Jurnal Teknik Informatika (JUTIF)
Vol. 4, No. 3, June 2023, pp. 493-501

DOI: https://doi.org/10.52436/1.jutif.2023.4.3.737
p-ISSN: 2723-3863
e-ISSN: 2723-3871

IMPROVING PERFORMANCE OF STUDENTS' GRADE CLASSIFICATION MODEL USES NAÏVE BAYES GAUSSIAN TUNING MODEL AND FEATURE SELECTION

M. Hafidz Ariansyah[*1], Esmi Nur Fitri[2], Sri Winarno[3]

[1,2,3]Information Systems Department, Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia
Email: [1]112201906146@mhs.dinus.ac.id, [2]112201906175@mhs.dinus.ac.id, [3]sri.winarno@dsn.dinus.ac.id

(Article received: December 07, 2022; Revision: December 28, 2022; published: June 26, 2023)

Abstract

Student grades are a relevant variable for predicting student academic performance. In achieving good and quality student performance, it is necessary to analyze or evaluate the factors that influence student performance. When a educator can predict students' academic performance from the start, the educator can adjust the way of learning so that learning can run effectively. The purpose of this research is to study how it is applied to determine the interrelationships between variables and find out which variables have an effect, then use it as a feature selection technique. Then, researchers review the most popular classifier, Gaussian Naïve Bayes (GNB). Next, we survey the feature selection models and discuss the feature selection approach. In this study, researchers will classify student grades based on existing features to evaluate student performance, so it can guide educators in selecting learning methods and assist students in planning the learning process. The result is that applying Gaussian Naïve Bayes (GNB) without feature selection has a lower accuracy of 10.12% while using feature selection the accuracy increases to 10.12%.

Keywords: *classification, features selection, gaussian naïve bayes, student grade.*

1. INTRODUCTION

Students' academic achievement in schools is extensively researched to address underachievement, delays in graduation, and other challenges [1]. In simple terms, student performance refers to the extent to which short-term and long-term goals are achieved in education [2]. However, academics measure student success from various perspectives, starting from final grade and student activity both inside and when participating in school activities such as extracurriculars [3]. The literature offers many computational efforts that seek to improve student performance in schools, especially those driven by data mining and learning analytic techniques [4]. However, there is still confusion regarding the effectiveness of existing technology and witty models. Precise classification of student performance enables the detection of underperforming students, thereby empowering educators to intervene early during the learning process and implement the necessary interventions. Advantageous interventions include, but are not limited to, student advising, monitoring of performance progress, development of intelligent tutoring systems, and policy-making [5]. This effort has been greatly driven by computational advances in data mining and learning analytics [6].

The lack of analysis that investigates the classification of student performance using student outcomes has motivated researchers to pursue the aims of this study. In a systematic literature review (i.e., SLR), the step-by-step protocol is executed to identify, select, and rate studies synthesized to answer specific research questions [2]. Here the researcher must deeply understand the intelligent approaches and techniques developed to predict student learning outcomes, which represent student academic performance. One way is to determine the dominant predictors (eg, factors and features) of student learning outcomes based on evidence from the synthesis. Identify research challenges and limitations facing current intelligent techniques for predicting academic performance using learning outcomes.

Therefore, the final grades of students remain an important thing to predict. The classification results can be used as a reference for schools in the process of making policies related to improving the quality of education in the future. Classification results can also be used as an initial evaluation of students in preparation for the actual test to get optimal results later [7].

This study utilizes the Gaussian Naïve Bayes algorithm with feature selection to classify the grades obtained by students so that educators intervene early during the learning process and apply the necessary interventions. This research will look for the best accuracy value of the Gaussian Naïve Bayes algorithm with and without feature selection. The purpose of this research is to compare the accuracy of the classification model pattern so that it can be the basis for decision-making for educators so that students can improve their academic achievement.

2. METODE PENELITIAN

The processes and structures involved in the research are broadly shown in Figure 1.

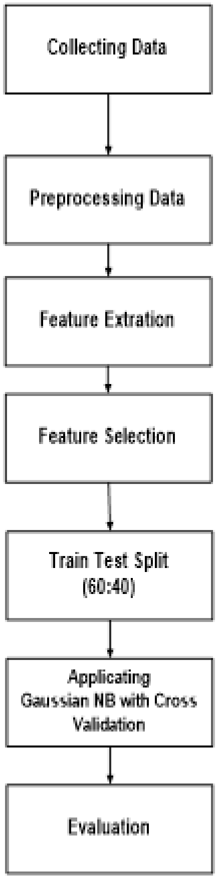

Figure 1. Research Overview

To uncover patterns or possibilities in various data sets, this study used data mining, feature engineering, and feature selection approaches. In a word, it is the research of a model to predict student grades. Selection of the best features from existing data using chi from the scikit-learn library to choose the best features; training a Gaussian Nave Bayes model with 60% of data in dataset and 10-fold cross-validation; and confusion matrix analysis to verify the model's performance accuracy. As indicated in Figure 1, the phases in this study comprised seven processes: data collection, data pre-processing, feature engineering, feature selection, dataset division into training and testing, algorithm implementation with cross validation, and evaluation.

2.1. Collecting Data

The main contribution of this research is to implement this elegant learning algorithm on the data set from the UCI machine learning repository to observe the best available model. Student data is the data used in classification and obtained from UCI [8]. The data that will be used has 20 features and 395 data. The description of the attributes and dataset (see Table 1) used is as follows:

1. Gender - Male/female
2. Age - (in years)
3. Address - Urban / Rural
4. Family Size - (number of people in the family)
5. Status of Parents - Living together/separated
6. Mother's Education
7. Father's education
8. Mother's job
9. Father's job
10. Family Relations - (from 1-very bad to 5-very good)
11. Health - (from 1-very bad to 5-very good)
12. Absences - The number of student absences
13. Internet - Yes / No (home access)
14. Free Time - (from 1-very little to 5-very much)
15. Playing Time - (from 1-very little to 5-very much)
16. Extracurriculars - Yes/no
17. Study Time - Weekly study hours
18. Grades1
19. Grades2
20. Grades3

Table 1. Dataset

Features	Value	….	Value
Sex	F	…	M
Age	18	…	19
Address	Urban	…	Urban
Sizefam	More than 3	…	Less than 3
StatusParent	Apart	…	Together
Mother’s Edu	University	…	Elementary
Father’s Edu	University	…	Elementary
Mother’s Job	Housewife	…	Other
Father’s Job	Teacher	…	Other
FamRelations	4	…	3
Health	3	…	5
Absences	6	…	5
Internet	No	…	Yes
Free Time	3	…	2
Playing Time	4	…	3
Extracurriculars	No	…	No
Study Time	2 - 5 hours	…	< 2 hours
Grades1	25	…	40
Grades2	30	…	45
Grades3	30	…	45

2.2. Preprocessing Data

Different types of problems affect the success of data mining on a particular problem. Two major and important issues are representation and data set quality. Knowledge discovery becomes a very difficult problem, especially if redundant and irrelevant or distracting and unreliable information is presented [9]. It is well known that machine learning tasks require significant processing time depending on the preparation of the data. In this study, the researcher used the control of the missing data values for the names of the classes such as the student records high dimensional variance deviations and unbalanced data. Therefore pre-processing techniques are used to prepare the data to improve the performance of the classification system [10].

2.3. Train-Test split

Evaluation of machine learning models with train/test splits is suitable for large datasets. The train/test split divides the dataset into a train set and a test set, or in other words, the data used for the training and testing processes are different data sets. This train/test split method will provide more accurate results for new data or data that has never been trained [11,12]. Because data testing is not used to train the model, the model does not know the outcome of the data [12]. It is called out-of-sample testing. A model is said to be good if it has high accuracy or is well for out-of-sample data because the primary purpose of making a model is to predict data correctly for which the outcome is unknown. The data in this study will be divided by 60% for training data and 40% for testing data. This data division aims to split the data into testing and training data so that the data can be evaluated more deeply. The researcher uses a 60:40 ratio to avoid overfitting which is a training condition where the test results on the trained data are excellent but tested by other data that is not used in training is very bad.

2.4. Feature Extraction

Model features are data that machine learning (ML) models use to make predictions during training and inference. The accuracy of machine learning models depends on the correct sets and combinations of features. Feature extraction and classification algorithms are an important part of classification work, which has a direct impact on the classification effect. For sample, in an application that recommends the best sellers the best items, and the others. Creating features requires extensive engineering effort [13]. Feature Extraction involves extracting and transforming variables from raw data such as price lists product details and sales volumes so that researchers can use the features for training and validations [13,14]. The steps required to design a feature include extracting and cleaning data and creating and storing features.

2.5. Univariate Feature Selection

Feature selection is an efficient data pre-processing technique in data mining to reduce data dimensions [15,16]. Identifying relevant features helps remove unnecessary and redundant attributes from the dataset and provides faster and better results. The univariate filter method is a type of method in which particular criteria are used to sort individual characteristics, and then the top N features are selected. For the univariate filter method, different types of ranking criteria are used such as fishery grades, mutual information, and feature variance. In the paper, the researcher selects the best features from the original feature total using the chi-square method.

Chi is a statistical test that measures the independence of features from class labels. It is a two-way metric. Form [17] notes that this test can act erratically when the feature count is low to predictability. With unbalanced dataset classes, this approach is relatively popular. To select the desired number of features with the best Chi-square grade, we measure the Chi-square between the target and each feature.

2.6. Classification

In Data Mining, there are various methods of classifying. Classification is an activity to find a model or function that can distinguish concepts or data classes for particular purposes [18]. Classification is the process of finding a model that explains or differentiates the concept of data, intending to be able to estimate which class the label is unknown [19]. The purpose of the classification algorithm is to predict a new class from a dataset that has a class [20,21,22]. Classification is a systematic grouping of objects, ideas, books, and other objects into particular groups or categories based on the same characteristics. In classification, there are two main jobs to be carried out, namely building a model as a prototype to be stored as memory and using it to perform recognition/classification/prediction on another data object so that it is known which group the data object belongs to in the model that has been stored.

2.7. Naïve Bayes Algorithm

The Naïve Bayes algorithm is a simple probabilistic algorithm in a classification technique that obtains its probability value based on calculating the frequency and combination values of related collections [16]. Naïve Bayes classifier is a probabilistic machine learning model that is used to classify based on Bayes' theorem [22,23]. This algorithm assumes that all attributes are independent [24]. Naïve Bayes works by predicting the probability that a data sample belongs to a particular class. The given data vector is in class C, referred to as the posterior probability, and is denoted by $P(C|F_1...F_n)$[25]. Therefore, Equation 1 is applied [25].

$$P(C|F_1 \ldots F_n) = \frac{\mathrm{P(C).P(F_1....F_n|C)}}{\mathrm{P(F_1....F_n)}} \qquad (1)$$

In Equation 1, the variable C is the class and the variable F1...Fn represents the characteristics needed to carry out the classification. Therefore, the probability of matching the data with a particular characteristic in class C (posterior) is the probability that class C appears multiplied by the sample probability of the characteristic in class C (likelihood) and then divided by the probability of the sample characteristic globally (evidence).

2.8. *Gaussian Naïve Bayes*

For features or attributes of numerical data type, the Gaussian distribution is usually chosen to represent the conditional probabilities of the continuous nature of the class, P(Xi|C), while the Gaussian distribution is characterized by two parameters: the mean, μ, and the variance, σ2. For all CJ classes, the conditional probability class is yj [26]. Equation 2 is applied as follows:

$$P(Xn = xn|Y = Yj) = \frac{1}{\sqrt{2\pi\sigma ij}} e^{-\frac{(Xi-\pi ij)}{2\sigma 2ij}} \qquad (2)$$

Explanation :
P : Probability
Xn : The nth feature
xn : Value of the nth feature
Y : The class you are looking for
μ : Mean, the average of all features.
σ : Standard deviation, stating the variance of all attributes.

2.9. Cross Validation

CV (cross-validation) is a statistical method used to evaluate the performance of a model or algorithm where the data is divided into two sections: training data and validation/evaluation [27,28]. A model or algorithm is trained by section training and validated by section validation. The choice of CV type (cross-validation) is based on the size of the data set. The CV (cross-validation) K double is usually used because the computation time can reduce the accuracy of the maintained estimate and it was developed because the previous model found a weakness.

2.10. Matrix Evaluation

This study tested model evaluation parameters including sensitivity specificity and Matthews correlation coefficient (MCC) [29]. The percentage of subjects classified correctly represents the accuracy in the training or test dataset, whereas the sensitivity represents the difference in the number of subjects classified correctly when it contains data on correctly classified student grades [30].

Table 2. *Confusion Matrix*

		Actual Value	
Predicted Value		Positive	Negative
	Positive	TP	FP
	Negative	FN	TN

Table 2 displays the ratio of correct absolute predictions (TP) to the number of positive data samples (sensitivity) and the total number of positive predictions (precision), as well as the ratio of correct negative predictions (TN) to the number of negative data specimens (specificity) and the total number of negative predictors (negative predictive value.

3. RESULT

3.1. Data Labeling Using Feature Extraction

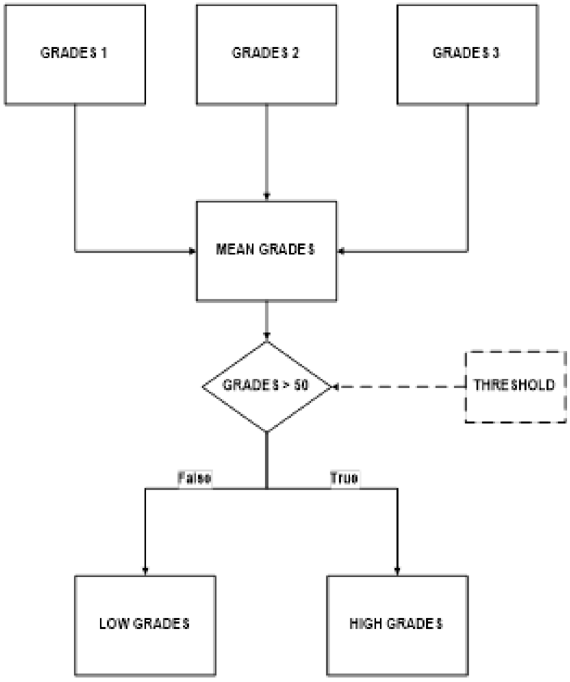

Figure 2. Labeling Process Using Feature Extraction

Data labeling is the process of adding one or more meaningful and informative labels to identify raw data and provide context for training a machine learning model. For example, a tag might indicate the presence of spoken words in a photo such as a bird or a car footage. Here the researcher uses the average value of features in Grade 1, Grade 2, and Grade 3 to produce a new one. The process can be seen in Figure 2, and the results of the averaging process (Grades Mean) can be seen in Table 3.

Table 3. Dataset With New Feature

Features	Value	….	Value
Sex	F	…	M
Age	18	…	19
Address	Urban	…	Urban
Sizefam	More than 3	…	Less than 3
StatusParent	Apart	…	Together
Mother's Edu	University	…	Elementary
Father's Edu	University	…	Elementary
Mother's Job	Housewife	…	Other
Father's Job	Teacher	…	Other
FamRelations	4	…	3
Health	3	…	5
Absences	6	…	5
Internet	No	…	Yes
Free Time	3	…	2
Playing Time	4	…	3
Extracurriculars	No	…	No
Study Time	2 - 5 hours	…	< 2 hours
Grades Mean	28	…	43

After the researcher obtained a new feature in the form of the average Grades 1, Grades 2, and Grades 3 (Grades Mean) feature values in Table 3, the researcher carried out a labeling process for these features. In the Grades Mean feature, grades below 50 (threshold) will be labeled as (Low Grades), and grades above 50 will get (High Grades). The researcher uses the scikit-learn library (LabelEncoder) for encoding labels. Code 0 is for Low Grades, while code 1 is for High Grades. This new feature will be the label for the final dataset. The result is 189 data with label 0 and 206 with label 1. This dataset is ideal because the number of labels is almost balanced. Table 4 is the final form of the dataset to be tested with the Gaussian Naïve Bayes algorithm and feature selection.

Table 4. Final Dataset

Features	Value	....	Value
Sex	F	...	M
Age	18	...	19
Address	Urban	...	Urban
Sizefam	More than 3	...	Less than 3
StatusParent	Apart	...	Together
Mother's Edu	University	...	Elementary
Father's Edu	University	...	Elementary
Mother's Job	Housewife	...	Other
Father's Job	Teacher	...	Other
FamRelations	4	...	3
Health	3	...	5
Absences	6	...	5
Internet	No	...	Yes
Free Time	3	...	2
Playing Time	4	...	3
Extracurriculars	No	...	No
Study Time	2 - 5 hours	...	< 2 hours
Label	0	...	0

3.2. Feature Selection

In this study, the cleaning and data transformation stages produced a dataset consisting of 17 input variables or features. To reduce unnecessary predictor variables, Feature Selection is used to identify inputs or predictor variables that are more substantial and thus can increase the overall model accuracy. Of the features identified as input features or predictor variables, the researcher only chose five features that were considered substantial because they had the highest score, while the others were rejected (see Table 5).

Table 5. Feature Rank

Features	Score	Rank
Sex	2.351854	5
Age	0.305375	11
Address	0.620337	9
Sizefam	1.506687	6
StatusParent	0.148160	14
Mother's Edu	5.795512	2
Father's Edu	5.380054	3
Mother's Job	1.360668	8
Father's Job	0.166120	13
FamRelations	0.187003	12
Health	0.001769	16
Absences	13.458909	1
Internet	0.357856	10
Free Time	0.007043	15
Playing Time	4.095381	4
Extracurriculars	0.000608	17
Study Time	1.389877	7

Two reasons researchers only chose five: researchers will only use features that have the most influence on the model and reduce the time used to input data in classifying.

From Table 5, five features are selected from the dataset that has the highest coefficient value according to the ranking in Table 5. Absence occupies the first position as a feature that has a substantial role in the dataset with a value of 13.458909, Mother's Education (Medu) has a second important role with a value of 5.795512, Father's Education (Fedu) has the third important role with a value of 5.380054, Playing Time has the fourth important role with a value of 4.095381, and Sex has a fourth important role with a value of 2.351854. These five features will later be modeled using the Gaussian Naïve Bayes algorithm.

3.3. Gaussian Naïve Bayes Algorithm Tuning

Researchers do this step to establish selected parameters that are used to find the optimal combination. The parameters used in this study were smoothing var. Var_smoothing is a coherence calculation that broadens (or smooths) the curve explaining more samples than the mean of the distribution. In this case np.logspace returns uniformly spaced on a logarithmic scale starting with -9 and ending with 0 giving 100 samples. Figures 3 and 4 are the best parameters of the two models:

```
▾ GaussianNB
GaussianNB(var_smoothing=0.0001)
```

Figure 3. Original Data Best Parameters

```
▾ GaussianNB
GaussianNB(var_smoothing=0.006579332246575682)
```

Figure 4. Feature Selection Data Best Parameters

3.4. Evaluation

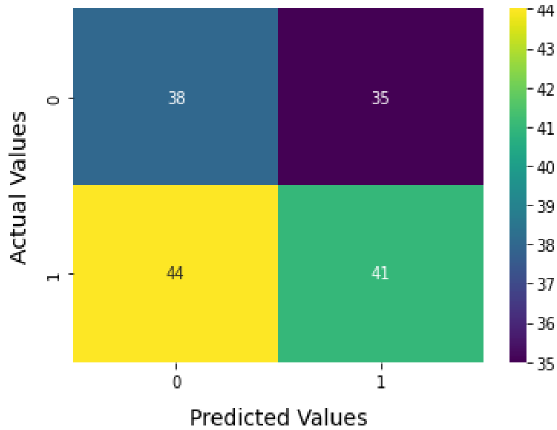

Figure 5. Confusion Matrix Original Data

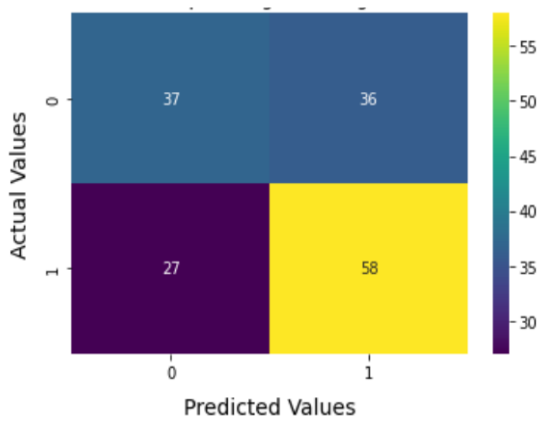

Figure 6. Confusion Matrix Selection Feature Data

Figures 5 and 6 show the difference between the two models. Then Figures 4 and 5 will be analyzed more deeply with accuracy, specificity, sensitivity, and MCC which can be seen in Table 6.

Table 6. Evaluation

Evaluation	Selection Feature	Original Data
Accuracy	60,12%	50%
Sensitivity	68,23%	48,23%
Spesificity	50,6%	52,05%
MCC	0.192	0.028

Evaluation using this metric proves that the performance in feature selection has a very significant difference. Researcher evaluates the model using accuracy, sensitivity, specificity, and MCC. The performance of models that use feature selection increases by 10% compared to those that don't. Evaluation using the Matthews Coefficient of Correlation (MCC) which is used to assess performance increased sharply from 0.028 to 0.192. It happens because the model becomes more concise so that the performance of the algorithm increases.

Researchers conducted tests for positive and negative value in ensuring the classification model. This test is carried out so that the classification carried out can work well and according to its function. The following is the calculation for the test:

1. Positive Test Value:

$$PTV = \frac{(\mathrm{TP})}{(\mathrm{TP+FP})} \tag{3}$$

Original Data:

$$PTV = \frac{41}{(41+35)} = 0{,}5394 \tag{4}$$

Feature Selection Data:

$$PTV = \frac{58}{(58+36)} = 0{,}6170 \tag{5}$$

The value above shows the probability that the classification of student scores will have characteristics according to the features if the results have been labeled 1 (High Grades). PTV of 53.94% means that if the classification results have 1 as the label, then the probability of a student's score having a high score is 53.94%. Meanwhile, in the model with the selection feature, PPV has a better value of 61.7%.

2. Negative Test Value:

$$NTV = \frac{(\mathrm{TN})}{(\mathrm{TN+FN})} \tag{6}$$

Original Data:

$$NTV = \frac{38}{(38+44)} = 0{,}4634 \tag{7}$$

Feature Selection Data:

$$NTV = \frac{(37)}{(37+27)} = 0{,}5781 \tag{8}$$

Negative test scores indicate the probability that student grades will have appropriate characteristics if the classification results are 0 (Low Grades). NTV of 46.34% means that if the result is 0, then the probability of a student's score having a low score is 46.34%. Meanwhile, in the model with the selection feature, NTV has a better value of 57.81%.

Overall, the model was better in identifying the scores of students labeled 1 or in identifying students who would have high scores at the end of the semester. If a teacher uses this model as a tool for early detection in determining students who will have high scores, then this model can fulfill its goals because it has a PTV > NTV value. In addition, the Matthews Correlation Coefficient (MCC) has a value of 0.192, which indicates a true binary classification because the MCC value is close to 1 not -1.

4. DISCUSSION

In research [31], the researcher uses an approach that enables feature selection and machine learning. The researcher proposes a classification and detection method for mesothelioma cancer. Feature selection acts as a filter, selecting just the traits which are relevant to the categorization. The accuracy of the categorization model is improved as a direct consequence of this. It has been discovered that the choice of features has a substantial influence on the accuracy of the categorization. In research [32], Feature Selection Information Gain showed better results than using only the Naïve Bayes method. Selected eight attributes used in the implementation using Naïve Bayes showed that the attribute of the source of funds had the highest weight, namely 1.00, and the feature of the parents' work received the smallest weight, namely 0.00. Research [33] provides an overview of current approaches to trait selection and classification systems for effective disease prediction. It also describes performance metrics used in medical diagnostic systems to measure the performance of classification methods. In addition, this study separately compares the strengths and weaknesses of different feature selection methods and classifiers. In addition, this paper summarizes the benefits of previous investigator work on common chronic disease datasets. This overview of feature selection methods shows that certain feature selection algorithms play a substantial role in improving accuracy in the classification of diseases. This study showed that the filter method is computationally efficient and offers better generality than other methods.

5. CONCLUSION

In this study, researchers can conclude that feature selection and feature engineering in student data classification have a significant role in improving the performance of the classification model. The proposed comparison of feature selection aims to compare the evaluation results between models with feature selection and without feature selection. It is evidenced by the improved model performance in terms of accuracy, sensitivity, and MCC compared to models that do not use feature selection. The MCC value increased almost nine times from 0.028 to 0.192, and it is good news because the model can classify better than the previous model. The selected features in this modeling are absence as the feature that ranks first, followed by the mother's education, father's education, playing time, and sex. These features were chosen because they have a higher value than other features. In future research, the application of several things can be used to improve research, namely using other algorithms to optimize parameters or other methods to reduce dataset dimensionality like PCA.

REFERENCES

[1] A. Zohair and L. Mahmoud, "Prediction of Student's performance by modelling small dataset size." *International Journal of Educational Technology in Higher Education*, vol.16, no.1, pp.1-18, 2019, DOI : https://doi.org/10.1186/s41239-019-0160-3.

[2] A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, ... and S.N. Liao, "Predicting academic performance: a systematic literature review". *In Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education*, pp. 175-199, 2018, DOI : https://doi.org/10.1145/3293881.3295783.

[3] L. Zhang, and K. F. Li, "Education analytics: Challenges and approaches". In 2018 32nd *International Conference on advanced information networking and applications workshops* (WAINA), pp. 193-198, IEEE, 2018, DOI : 10.1109/WAINA.2018.00086.

[4] J. V. Macayan, "Implementing outcome-based education (OBE) framework: Implications for assessment of students' performance". *Educational Measurement and Evaluation Review*, vol.8, no.1, pp. 1-10, 2017.

[5] M. Mahajan, and M. K. S. Singh, "Importance and benefits of learning outcomes". *IOSR Journal of Humanities and Social Science*, vol.22, no.03, 65-67. 2017, DOI : 10.9790/0837-2203056567.

[6] A. E. Tatar, and D. Düştegör, "Prediction of academic performance at undergraduate graduation: Course grades or grade point average?", *Applied sciences*, vol.10, no.14, pp. 4967, 2020, DOI : https://doi.org/10.3390/app10144967.

[7] E. T. Lau, L. Sun, and Q. Yang, "Modelling, prediction and classification of student academic performance using artificial neural networks". *SN Applied Sciences*, vol.1, no.9, pp. 1-10. 2019, DOI : https://doi.org/10.1007/s42452-019-0884-7.

[8] UCI Machine Learning Repository, "Student Performance Data Set", 2014. https://archive.ics.uci.edu/ml/datasets/student+performance [accessed Nov. 2, 2022].

[9] S. A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, "Data preprocessing in predictive data mining", *The Knowledge Engineering Review*, 34, 2019, DOI : https://doi.org/10.1017/S026988891800036X.

[10] H. Benhar, A. Idri, and J. L. Fernández-Alemán, "Data preprocessing for heart disease classification: A systematic literature review", *Computer Methods and Programs in Biomedicine*, vol.195, pp.105635, 2020, DOI : https://doi.org/10.1016/j.cmpb.2020.105635.

[11] C. Thapa, P. C. M. Arachchige, S. Camtepe, and L. Sun, "Splitfed: When federated learning meets split learning". *In Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, pp. 8485-8493, 2022, DOI : https://doi.org/10.1609/aaai.v36i8.20825.

[12] R. Y. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang, and J. P. Campbell, "Introduction to machine learning, neural networks, and deep learning". *Translational Vision Science & Technology*, vol.9, no.2, pp.14-14, 2020, DOI : https://doi.org/10.1167/tvst.9.2.14.

[13] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform", *Journal of Big Data*, vol.6, no.1, pp. 1-24,

2019, DOI : https://doi.org/10.1186/s40537-019-0191-6.

[14] R. Baboota, and H. Kaur, "Predictive analysis and modelling football results using machine learning approach for English Premier League", *International Journal of Forecasting*, vol.35, no.2, pp.741-755, 2019, DOI : https://doi.org/10.1016/j.ijforecast.2018.01.003.

[15] N. Sneha, and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection". *Journal of Big Data*, 6(1), pp. 1-19, 2019, DOI : https://doi.org/10.1186/s40537-019-0175-6.

[16] L. Ali, S. U. Khan, N. A. Golilarz, I. Yakubu, I. Qasim, A. Noor, and R. Nour, "A feature-driven decision support system for heart failure prediction based on statistical model and Gaussian naive bayes", *Computational and Mathematical Methods in Medicine*, 2019, DOI : https://doi.org/10.1155/2019/6314328.

[17] M. M. Saritas, and A. Yasar, "Performance analysis of ANN and Naive Bayes classification algorithm for data classification". *International Journal of Intelligent Systems and Applications in Engineering*, vol.7, no.2, pp. 88-91, 2019, DOI : https://doi.org/10.18201//ijisae.2019252786 .

[18] A. M. Jiménez-Carvelo, A. González-Casado, M. G. Bagur-González, and L. Cuadros-Rodríguez, "Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity–A review", *Food research international*, vol.122, pp. 25-39, 2019, DOI: https://doi.org/10.1016/j.foodres.2019.03.063.

[19] Z. A. Bukhsh, A. Saeed, I. Stipanovic, and A. G. Doree, "Predictive maintenance using tree-based classification techniques: A case of railway switches", *Transportation Research Part C: Emerging Technologies*, vol.101, pp.35-54, 2019, DOI : https://doi.org/10.1016/j.trc.2019.02.001.

[20] J. Bi, and C. Zhang, "An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme", *Knowledge-Based Systems*, vol.158, pp. 81-93, 2018, DOI : https://doi.org/10.1016/j.knosys.2018.05.037

[21] J. A. Sidey-Gibbons, and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction", *BMC medical research methodology*, vol.19, no.1, pp. 1-18, 2019, DOI : https://doi.org/10.1186/s12874-019-0681-4.

[22] A. Tharwat, "Classification assessment methods", Applied *Computing and Informatics,* vol. 17, no. 1, pp. 168-192, 2020, DOI : https://doi.org/10.1016/j.aci.2018.08.003.

[23] R. Venkatesh, C. Balasubramanian, and M. Kaliappan, "Development of big data predictive analytics model for disease prediction using machine learning technique". *Journal of medical systems*, vol.43, no.8, pp. 1-8, 2019, DOI : https://doi.org/10.1007/s10916-019-1398-y.

[24] A. E. Maxwell, T. A. Warner, and F. Fang, "Implementation of machine-learning classification in remote sensing: An applied review". *International Journal of Remote Sensing*, vol.39, no.9, pp. 2784-2817, 2018, DOI : https://doi.org/10.1080/01431161.2018.1433343.

[25] M. F. A. Saputra, T. Widiyaningtyas, and A.P. Wibawa, "Illiteracy classification using K means-Naïve Bayes algorithm", *JOIV: International Journal on Informatics Visualization*, vol.2, no.3, pp. 153-158, 2018, DOI : https://doi.org/10.30630/joiv.2.3.129.

[26] A. A. Rafique, A. Jalal, and A. Ahmed, "Scene Understanding and Recognition: Statistical Segmented Model using Geometrical Features and Gaussian Naïve Bayes", *In IEEE conference on International Conference on Applied and Engineering Mathematics*, vol. 57, 2019.

[27] C. A. Ramezan, T. A. Warner, and A. E. Maxwell, "Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification", *Remote Sensing*, vol.11, no.2, pp. 185, 2019, DOI : https://doi.org/10.3390/rs11020185.

[28] I. Tougui, A. Jilbab, and J. El Mhamdi, "Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications", *Healthcare informatics research*, vol.27, no.3, pp. 189-199, 2021, DOI : https://doi.org/10.4258/hir.2021.27.3.189.

[29] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation", *BioData mining*, vol.14, no.1, pp. 1-22, 2021, DOI : https://doi.org/10.1186/s13040-021-00244-z.

[30] G. Zeng, "On the confusion matrix in credit scoring and its analytical properties",

Communications in Statistics-Theory and Methods, vol.49, no.9, pp. 2080-2093, 2020, DOI : https://doi.org/10.1080/03610926.2019.1568485.

[31] M. Shobana, V. R. Balasraswathi, R. Radhika, Ahmed Kareem Oleiwi, Sushovan Chaudhury, Ajay S. Ladkat, Mohd Naved, and Abdul Wahab Rahmani, "Classification and Detection of Mesothelioma Cancer Using Feature Selection-Enabled Machine Learning Technique", *BioMed Research International*, vol. 2022, 2022. DOI : https://doi.org/10.1155/2022/9900668.

[32] S. Hartati, N. A. Ramdhan, and H. A. SAN, "Prediksi Kelulusan Mahasiswa Dengan Naïve Bayes Dan Feature Selection Information Gain ", *Jurnal Ilmiah Intech*, vol. 4, no. 02, pp. 223–234, 2022, DOI : https://doi.org/10.46772/intech.v4i02.889.

[33] D. Jain, and V. Singh, "Feature selection and classification systems for chronic disease prediction: A review", *Egyptian Informatics Journal*, vol.19, no.3, pp. 179-189, 2018, DOI : https://doi.org/10.1016/j.eij.2018.03.002.