galimberti, G. Schiavo, S. Mastrangelo, R. Di Gerlando, M. G. Strillacci, A. Bagnato, B. Portolano, and L. Fontanesi, “Preselection Statistics and Random Forest Classification Identify Population Informative Single Nucleotide Polymorphisms in Cosmopolitan and Autochthonous Cattle Breeds,” *Journal of Animal*, vol. 12, no. 1, pp. 12–19, 2018.

[12] N. Batnyam, A. Gantulga, and S. Oh, “An Efficient Classification for Single Nucleotide Polymorphism (SNP) Dataset,” in *Studies in Computational Intelligence*. Heidelberg: Springer, 2013, vol. 493, pp. 171–185.

[13] S. N. Kamalina, “Identifikasi Single Nucleotide Polymorphism (SNP) pada Genom Kedelai Menggunakan Algoritme C5.0,” Ph.D. dissertation, 2018.

[14] Paiman, *Korelasi dan Regresi Ilmu-Ilmu Pertanian*. Yogyakarta: UPY Press, 2019.

[15] P. Schober, L. A. Schwarte, and C. Boer, “Correlation Coefficients: Appropriate Use and Interpretation,” *Journal of Anesthesia and Analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018.

[16] D. Napitupulu, R. Rahim, D. Abdullah, M. I. Setiawan, L. A. Abdillah, A. S. Ahmar, J. Simarmata, R. Hidayat, H. Nurdiyanto, and A. Pranolo, “Analysis of Student Satisfaction Toward Quality of Service Facility,” *Journal of Physics: Conference Series*, vol. 954, pp. 1–7, 2018.

[17] M. R. Faisal and D. T. Nugrahedi, *Belajar Data Science: Klasifikasi dengan Bahasa Pemrograman R*. Banjarbaru: Scripta Cendekia, 2019.

[18] R. T. Vulandari, *Data Mining Teori dan Aplikasi Rapidminer*. Yogyakarta: Gava Media, 2017.

[19] D. Berrar, “Bayes’ Theorem and Naive Bayes Classifier,” in *Encyclopedia of Bioinformatics and Computational Biology*. Elsevier, 2018, vol. 1, pp. 403–412.

[20] F. Gorunescu, *Data Mining Concepts, Models and Techniques*. New York: Springer-Verlag Berlin Heidelberg, 2011.

[21] D. Srianto and E. Mulyanto, “Perbandingan K-Nearest Neighbor dan Naive Bayes,” *Jurnal Techno.COM*, vol. 15, no. 3, pp. 241–245, 2016.

[22] A. Indriani, “Analisa Perbandingan Metode Naïve Bayes Classifier dan K-Nearest Neighbor Terhadap Klasifikasi Data,” *Jurnal SEBATIK*, vol. 24, no. 1, 2020.

[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,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 4, pp. 427–434, 2018.

[24] A. Ashari, I. Paryudi, and A. M. Tjou, “Performance Comparison Between Naïve Bayes, Decision Tree and K-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool,” *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 11, pp. 33–39, 2013.

[25] M. K. Anam, B. N. Pikir, M. B. Firdaus, S. Erlinda, and Agustin, “Penerapan Na ve Bayes Classifier, K-Nearest Neighbor (KNN) dan Decision Tree untuk Menganalisis Sentimen pada Interaksi Netizen dan Pemeritah,” *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 1, pp. 139–150, 2021.

[26] Y. F. Safri, R. Arifudin, and M. A. Muslim, “K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor,” *Scientific Journal of Informatics*, vol. 5, no. 1, pp. 10–18, 2018.