Journal of Informatics and Advanced Computing (JIAC)
Vol. 5 No.1, Mei 2024
E-ISSN : 2722-0346

Use Of Support Vector Machine Algorithm To Determinate Recommendations Scholarship Recipient At Sma Negeri 8 Kota Bogor

Muhammad Rafly
Informatics Engineering / Data Science
Pancasila University[1]
rafly.muhammad72@gmail.com[1]

***Abstract*—Scholarship is an assistance program in the form of educational funding in the form of being given to assist students in obtaining proper education. SMA Negeri 8 Bogor City as an educational unit that organizes formal education, participates in providing scholarships for its students. The scholarship program at SMA Negeri 8 Bogor City has not used an effective and efficient method in determining the eligibility of students to receive scholarships so that errors often occur such as scholarships that are not on target. To overcome this, a website-based scholarship prediction application was created to predict the eligibility status of students to receive scholarships. Prediction is done using the classification method with the Support Vector Machine model. The results in the composition of 70% training data and 30% testing data get the highest evaluation results with Accuracy 89%, Precision 91%, Recall 95% .**

Keyword : Prediction, Support Vector Machine, Accuracy, Precission, Recall

1. INTRODUCTION

Education is a learning process that is a primary need that must be fulfilled, with education, a country will be able to progress and increase rapidly because education is the milestone of a nation's progress. rapidly because education is the milestone of a nation's progress. The purpose of education is to develop abilities and shape the character and civilization of a dignified nation in order to civilization in order to educate the nation's life, aiming for the development of potential of learners in order to become human beings who have faith and devotion to God Almighty, have noble character, healthy, knowledgeable, capable, and competent.[1] This is stated in the 1945 Constitution Chapter XIII on Education and Culture article 31, namely “every citizen has the right to education”.

This indicates that every citizen of school age, from primary to tertiary level, must receive an education. This indicates that every citizen of school age from primary level to tertiary level must receive an education. In reality, many citizens of school age education, especially higher education. The high cost of education is a problem faced by people who are below the poverty line, this is due to their income which does not cover their living expenses.

Therefore, economic problems are also a major factor in the problem of education in Indonesia. Scholarships are funds that can be provided by the government, private companies, embassies, universities or from workplaces to improve the quality of human resources through education[2]. Scholarships can be received by students from elementary school, junior high school, high school, and higher education levels. There are various types of scholarships available such as award scholarships given to students who have academic achievements, athletic scholarships given to students who have athletic achievements and assistance scholarships given to students who have difficulties in education costs.

To implement the above point, the government has pursued various programs, one of which is the scholarship program at various levels of education. Likewise, in high school education, the government provides scholarships through the Program Indonesia Pintar[3]. In Indonesia, education is enshrined in the Constitution where the State guarantees the right of every citizen to education. But in its implementation, there are many obstacles such as the relatively low quality of education, inadequate facilities and equal distribution of education due to economic disparities.

Therefore, the government issued a policy in this regard in the form of the Smart Indonesia Program which is expected to help children continue their education[4]. In accordance with the Regulation of the Secretary General of the Ministry of Education and Culture No. 8 of 2020 concerning PIP Implementation Guidelines, it states that this scholarship is given to students from the primary education level to the secondary education level in order to obtain educational services until they graduate from the education unit pursued by the age of the students, namely 6 years to 21 years.

SMA Negeri 8 Kota Bogor took part in the distribution of the PIP scholarship program. So far, the school has determined the scholarship recipients by manually analyzing the aspects that are taken into consideration in determining the scholarship award. Because there is no definite method in the selection process of prospective scholarship recipients, it causes target

E-ISSN : 2722-0346

errors and inaccurate targeting of scholarship distribution. Scholarships that are not on target reduce the applicant's level of trust in the scholarship provider. Muhammad Abdul Karim conducted research on student eligibility for scholarships using the Support Vector Machine and K-nearest Neighbors models. The study analyzed data such as grade point average, poverty status, orphan status, parents' income, and scholarship eligibility. The study found that both models performed best with a training data composition of 70% and testing data of 30%. The K-nearest Neighbors algorithm had an optimal K value of 3. These results show that both models are effective in determining student eligibility for scholarships. Another researcher, Apandi Sugianto, developed a classification system to predict satisfaction with e-KTP recording services using the Naïve Bayes and Decision Tree algorithms. The study found that the Naïve Bayes algorithm was more accurate, with a 91.70% accuracy rate and a 93.92% f-measure test value, compared to the Decision Tree algorithm's 65.90% accuracy rate and 79.26% f-measure test value. The Naïve Bayes algorithm also had a faster execution time of 0 seconds, while the Decision Tree algorithm took 57 seconds.

This research explores the use of Python, specifically the Support Vector Machine algorithm, to predict the admission status of students for scholarships at SMA Negeri 8 Bogor City. Python was selected due to its extensive libraries and frameworks that facilitate the development of data-centered applications, including data classification. The research resulted in a web-based application that can accurately categorize which students are eligible for scholarships at the school. This application aims to provide scholarship providers with a useful tool for making informed decisions when selecting students to receive scholarships. By using this application, decisions can be justified and scholarships can be more effectively targeted to the most deserving students.

Prediction, also known as forecasting, involves making guesses or predictions about future events[5]. In machine learning, predictions are made using trained models that can identify patterns or relationships between input and output variables. There are three types of forecasting: economic, technological, and demand forecasting. Good forecasting involves following a set of preparation steps or procedures, such as determining the purpose, selecting forecasting models, obtaining data, validating the model, making forecasts, implementing the results, and monitoring their reliability. Time series analysis, which focuses on the relationship between dependent and independent variables over time, involves methods like smoothing, Box Jenkins, and trend projection with regression. Smoothing reduces irregularities in past data, while Box Jenkins uses mathematical models for short-term forecasting. These methods are used to make accurate predictions in various industries.

Scholarships are a form of financial assistance that does not come from personal or family funding. They can be awarded for study at universities or educational institutions both domestically and internationally[6]. Scholarships usually cover educational fees for students actively attending university lectures. They can also be seen as student funds or obligations to the community. Scholarships are provided to support individuals, particularly those still in school or college, in completing their educational goals. The aim is to enable them to seek knowledge and successfully complete their education. Scholarships can be awarded by government agencies, companies, or foundations. They can be categorized as either free gifts or gifts with work obligations, known as service ties, upon completion of education. The length of the service tie varies depending on the scholarship provider.

Data Science is a multidisciplinary field that aims to derive value from data through various processes such as extraction, preparation, exploration, transformation, and presentation[7]. It involves techniques like machine learning, deep learning, computer vision, and natural language processing to analyze and classify large amounts of data. By using ensemble learning algorithms, predictions from multiple sensors can be aggregated to minimize model selection errors. The data science lifecycle follows stages like data ingestion where data is collected from various sources, including structured and unstructured data. Data storage and processing involve selecting appropriate storage systems for different types of data, as well as cleaning, transforming, and merging the data. These stages enable analysts to gain actionable insights and provide a foundation for data-driven decision making.

Support Vector Machine (SVM) is a statistical-based method used in machine learning that has a clear theoretical foundation and is not considered a black box. Implementing SVM is relatively straightforward as the process of determining the Support Vectors can be framed as a quadratic programming problem, making it easier to automate if there is a library available to solve such problems[8]. SVM has good generalization ability with small sample sizes but struggles to choose the most appropriate parameters, which can affect its performance. Finding the right parameter values, such as the soft-margin coefficient C and the kernel function parameters, is crucial for increasing the accuracy of SVM. SVM works by identifying a hyperplane with the largest margin that separates the data between classes. The Support Vectors, which are the closest data points to the hyperplane in each class, play a crucial role in the classification process.

SVM has the main principle of finding the best hyperplane that serves to separate two classes in the input space. The hyperplane can be a line in two-dimension and can be a flat plane in multiple planes.

Finding the optimal hyperplane location is the core of the SVM method. It is assumed that there is learning data with data points xi (i = 1,2,....m) has two classes y¬i = ±1, namely positive (+1) and negative (-1) classes so that the decision function will be obtained as in Equation (1).

$$f(x) = sign\,(w \cdot x + b) \quad (1)$$

With (·) being a scalar such that in Equation (2).

$$w.x = w^t\, x \quad (2)$$

Description:
w = hyperplane weight
x = input data vector
b = relative field

$$w.x_i + b = 0, for\ all\ i = 1,2,3,\dots,n \quad (3)$$

Description:
w = weight vector
x_i = input data
b = bias value
n = a lot of data

to be able to fulfill Equation (3), the values of w and b can be known by Equation (4), where two-class data can be +1 or -1.

$$y_1((w_1.x_1) + (w_2.x_2) + \dots + (w_i.x_i) + b) \geq 1 \quad (4)$$

Description:
y_i = class of the i-th data sample
w_1 = i-th weight vector
x_i = feature input data

A. *Normalization*

Normalization in the context of machine learning refers to the process of transforming feature values into a specific scale, typically between 0 and 1, without altering the data distribution. This is done to handle variations in the range of values present in the data, as a significant difference in range can impact the algorithm's classification performance. Common normalization methods include Min-Max Scaling, Standard Scaling, and Robust Scaling. Min-Max Scaling, for example, utilizes the minimum and maximum values of each attribute to transform values within a range of 0-1. This approach helps maintain value comparisons between input variables and avoids scale imbalances in the data. The formula utilized in Min-Max Scaling is (normalized x = (x - min) / (max - min)). Normalization is crucial in machine learning as it ensures that all features carry equal weight when employed in algorithms [9].

B. *K-Fold Cross Validation*

Cross Validation is a method used in the application of machine learning to evaluate the ability of the machine learning model on the data used by dividing the data into subsets. The type of Cross Validation test used in this research is K-Fold Cross Validation. K-Fold Cross Validation serves to measure the performance of the algorithm model built by dividing the data set randomly and grouping the data as much as the K value in K-Fold [10]. In the K-Fold Cross Validation method scheme, the data set is divided into several smaller sets randomly. The data subsets are processed K times with each experiment using the Kth subset as testing data and using the rest of the subset as training data.

C. *Confusion Matrix*

Confusion Matrix is a measurement method in the form of a table used to evaluate the performance of the machine learning model created by comparing the model's prediction results with the actual values [11]. There are four components in confusion matrix, including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positive (TP) is the amount of data that is predicted to be positive with the actual condition also being positive. True Negative (TN) is the amount of data that is predicted to be negative with the actual condition also being negative. False Positive (FP) is the amount of data that is predicted positive but with negative actual conditions and False Negative (FN) is the amount of data that is predicted negative but with positive actual conditions. The results of the confusion matrix are used in the calculation of three accuracy metrics, including accuracy, precision, and recall. These three matrices are then used to measure the performance of the machine learning model built. The confusion matrix table is depicted in Table 1.

Table 1 Confusion matrix

		Aktual	
		Positif	Negatif
Prediksi	Positif	TP	FP
	Negatif	FN	TN

1) Accuracy

Accuracy is a measurement of the closeness between the predicted value and the actual value. The accuracy value is known by calculating the amount of data that is predicted correctly from all model predictions. Equation 5 is an equation to determine the accuracy value.

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

2) Precision

Precision is the result of the comparison between the amount of data predicted True Positive with the total amount of data predicted positive. Equation 6 is an equation to determine the precision value.

$$Precision = \frac{TP}{FP + TP}$$

3) Recall

Head and shoulders shots of authors which appear at the end of our papers.

$$Recall = \frac{TP}{TP + FM}$$

E-ISSN : 2722-0346

2. METHODOLOGY

This research follows a structured workflow that includes a number of organized stages and methods. Figure 1 is the flow of scholarship prediction classification research using Support Vector Machine.

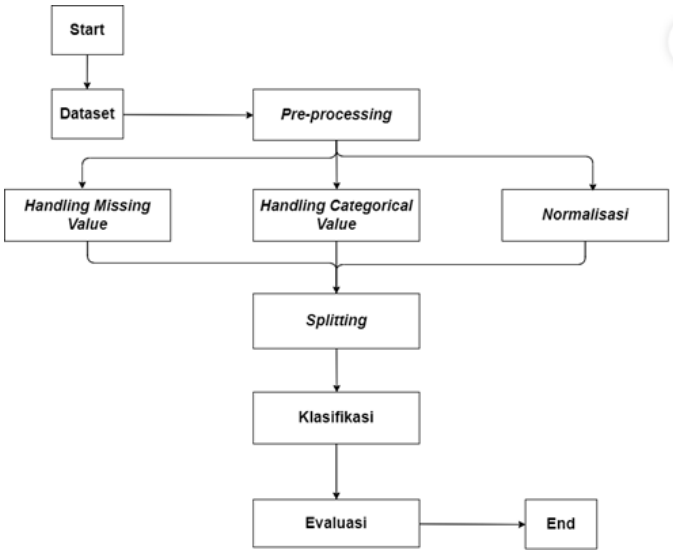

Figure 1 Research Flow

A. Dataset

A data set is a structured collection of data that can include various types of information, such as numbers, text, and images. These data sets are collected by professionals from different sources and are processed through stages like data cleaning and categorization. Data sets are categorized based on their variables and are used in various industries and disciplines. In the context of research on predicting scholarship acceptance status, the data collection stage is crucial. It involves collecting data with the goal of training a system to make accurate predictions about scholarship acceptance. This stage requires both quantity and quality of data, with a focus on gathering data that specifically relates to scholarship prediction. By ensuring a thorough and accurate data collection stage, this research can establish a strong foundation for developing a successful scholarship prediction model.

B. Pre-processing

Data pre-processing is an important step in data analysis that involves organizing data in a consistent format and managing missing values and categorical variables. The process includes handling missing values, which involves dealing with data that is absent or has errors or failures. Proper handling of missing values is crucial for reliable data analysis. Categorical variables, which group data with similar characteristics, need to be transformed into numeric values to be processed by machine learning algorithms. Normalization involves reorganizing a database by removing duplicate data, creating a standard format for writing, and establishing relevant correlations among data in tables. Data normalization simplifies analysis, decision-making, and data modification. Overall, data pre-processing is essential for preparing data and ensuring its integrity and reliability in further analysis.

C. Splitting

Data splitting is a crucial technique in data science for developing accurate and usable data models. It involves dividing the data into subsets, typically for training and testing purposes. The training dataset is used to build and train the model, allowing the machine learning algorithm to adjust its parameters based on the provided data. The model's performance is then evaluated using the validation dataset to assess its ability to recognize patterns. On the other hand, the testing dataset is used to assess the model's performance after the training process is completed. A high-performing machine learning algorithm should be able to correctly categorize new data that was not part of the training process, demonstrating its ability to generalize. The recognition rate of an algorithm is typically higher when it has a higher level of generalization.

D. Classification

Data classification is the process of organizing data into relevant categories to make it more efficient and retrievable. It plays a crucial role in risk management, compliance, and data security. This involves tagging data for easier search and tracking. Data classification serves a variety of purposes such as maintaining regulatory compliance, facilitating access, and meeting business or personal needs. It helps organizations manipulate, track, and analyze data systematically. While data classification is commonly used for structured data, it is particularly important for unstructured data, which lacks clear labels. By classifying unstructured data, it becomes more useful and easier to search or query.

E. Evaluation

Evaluation is a descriptive, informative and predictive data process. Evaluation is also a series of activities or activities that aim to measure the level of success in a program. Evaluation aims to provide input for program planning, provide input to modify the program and obtain information about the supporting and inhibiting factors of the program.

The information gathered from the evaluation process can improve the level of performance of ongoing activities get interruptions that occur from the beginning to the evaluation, and realize what to do in the future to avoid problems to remain productive.

3. RESULTS AND DISCUSSION

The data used is data on students who apply for scholarships at SMA Negeri 8 Bogor City in the academic year 22/23 classes XI IPA 1-6 and XI IPS 1-3. The data obtained consists of 314 (three hundred and fourteen) student data with 11 (eleven) attributes or features with 8 (eight) attributes of string type and 3 (three) attributes of numeric type. These attributes are student name, year of enrollment, average score per semester, orphan information, means of transportation, parental dependents, distance, type of house, father's salary, mother's salary and admission status. dimensionally. If you must use mixed units, clearly state the units for each quantity

in an equation.

Of the 11 existing attributes, not all can be used in the classification process. There are 9 attributes that are used in the classification process, namely Average score per semester, Orphan's Statement, Means of Transportation, Distance, Father's salary, Mother's salary, Type of House, Parental Dependents, Admission Status that determine the eligibility of prospective students to receive scholarships as labels. Meanwhile, 2 other attributes namely student name and year of enrollment are not used.

A. Research Stages

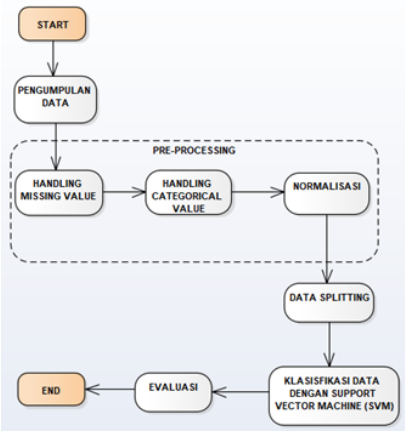

Figure 2 Research Stages

This Figure 2 describes the stages of a research project. The research consists of seven stages, starting with data collection. The data collection process involves observing students from SMA Negeri 8 Bogor who applied for scholarships in the academic year 22/23. The collected data includes information about the students in classes XI IPA 1-6 and XI IPS 1-3, totaling 314 students. After data collection, the researchers proceed with data pre-processing, which involves three processes: handling missing values, handling categorical values, and normalization. Handling missing values is important because most machine learning models have issues with null data. This can be addressed through deletion or imputation. Handling categorical values involves converting string categorical data into numerical form using techniques like One Hot Encoding and Ordinal Encoding. The attributes in the study are determined based on factors such as average grades per semester and orphan status, which have been identified as important criteria for scholarship acceptance.

B. Scenario of Application of the Methods Used

The text describes a simulation conducted using Microsoft Excel to understand the stages of the data processing process. The simulation involved applying a scenario to a dataset of scholarship applicant student data, with a total of 314 rows. The stages in the scenario include data collection and data pre-processing. In data pre-processing, missing values are handled by checking for null values, and since none are found, the process proceeds. Categorical values are then handled by converting them to numerical form. The orphan information column is converted to 1 for "Yes" and 0 for "No". The transportation equipment column is converted to numerical values based on different categories of transportation methods. The parental dependents column is converted to 1 for "With parents" and "Orphanage", and 0 for "Boarding school", "Guardian", and "Dormitory". The distance column already has a numerical data type, so it is entered accordingly. The house type column has a nested string value.

C. Evaluation Technique

To determine the performance of the *Support Vector Machine* algorithm in the classification of students who are eligible to receive scholarships, an evaluation is carried out using *K-Fold Cross Validation* with a value of K = 10. In this process, the data is divided into 10 subsets. Then 10 trials were conducted, where in each trial, the Kth subset was used as testing data and the other subset was used as training data. *K-Fold Cross Validation* is implemented using the *cross_val_score* module of *sklearn*. The following are the results of the *K-Fold Cross Validation* performed on the algorithm.

From the experiments carried out by applying *K-Fold Cross Validation*, the results obtained for the *Support Vector Machine* (SVM) algorithm have the highest permutation at K = 5, K = 6 and K = 8 as described in Table 2

Table 2 K-Fold Cross Validation

Fold	Skor *Cross Validation*
1	0.90625
2	0.875
3	0.75
4	0.83870968
5	0.93548387
6	0.93548387
7	0.90322581
8	0.93548387
9	0.87096774
10	0.87096774

From the classification results with *Support Vector Machine* (SVM) obtained as in table 2, the confusion matrix is obtained in table 3.

Table 3 Confusion Matrix

		Actual	
		Negative	Positive
Prediction	Negative	65	7
	Positive	12	42

Based on the *Confusion Matrix*, the *accuracy*, *precision*, and *recall* values are as follows.

1) Accuracy is calculated based on Equation 5

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN} = \frac{65 + 42}{65 + 7 + 42 + 12} = 0.84$$

2)

3) Precision is calculated based on Equation 6

$$Precision = \frac{TP}{FP + TP} = \frac{42}{7 + 42} = 0.85$$

4) Recall is calculated based on Equation 7

$$Recall = \frac{TP}{TP + FN} = \frac{42}{42 + 12} = 0.77$$

For comparison of the Accuracy, Precision and Recall values obtained based on the composition of different training data and testing data for predictions made using the Support Vector Machine algorithm. From table 5.5, it is known that for a total of 314 (three hundred and fourteen) student data, the composition of training data and testing data that produces the best value in each Accuracy metric is obtained in the composition of 70% training data and 30% testing data. This composition produces an Accuracy value of 89.36%, Precision of 91% and Recall of 95% as in the following Table 4

Table 4 Model Evaluation with Different Data Composition

Komposisi dalam persen (%)		Metrik Akurasi		
Data Training	*Data Testing*	Akurasi	Presisi	*Recall*
90	10	81.25%	0.85%	0.88%
80	20	85.71%	0.91%	0.94%
70	30	89.36%	0.91%	0.95%
60	40	88.89%	0.92%	0.89%
Rata-rata		86.30%	0.89%	0.91%

D. Implementation

After doing the implementation that has been done from data pre-processing, data splitting, classification, evaluation, and prediction of new data. To make predictions with new data by entering into a form in the form of a web-based application. The implementation consists of input implementation and output implementation. The input consists of entering the scholarship applicant data into the form of the scholarship prediction process. Figure 3 and Figure 4 are the inputs of the prediction process.

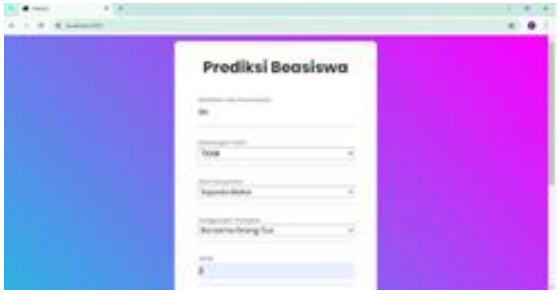

Figure 3. Implementasi data *input*

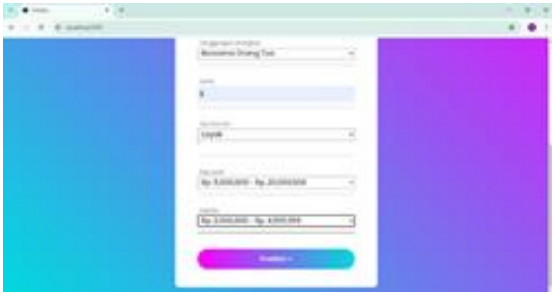

Figure 4. Implementasi data *input*

Implementasi *output* terdiri dari output proses prediksi beasiswa. Gambar 5 merupakan *output* dari proses prediksi 5 merupakan *output* dari prediksi beasiswa.

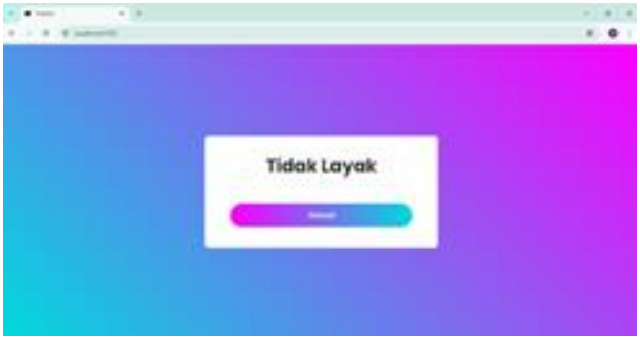

Figure 5. Implementasi *output*.

4. CONCLUSION

Based on the results obtained in research on the comparison of Support Vector Machine algorithm in predicting student eligibility to receive scholarships at SMA Negeri 8 Bogor City, it can be concluded that Regarding the economic attributes that are currently running, with the existence of web-based machine learning with Support Vector Machine modeling, all attributes including average score per semester, orphanage information, means of transportation, parental dependents, distance, type of house, father's salary and mother's salary can make SMA Negeri 8 Bogor City more accurate and on target in recommending the eligibility status of students receiving scholarships. Based on the results and evaluation tests above, the algorithm performance value of Support Vector Machine modeling with the highest value in the division of training data as much as 70% and 30% testing data with an Accuracy value of 89%, precision 91%, Recall 95% and K-Fold Cross Validation with the highest value at K = 5, K = 6 and K = 8 produces a value of 0.93548387. With this evaluation value, it can help SMA Negeri 8 Bogor City in determining the eligibility of students to receive scholarships.

REFERENCES

[1] A. Razak and S. Wiguna, "The Effect of the Indonesia Pintar Program (PIP) Scholarship on the Learning Interest in Aqidah Akhlak of Grade VIII Students at MTS Alwashliyah, Babalan District," J. Keguruan Ilmu Pendidik., vol. 1, p. 249, 2022.
[2] T. Hidayatulloh, S. Suhada, E. Nursyifa, and L. Yusuf, "Decision Making for High School Scholarship Recipients Using Fuzzy Multiple Attribute Decision Making Model Weighted Product," J. Pilar Nusa Mandiri, vol. 14, p. 247, 2018.
[3] R. D. Gunawan, T. Oktavia, and R. I. Borman, "Design of an Online-Based Scholarship Information System for the Indonesia Pintar Program (PIP) (Case Study: SMA N 1 Kota Bumi)," J. Mikrotik, vol. 8, p. 43, 2018.
[4] N. Mutiah and R. P. Sari, "Determination of PIP Scholarship Acceptance Using the MOORA Method at SD Negeri 11 Sandai," J. Khatulistiwa Inform., vol. 11, p. 43, 2023.
[5] F. D. Ramadhani and M. Ardiansyah, "Definition of Prediction," in Sales Prediction System Using Single Exponential Smoothing and Parabolic Trend Method, South Tangerang, Indonesia: Pascal Books, PT. Mediatama Digital Cendekia, 2021, p. 5.
[6] A. Gafur, S. Yulianti, and N. Hidayat, "Easy Ways to Get Scholarships," in Beasiswa, Jakarta, Indonesia: Penebar Plus, 2008.
[7] M. S. Haris, Q. Nurlaila, Ratnadewi, F. John, *Introduction to Data Science*, Makassar, Indonesia: CV. Tohar Media, 2019.
[8] S. Yahya, *Data Mining*, Sukabumi, Indonesia: CV Jejak, 2022.
[9] Rumah Coding, "Normalisasi dan Standarisasi Data dengan Scikit-Learn," *Rumah Coding*, Jun. 14, 2024. [Online]. Available: https://rumahcoding.id/blog/normalisasi-dan-standarisasi-data-dengan-scikit-learn/. [Accessed: Sep. 5, 2024].
[10] R. R. R. Arisandi, B. Warsito, and A. R. Hakim, "Application of Naive Bayes (NBC) for Classification of Stunting Nutrition Status in Toddlers with K-Fold Cross Validation Testing," J. Gaussian, vol. 11, 2022.
[11] A. H. Al Kabir, S. Basuki, and G. W. Wicaksono, "Sentiment Analysis of Feedback Data from Information Technology Training Using Support Vector Machine Algorithm," *Repositor*, 2019.

Muhammad Rafly[1] (24 November 2000) born in Bogor and he is studying at Pancasila University of informatics engineering from 2019 to 2024 where he takes a focus of interest on data science. He graduated from Pancasila University with Engineering Degree (S.T) from informatics engineering department. and he works under a CV. Terang Megah as IT support (2022).

He works under CV.Terang Megah as IT support in inter-school teaching aids located in Bantarkalong, Kec. Warung Kiara, Kab. Sukabumi, Prov. West Java. he has been a practicum assistant in algorithm and programming courses with C++, web-based programming with laravel, laragon and Golang, and finally mobile-based programming with Flutter. He has been a practicum assistant in algorithm and programming courses with C++, web-based programming with laravel, laragon and Golang, and finally mobile-based programming with Flutter.