Paradigma, Vol. 26, No. 1, Maret 2024, P-ISSN 1410-5063, E-ISSN: 2579-3500, Page. 24-29
Published: LPPM Universitas Bina Sarana Informatika
PARADIGMA
website: http://jurnal.bsi.ac.id/index.php/paradigma/

Comparison of Supervised Learning Classification Methods on Accreditation Data of Private Higher Education Institutions

Noviyanto[1], Mochamad Wahyudi[2], Sumanto[3]

[1,2] Department of computer science, Universitas Gunadarma, Depok, Indonesia
[2,3] Department of computer science, Universitas Bina Sarana Informatika, Jakarta, Indonesia

ARTICLE INFORMATION

Artikel History:
Received: March 9, 2024
Revised: March 13, 2024
Accepted: March 20, 2024

Keyword:
Accreditation Data, logistic regression, K-nearest neighbor, naive bayes, super vector machine, random forest.

ABSTRACT

This research aims to analyze and compare supervised learning classification methods using a case study of accreditation data for private higher education institutions within the LLDikti Region III contained in BAN-PT. In addition, The problem in this research is how each supervised learning classification method operates and how accurate and precise the results given by each method are. this research also uses Weka machine learning software in its calculations. The initial step taken is to prepare the software used for supervised learning analysis, then pre-processing the data, namely labeling data that has a categorical data type, after that determining data for testing data. The next step is to test each classification method. The methods used for comparison are logistic regression, K-nearest neighbor, naive bayes, super vector machine, and random forest. Based on the calculation results, the Kappa Statistic and Root mean squared error values obtained are 1 and 0 for the logistic regression method, 0.979 and 0.0061 for the K-nearest neighbor method, 1 and 0.2222 for the super vector machine method, 0.969 and 0.0341 for the naive bayes method, 1 and 0 for the decision tree method, and 0.5776 and 0.1949 for the random forest method, respectively. The logistic regression and decision tree methods in this study get Kappa Statistic and Root mean squared error values of 1 and 0 respectively so that they are said to be good and acceptable, thus the two classification methods are the most appropriate methods and are considered to have the highest accuracy.

Corresponding Author:
Noviyanto,
Department of computer science,
Universitas Gunadarma,
Jl. Margonda No.100 Kampus D, Pondok Cina, Kecamatan Beji, Kota Depok, Jawa Barat 16431
Email: viyan@staff.gunadarma.ac.id

INTRODUCTION

In order to guarantee and improve the quality of national education in a gradual, planned, and measurable manner as mandated by Law Number 12 of 2012 concerning Higher Education Article 55 paragraphs 1 to 8, the Government conducts accreditation to assess the feasibility of Study Programs and Higher Education on the basis of criteria that refer to the National Higher Education Standards. In this regard, the government has established the National Higher Education Accreditation Board as an accreditation agency authorised by the government to improve the quality of higher education.

Classification is a science found in machine learning. Classification is a method that can handle big data. Classification in machine learning is the grouping of data where the data used have a label or target class. Therefore, algorithms used to solve classification problems are categorised as supervised learning. There are many methods that exist in supervised learning classification, including logistic regression, K-nearest neighbour, super vector machine, naive Bayes, decision tree, and random forest.

The formulation of the problem in this study is because each supervised learning classification method has its own advantages, disadvantages, and accuracy. Through university accreditation data contained on the National Accreditation Board for Higher Education (BAN-PT) page, it is known how the supervised learning classification method works and how accurate

DOI: https://doi.org/10.31294/p.v26i1.3306

and precise it is. So that the discussion in this study is not too broad and more focused, it is limited to classifying accreditation data for Private Higher Education institutions within the Higher Education Service Institution (LLDikti) Region III using the Weka 3.8.6 application. This study aims to discuss the accuracy of the classification of each method using accreditation data of private higher education institutions in the LLDikti Region III environment obtained from the BAN-PT website. From the results of the classification, the most appropriate and accurate classification method was determined.

1. WEKA Machine Learning Software

Weka is a practical machine learning tool. "Waikato Environment for Knowledge Analysis" or known as WEKA was created at the University of Waikato, New Zealand, which is devoted to supporting the fields of research, education and various applications in data mining. The software is built using Java classes with object-oriented methods and can be run on almost all platforms. Weka is easy to apply at several different levels. Weka provides an implementation of state-of-the-art learning algorithms that can be applied to datasets from the command line. In WEKA, there are tools that are useful for data preprocessing, classification, regression, clustering, association rules, and visualisation. It can be used to preprocess data, incorporate it into a learning scheme, and analyse the classifier generated by its performance, all without writing the program code. One example of using WEKA is applying a learning method to a dataset and analysing the results to gain information about the data, or applying several methods and comparing their performance to select the best one.

2. Logistic Regression

Logistic regression is a statistical analysis method used to describe the relationship between a response variable (dependent variable) that has two or more categories and one or more explanatory variables (independent variables) on a categorical or interval scale (Hosmer and Lemeshow, 2000). Logistic regression is a non-linear regression used to explain the relationship between X and Y that is not linear, the non-normality of the Y distribution, and the non-constant response diversity that cannot be explained by ordinary linear regression models (Agresti, 1996).

3. *K-Nearest Neighbor*

The KNN method is an easy classification method. This method works by finding k patterns (among all training patterns in all classes) that are closest to the input pattern and then determining the decision class based on the largest number of patterns (Suyanto, 2018). The KNN training process produces k, which provides the highest accuracy in generalising future data. The problem is that, until now, k cannot be determined mathematically. Therefore, the training process is basically observing a number of k until the most optimum k is produced.

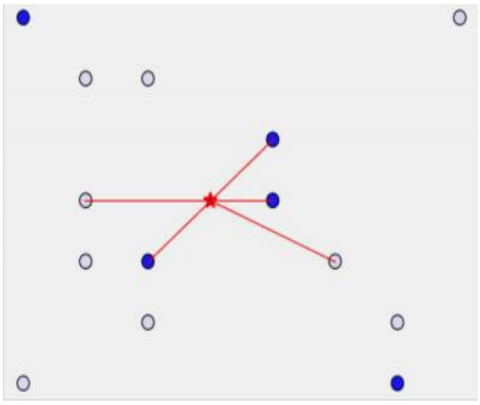

Figure 1. K-Nearest Neighbor Visualization

4. Super Vector Machine

The SVM method aims to determine the optimal hyperplane. Hyperplane that can divide the two classes with the furthest margin distance between the classes. The margin is the distance between the hyperplane and the closest pattern from each class. This closest instance is called a support vector. The red line above the thick black line can be an instance with a "+" sign which is the support vector for the Men class. While on the red line below the thick black line there is an instance with an "o" sign which is the support vector for the Women class. Therefore, it can be concluded that the main purpose of SVM is to find the best hyperplane with the help of support vectors from each class so that the optimal hyperplane is obtained.

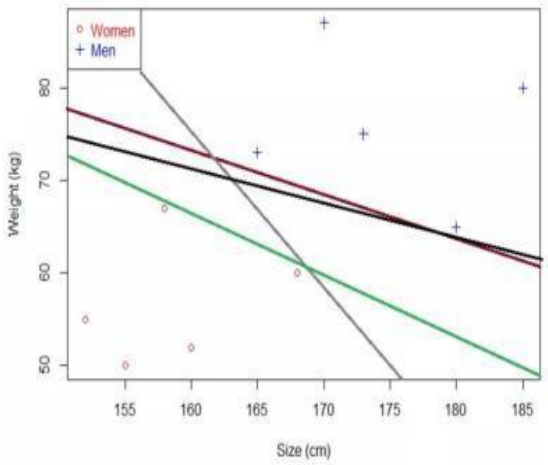

Figure 2a. Hyperlane visualization is not optimal

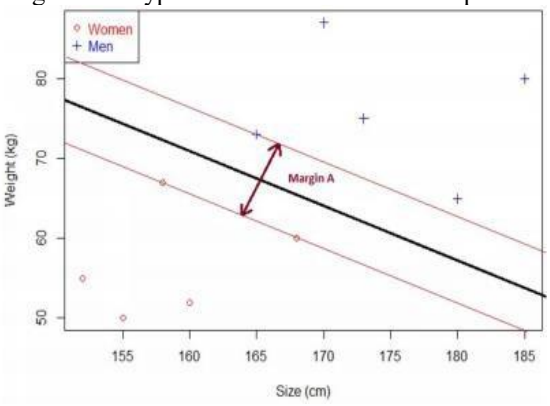

Figure 2b. Maximal hyperlane visualization

4. Naive Bayes

The Bayes classifier is a statistical classifier that can predict the probability of class membership of a data tuple that will enter a certain class, according to probability calculations. The Bayes classifier was based on the Bayes theorem discovered by Thomas Bayes in the 18th century. In the study of classification algorithm comparison, a simple Bayesian or Naïve Bayes classifier has been found. Naïve Bayes classifier shows high accuracy and speed when applied to large databases. This method is often used in solving problems in the field of machine learning because it is known to have a high level of accuracy with simple calculations.

5. Decision Tree

A decision tree is a data-mining classification method. A decision tree in learning terms is a tree structure in which each tree node represents an attribute that has been tested. Each branch is a division of test results, and leaf nodes represent certain class groups. [5]. The top level node of a Decision Tree is the root node which is usually the attribute that has the greatest influence on a particular class. In general, a Decision Tree performs a top-down search strategy for the solution. In the process of classifying unknown data, attribute values will be tested by tracing the path from the root node to the final node (leaf), and then the class belonging to a particular new data will be predicted.

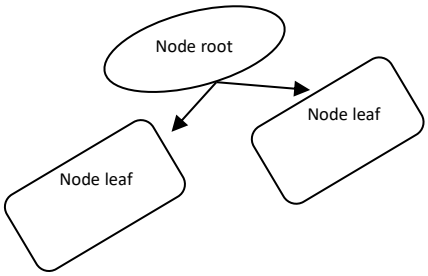

Figure 3. Decision Tree Visualization

6. Random Forest

Random Forest is a development of the Decision Tree method that uses several Decision Trees, where each Decision Tree has been trained using individual samples and each attribute is broken down on a tree selected between random subset attributes. Random Forest has several advantages, namely it can improve accuracy results if there is missing data, it can resist outliers, and it is efficient for storing data. In addition, Random Forest has a feature selection process where it is able to take the best features so that it can improve the performance of the classification model. With feature selection, Random Forest can effectively work on big data with complex parameters.

7. 10-Fold Cross-Validation

Cross-validation is a statistical method to evaluate and compare learning algorithms by dividing the data into two segments: one used to learn or train the model and the other used to validate the model. The way K-Fold Cross-validation works is as follows:

1. The entire dataset is divided into K parts.
2. The 1st fold is when the 1st part becomes the testing data and the rest becomes the training data. Next, we calculated the accuracy based on that portion of the data.
3. The 2nd fold is when the 2nd part becomes the testing data and the rest becomes the training data. Next, we calculated the accuracy based on that portion of the data.
4. This process continues until it reaches the kth fold.
5. The average accuracy of the N accuracies is calculated.
6. The average accuracy is the final accuracy.

Figure 1 illustrates the scheme of 10-fold cross-validation. The data are divided into 10 folds of equal size, so we have 10 subsets of data to evaluate the performance of the algorithm. For each of the 10 subsets of data, cross-validation will use nine folds for training data and one fold for test data..

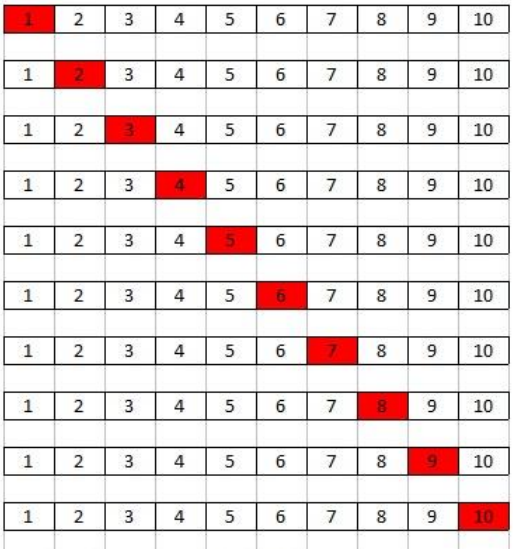

Figure 4. 10-Fold Cross-validation Scheme

8. Kappa Statistik

The kappa statistic is a numerical index of agreement between two raters classifying the same item, where it ranges between 0 (no agreement) and 1 (perfect agreement) and takes into account chance agreement. Based on Fleis (1981), the interpretation of Kappa values is presented in the following table:

Table 1. Kappa Interpretation

The Kappa Index	Agreement
< 0.40	Bad
0.40 - 0.60	Fair
0.60 - 0.75	Good
> 0.75	Excellent

9. Root mean squared error

The Root Mean Square Error (RMSE) is the magnitude of the prediction error rate, where the smaller (closer to 0) the RMSE value, the more accurate the prediction results. The RMSE value can be calculated using the following equation:

$$RMSE=\sqrt{\frac{\sum(X-Y)^2}{n}} \quad (1)$$

where n is the number of data points.

RESEARCH METHOD

In this research design, the author will describe the methodology and framework of the research work used in solving research problems. This research methodology is used systematically to obtain a good workflow so that the results achieved do not deviate from the desired goals and are carried out properly, in accordance with the predetermined objectives. The following research model flow chart is presented in a flowchart design which can be seen from the following figure:

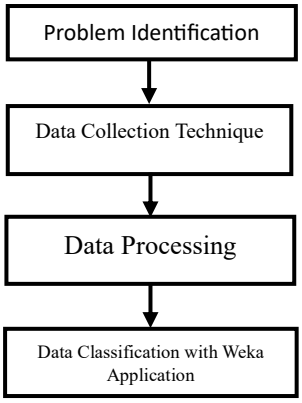

Figure 5. Flowchart of Research Framework

The flowchart in Figure 2 is as follows:

1. Problem Identification

Introduction to a problem and the initial stage in the research process. Identifying data on Accreditation of Private Higher Education Institutions in the LLDikti Region III environment in the Higher Education Database (PDDikti).

2. Data Collection Technique

The research data are obtained from the following page with the URL address:
https://www.banpt.or.id/direktori/institusi/pencarian_institusi.php
which is a data report on the results of accreditation of Higher Education institutions.

3. Performing Data Processing

At this stage, data processing is carried out to obtain results, which can then be managed to the next stage to produce the right information.

4. Data classification with the Weka application

In testing the data, a Weka application is used which is a series of machine learning software. Using the Weka application, the results of data processing will be compared between several classification methods, namely logistic regression, K-nearest neighbour, naive Bayes, super vector machine, and random forest.

RESULTS AND DISCUSSION

1. Data Import and Labeling

The first thing to do for analysis is to import the accreditation data of Higher Education institutions in *. xlsx format into the Weka application and then do labeling on categorical data, especially on the target variable (exited):

Table 2. Labeling data by giving scores on accreditation of higher education institutions

Accreditation	Score
Superior	4
A	3,5
Excellent	3
B	2,5
Good	2
-	0

Table 3. Accreditation Data of Higher Education Institutions

NO	KODE PT	NAMA PT	JENIS PT	BENTUK PT	STATUS PT	AIPT	SKOR
1	031001	Universitas Ibnu Chaldun	Akademik	Universitas	A	Baik	2
2	031003	Universitas Islam Jakarta	Akademik	Universitas	A	B	2,5
3	031005	Universitas Jakarta	Akademik	Universitas	A	Baik	2
4	031006	Universitas Jayabaya	Akademik	Universitas	A	B	2,5
5	031007	Universitas Katolik Indonesia Atma Jaya	Akademik	Universitas	A	Unggul	4
.....	.....	.....	.....	.....	.....	.....	.....
264	035018	Politeknik Jakarta Internasional	Vokasi	Politeknik	A	Baik	2
265	035020	Politeknik Tempo	Vokasi	Politeknik	A	-	0
266	035021	Politeknik Astra	Vokasi	Politeknik	A	B	2,5
267	035022	Politeknik Multimedia Nusantara	Vokasi	Politeknik	A	-	0
268	035023	Politeknik Kartini Jakarta	Vokasi	Politeknik	A	Baik	2
269	035024	Politeknik Kreatif Indonesia	Vokasi	Politeknik	A	-	0
270	035025	Politeknik Prestasi Prima	Vokasi	Politeknik	A	-	0
271	036001	Akademi Komunitas Kosmetik Ristra	Vokasi	Akademi Komunitas	A	-	0
272	036002	Akademi Komunitas Bisnis Internasional	Vokasi	Akademi Komunitas	A	-	0

From the table above, we preprocessed the data and obtained a visualisation, as shown in Figure 2 below:

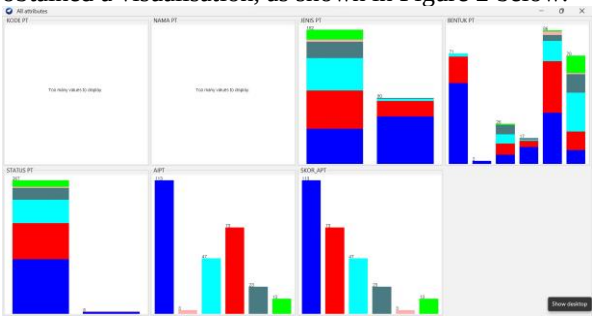

Figure 6 Visualization of accreditation data preprocessing

1. **Testing data for each classification method.**
   The steps taken in the Weka application to perform calculations after preprocessing the data are as follows:

a. Select the Classify menu and the classification method on the Choose button in the Classifier window. For example, select the function directory and the Random Forest classification method to perform the classification process using the logistic method.

b. In the Test options checkbox button, select Cross-validation with Folds 10.

c. Select the start button and run the calculation to produce Run Information as follows:

```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomForest -P 100 -I 100 -
num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Relation:     WekaExcel
Instances:    272
Attributes:   7
              KODE PT
              NAMA PT
              JENIS PT
              BENTUK PT
              STATUS PT
              AIPT
              SKOR_APT
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-
not-check-capabilities

Time taken to build model: 0.07 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         198               72.7941 %
Incorrectly Classified Instances        74               27.2059 %
Kappa statistic                          0.5776
Mean absolute error                      0.1949
Root mean squared error                  0.2827
Relative absolute error                 81.3127 %
Root relative squared error             81.8136 %
Total Number of Instances              272
```

e. Perform the same process according to the steps in letters a to c to perform the classification calculation process with logistic regression, K-nearest neighbour, naive Bayes, super vector machine, and decision tree methods. After performing all the above processes, we obtained a visualisation image of classifier errors, visualisation tree, and data table of test results with 10-Folds Cross Validation with each classification method, as shown in the figure and table below:

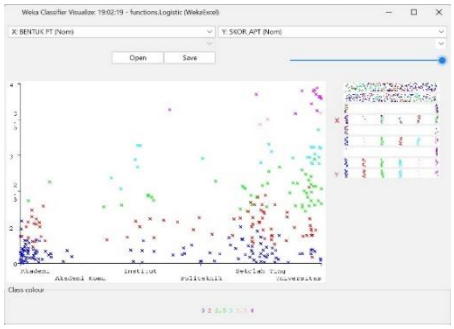

Figure 7. Visualisation of classifier errors for one of the classification methods: Logistic Regression.

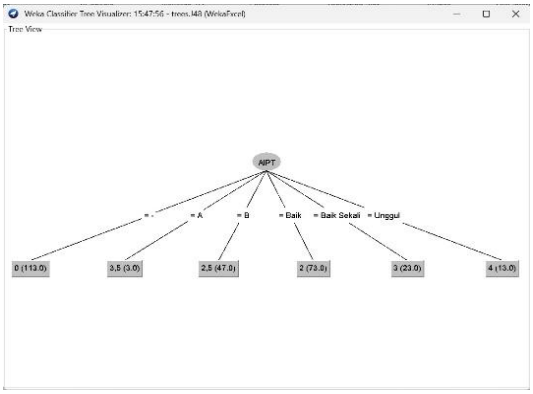

Figure 8. Visualization of the tree in the Decision tree classification method

Table 4. Test Result Data with 10-Folds Cross-validation

Metode Klasifikasi	Correctly Classified Instances		Incorrectly Classified Instances		Kappa statistic	Root mean squared error
Random Forest	198	72,7941%	74	27,2059%	0,5776	0,2827
Regresi Logistik	272	100%	0	0%	1	0
K-Nearest Neighboor	268	98,5294%	4	1,4706%	0,9794	0,0572
Super Vector Machine	272	100%	0	0%	1	0,3103
Naive bayes	266	97,7941%	6	2,2059%	0,969	0,0807
Decision Tree	272	100%	0	0%	1	0

From Table 4, the Kappa Statistic and Root mean squared error values are, respectively, 1 and 0 for the logistic regression method, 0.979 and 0.0061 for the K-nearest neighbour method, 1 and 0.2222 for the super vector machine method, 0.969 and 0.0341 for the naive Bayes method, 1 and 0 for the decision tree method, and 0.5776 and 0.1949 for the random forest method.

CONCLUSION

Based on the discussion, the best method for classifying accreditation data for private higher education institutions in the LLDikti Region III Jakarta environment is the Logistic Regression and Decision Tree classification method. This is because the Kappa Statistic and Root mean squared error values are 1 and 0, respectively, so that they are said to be good and acceptable. Thus, the two classification methods are the most appropriate and are considered to have the highest accuracy.

REFERENCES

Fleiss, J. L. (1981) Statistical methods for rates and proportions. 2nd ed. (New York: John Wiley)

Fultriasantri, Indah, & Fajrin (2023). Pemanfaatan Penginderaan Jauh Untuk Mengidentifikasi Kepadatan Bangunan Menggunakan Interpretasi Hibrid Citra Sentinel-2a Di Kota Padang. Jurnal Environmental Science, 6(1), https://doi.org/10.35580/jes.v5i2.43339

I. Kusniawati, S. Subiyanto, & F. J. Amarrohman, "Analisis Model Perubahan Penggunaan Lahan

Menggunakan Artificial Neural Network Di Kota Salatiga," Jurnal Geodesi Undip, 9(1), pp. 1-11, Dec. 2019. https://doi.org/10.14710/jgundip.2020.26026

Handayani, F. & Pribadi, F. (2017). Implementasi Algoritma Naive Bayes Classifier dalam Pengklasifikasian Teks Otomatis Pengaduan dan Pelaporan Masyarakat melalui Layanan Call Center 110. Jurnal Teknik Elektro, 7(1), 19-24. doi:https://doi.org/10.15294/jte.v7i1.8585

Kharisudin, Iqbal., Fajar, Sodik Pamungkas., & Prasetya, Bayu Dwi. Perbandingan Metode Klasifikasi Supervised Learning pada Data Bank Customers Menggunakan Python, PRISMA-Prosiding Seminar Nasional Matematika, vol. 3, pp. 692-697, Mar. 2020.

Nasrullah, Asmaul Husnah. (2021). Implementasi Algoritma Decision Tree Untuk Klasifikasi Produk Laris. Jurnal Ilmiah Ilmu Komputer Fakultas Ilmu Komputer Universitas Al Asyariah Mandar, 7(2). https://doi.org/10.35329/jiik.v7i2.203

Napitupulu, D. B. (2015). Studi Validitas Dan Realibilitas Faktor Sukses Implementasi E-Government Berdasarkan Pendekatan Kappa. Jurnal Sistem Informasi, 10(2), 70 - 74. https://doi.org/10.21609/jsi.v10i2.388

Sanjaya, Fadil Indra., Heksaputra, Dadang. (2020). Prediksi Rerata Harga Beras Tingkat Grosir Indonesia dengan Long Short Term Memory. JATISI (Jurnal Teknik Informatika dan Sistem Informasi), 7(2), 163-174. https://doi.org/10.35957/jatisi.v7i2.388

Setiawati, Noor Lusty Putri.& Utomo, Agung Priyono. (2017). Model Regresi Logistik Untuk Melihat Pengaruh Faktor Demografis, Self Efficacy, Terhadap Perilaku Mencontek, 6(2). http://dx.doi.org/10.15408/jp3i.v6i2.9172

Supriyadi, R., Gata, W., Maulidah, N., Fauzi, A. (2020). Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah. E-Bisnis: Jurnal Ilmiah Ekonomi Dan Bisnis, 13(2), 67 - 75. https://doi.org/10.51903/e-bisnis.v13i2.247

Sutoyo, I. (1). Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik. Jurnal Pilar Nusa Mandiri, 14(2), 217-224. https://doi.org/10.33480/pilar.v14i2.70

Syarli., & Muin, Asrul Ashari. (2016). Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus: Data Mahasiswa Baru Perguruan Tinggi), Jurnal Ilmiah Ilmu Komputer Fakultas Ilmu Komputer Universitas Al Asyariah Mandar, 2(1). https://ejournal.fikom-unasman.ac.id/index.php/jikom/issue/view/3

Witten, Ian H. The WEKA Workbench Fourth Edition Data Mining Practical Machine Learning Tools and Techniques, 2016

https://www.banpt.or.id/direktori/institusi/pencarian_institusi.php

https://c3.ai/glossary/data-science/root-mean-square-error-rmse/