Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 5 No. 2 2025, 106 - 113
DOI: *https://doi.org/10.22487/27765660.2025.v5.i2.17882*
eISSN: 2776-5660

THE CONTRIBUTION OF VOCATIONAL EDUCATION TO PREDICTING YOUTH UNEMPLOYMENT IN SOUTHEAST SULAWESI: A MACHINE LEARNING APPROACH

Fais Jefli

Politeknik Statistika STIS

ORCID ID: https://orcid.org/0009-0005-0403-1570

***e-mail**: *212313072@stis.ac.id*

ABSTRACT

Youth unemployment remains a major issue in Indonesia, including in Southeast Sulawesi Province. Although the overall open unemployment rate in this province is relatively low, the unemployment rate among young people is still quite high. One contributing factor is the mismatch between educational outcomes and labor market needs, especially for those entering the workforce for the first time. In this context, vocational education is expected to enhance youth employability. Therefore, this study aims to classify youth employment status and identify the predictor that contribute most to the prediction results, particularly vocational education, using SHapley Additive exPlanations (SHAP) values to interpret model decisions. Several machine learning classification methods were evaluated, including naïve Bayes and random forest, with logistic regression used as the baseline comparison model. The findings indicate that the random forest model provides the best classification performance. Based on the analysis, vocational education and age group are the most influential predictors in classifying youth employment status in Southeast Sulawesi Province. Thus, vocational education serves as a key predictor that enhances the model's ability to classify employment status and is associated with a higher model-predicted probability of being employed.

Keywords: *Youth Unemployment; Vocational Education; Machine Learning; Random Forest*

Cited: Jefli, F. (2025). The Contribution of Vocational Education to Predicting Youth Unemployment in Southeast Sulawesi: A Machine Learning Approach. *Parameter: Journal of Statistics*, 5(2), 106–113. https://doi.org/10.22487/27765660.2025.v5.i2.17882

*Received: Oct 2025; Reviewed: Nov 2025; Published: Dec 2025

INTRODUCTION

Employment is one of the main pillars that directly affects the economy and the welfare of society, especially in developing countries like Indonesia. One of the key indicators for measuring employment is the unemployment rate. Unemployment refers to a condition in which individuals within the labor force are unable to secure employment opportunities (World Bank, 2025). Elevated unemployment levels can obstruct a nation's developmental progress and, in turn, create adverse consequences for both economic stability and broader social well-being (Mitsi, 2023). In Indonesia, the youth group (aged 15–24 years) constitutes the largest contributor to the open unemployment rate, indicating that this productive age group is, in fact, the group most vulnerable to unemployment (BPS, 2024a). A similar situation occurs in several Indonesian provinces, including Southeast Sulawesi. Although Southeast Sulawesi has the third-lowest open unemployment rate (TPT) in Sulawesi Island at 3.09%, the TPT among young people in this province remains relatively high.

According to the August 2024 Sakernas data, the youth group (aged 15–24 years) contributed the most to the open unemployment rate in Southeast Sulawesi. This contrasts with other age groups, whose contributions were relatively smaller. This situation is a serious concern, as young people represent a valuable asset for Indonesia's future economic growth and development. However, in reality, this age group has instead become an economic burden.

The *link and match* issue is one of the main causes of unemployment. This concept refers to the alignment between skills and labor market needs (Ardhana et al., 2025). Although job vacancies (demand) exist, job seekers (supply) often lack the skills, education, or experience required by employers. This mismatch is known as the *gap* between the demand and supply sides of the labor market (Adriyanto et al., 2020). Such problems frequently occur among young people, especially recent graduates from schools or universities who fail to enter the workforce because their competencies do not meet industry requirements. This situation arises from several factors, including less adaptive curricula, insufficient practical training, and weak collaboration between educational institutions and industries (Pramesti et al., 2024).

Vocational education emerges as a solution to youth unemployment. According to Presidential Regulation No. 68/2022, vocational education is an educational system designed to prepare graduates to work or become entrepreneurs according to their expertise, by developing skilled, efficient, and competitive human resources. Research by Yoana et al. (2024) shows that vocational education reduces individuals' likelihood of becoming unemployed, as it helps bridge the gap between the education system and the labor market (Sari et al., 2024). However, a study by Ohara et al. (2020) found that the unemployment rate among vocational school graduates (secondary vocational education) is higher than that among general high school graduates.

Based on the previous explanation, there remains a knowledge gap regarding the role of vocational education in reducing youth unemployment. Therefore, this study aims to address this gap by employing machine learning methods that are capable of generating more accurate predictions. In addition, this research introduces the use of SHapley Additive exPlanations (SHAP) Values, which provide clearer insights into the directional contribution of each variable in the prediction, thereby enhancing the interpretability of machine learning models that are generally considered "black box", although not intended to indicate causal relationships. The objective of this study is to analyze how vocational education is associated with other variables in predicting the likelihood of youth employment, supported by SHAP Values analysis. The findings of this study are expected to provide insights for policymakers regarding the extent to which vocational education is associated with the prediction of youth unemployment levels.

MATERIALS AND METHODS

Data and Research Scope

This study utilizes secondary data in the form of raw data obtained from the Central Bureau of Statistics (BPS). The data used are sourced from the National Labor Force Survey (Sakernas) conducted in August 2024. The research focuses on Southeast Sulawesi Province, analyzing a total of 2371 individuals.

Operational Definition

This study predicts individuals' unemployment status by using vocational education as the main predictor variable, along with other supporting variables. A detailed explanation of the variables used in this study is presented in the following table.

Table 1. Operational Definition and Formation of Response and Predictor Variables

Variable	Notation	Criteria	Reference
Response Variable (Y)			
Youth Unemployment	Unemployment	"Individuals aged 15–24 years who are actively seeking employment, including those preparing a new business/job, those who are not looking for work because they believe no job is available, and those not actively seeking work because they already have a job but have not yet started working."	BPS (2024b)
Predictor Variable (X)			
Education Level	Vocational	1: Vocational education 0: Non-vocational education	Maulana & Suryaningrum (2023)
Age	Age	1: 15–19 years 0: 20–24 years	Febryanna (2022)
Area of Residence	Rural	1: Rural area 0: Urban area	Yanindah (2021)
Gender	Female	1: Female 0: Male	Suhaeri (2021)
Marital Status	Single	1: Never married 0: Ever married	Alharis & Yuniasih (2022)
Certified Training	Training	1: Attended 0: Never attended	Alharis & Yuniasih (2022)

Classification Method

Classification is the process of predicting a categorical response variable based on several predictor variables, with the aim of assigning each observation to one of the defined categories. Before performing classification, the dataset is divided into two groups: training data and testing data. The training data are used to build the classification model, while the testing data are used to evaluate the model's performance. In this study, 70% of the data are used as training data and 30% as testing data.

In this research, several machine learning methods are compared to determine the best model for predicting an event. The methods used include Naive Bayes and Random Forest (James et al., 2021). As a baseline comparison, a binary logistic regression model is also constructed. This model is used to predict binary outcomes (Yes/No). The logit function used is expressed as follows:

$$p(x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} \quad (1)$$

Equation (1) is adjusted using the log-odds link function as follows:

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad (2)$$

Naive Bayes Classifier is a method used to predict events based on Bayes' theorem. This model makes a simplifying assumption that, within a given category, all predictor variables are mutually independent hence the term "naive". Under the Naive Bayes assumption, the posterior probability formula is expressed as:

$$\Pr(Y = k \mid X = x) = \frac{\pi_k \cdot f_{k1}(x_1) \cdot f_{k2}(x_2) \cdots f_{kp}(x_p)}{\sum_{l=1}^{K} \pi_l \cdot f_{l1}(x_1) \cdot f_{l2}(x_2) \cdots f_{lp}(x_p)} \quad (3)$$

For k = 1, …, K

Random Forest is a model based on decision trees, which partition the predictor space using a set of rules that can be visualized as a tree structure. The Random Forest method constructs multiple decision trees from bootstrap samples of the training data, where each split considers only a random subset of $m$predictors out of the total p predictors. Commonly, $m \approx \sqrt{p}$ (the square root of the total number of predictors).

SMOTE

Resampling techniques are methods used to balance the number of observations between majority and minority categories. This process is crucial because most machine learning algorithms tend to struggle in accurately classifying minority categories. One such method is the Synthetic Minority

Oversampling Technique (SMOTE). SMOTE works by increasing the number of observations in the minority class to match the majority class by generating synthetic samples based on their nearest neighbors (Ihfa & Harsanti, 2021).

SHAP Values

SHAP (SHapley Additive exPlanations) values represent the magnitude and direction of each predictor variable's contribution to the model's prediction. A positive SHAP value indicates that the predictor variable increases the likelihood of an observation being classified as "success," while a negative SHAP value indicates the opposite effect (Rodríguez-Pérez & Bajorath, 2020).

The modeling stages in this study are as follows.

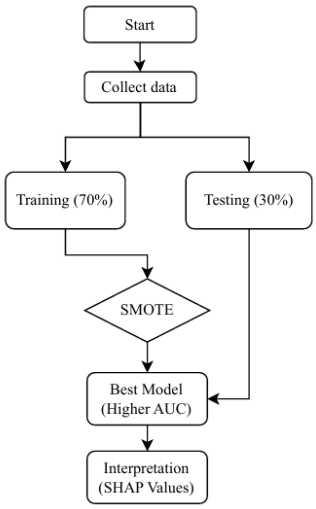
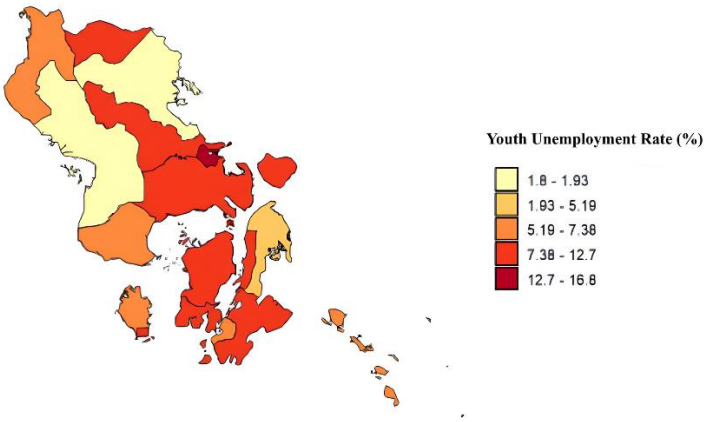


Figure 1. Modeling Flowchart

The modeling process begins by splitting the data into training and testing sets. If the training data exhibits class imbalance, a balancing procedure is performed using SMOTE. The next step is selecting the best model based on the highest Area Under the Curve (AUC) value in both the training and testing data. The final stage is interpreting the best model using SHAP values.

RESULTS AND DISCUSSION

Overview of by Regency/City in Southeast Sulawesi



Source: Sakernas August 2024, processed

Figure 2. Percentage of Youth Unemployment by Regency/City in Southeast Sulawesi

Based on Figure 2, there are clear variations in youth unemployment rates across regencies and cities in Southeast Sulawesi. The youth unemployment rate ranges from 1.8% to 16.83%, indicating that unemployment among young people remains relatively high in several areas. The pattern shows that regions with higher unemployment rates tend to be concentrated in the southwestern part of the province.

Model Development

Table 2. Number of Training and Testing Data

Category	Youth Unemployment Status	
	Training Data (70%)	Testing Data (30%)
Unemployed	118	54
Not Unemployed	1541	658

The training data were used to build the models, namely Logistic Regression, Naive Bayes, and Random Forest. Since the number of observations in the unemployment category was relatively small, data balancing was performed using the SMOTE method on the minority class, i.e., unemployed youth. The number of training data after applying the SMOTE method is presented below.

Table 3. Number of Training Data and SMOTE Data

Category	Youth Unemployment Status	
	SMOTE Data	Testing Data
Unemployed	1540	54
Not Unemployed	1541	658

From Table 3, it can be seen that the number of observations between the two groups became balanced after applying the SMOTE process. Subsequently, the performance of each model was evaluated based on the highest AUC value from both the SMOTE data and the testing data to identify the best-performing model.

Best Model

Table 4. Comparison of Machine Learning Model Performance

Model	Area Under Curve (AUC)	
	SMOTE Data	Testing Data
Logistic Regression	0,678	0,558
Naive Bayes	0,676	0,546
Random Forest	0,704	0,606

Based on Table 4, the Random Forest model achieved the highest AUC values for both the SMOTE data and the testing data compared to Logistic Regression and Naive Bayes. Therefore, the Random Forest model was selected for further interpretation and analysis.

Model Interpretation

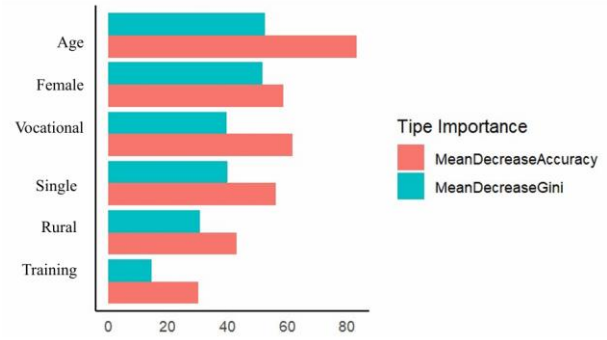


Figure 3. Variable Importance Based on Gini and Accuracy

Based on Figure 3, it can be seen that age, education level (vocational), and gender (female) are the most influential variables in determining the model's predictions. A higher importance value indicates a greater contribution to the model's performance. Subsequently, the contribution and direction of these key variables were analyzed using SHAP values.

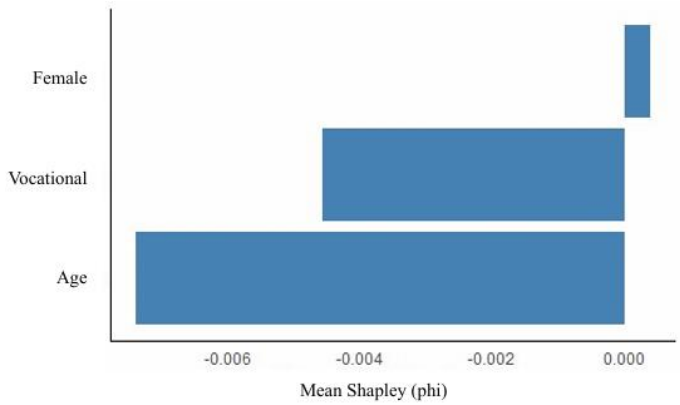


Figure 4. Average SHAP Values for Main Predictor Variables

According to Figure 4, the SHAP values show that individuals aged 15–19 years and those with vocational education tend to have higher probabilities of being classified as not unemployed. Meanwhile, the predictive strength of the female gender variable is relatively weak.

Based on Figure 3, negative SHAP values indicate that vocational education is associated with a higher likelihood of young individuals being classified as employed. This finding aligns with Yoana et al. (2024), who found that individuals with a vocational education background are less likely to be unemployed because they possess greater competitiveness in the labor market. For example, vocational high schools (SMK) equip students with practical skills aligned with job requirements and provide work experience through internship programs. Similarly, Pramesti et al. (2024) revealed that a gap still exists between the knowledge taught in conventional universities and the actual needs of the labor market, as reflected by the high rate of educated unemployment largely due to an emphasis on cognitive skills rather than life skills.

As a result, vocational education serves to supply job-ready workers for companies. Employers tend to prefer vocational graduates because they do not require significant time or resources for additional training (Subiyantoro et al., 2023). Thus, vocational education helps reduce the link and match gap between educational institutions and labor market needs, which is one of the main factors contributing to unemployment.

Based on Figure 3, the negative SHAP value also indicates that being in the 15–19 year age group is associated with a higher likelihood of being classified as employed. This finding supports Alam (2016), who reported that individuals aged 15–19 most of whom are high school graduates tend to have lower wage expectations and are therefore more willing to accept low-paying jobs. In contrast, university graduates are typically more selective regarding salary levels and job types, making them more prone to unemployment.

CONCLUSION

Based on the research findings, it can be concluded that the youth unemployment rate in Southeast Sulawesi Province is significantly higher than that of other age groups, with higher unemployment levels concentrated in the southwestern regencies and cities of the province. One of the main causes of high youth unemployment is the weak link and match between the education system and labor market needs. Therefore, this study compares several machine learning models to predict the unemployment status of young individuals in Southeast Sulawesi. The analysis results indicate that the Random Forest model performs best, achieving the highest AUC value compared to other models. Based on the variable importance results, age, education level, and gender are identified as the most influential predictors in determining the unemployment status of young individuals. Furthermore, the SHAP value analysis shows that vocational education and being in the 15–19 age group (compared to 20–24 years old) are associated with a higher model-predicted probability of being employed. These findings suggest that vocational education may serve as an important predictor in improving the prediction of youth employment outcomes and reflects its potential role in aligning education with labor market needs in Southeast Sulawesi Province.

REFERENCES

Ardhana, A. Y. A., Syazeedah, H. N. U., Fitriyaningrum, R. I., & Gunawan, A. (2025). Analisis ketidaksesuaian antara pendidikan dengan kebutuhan dunia kerja di Indonesia. *Kompeten: Jurnal Ilmiah Ekonomi dan Bisnis, 3*(4), 1020–1026.

Adriyanto, A., Prasetyo, D., & Khodijah, R. (2020). Angkatan Kerja dan Faktor yang Mempengaruhi Pengangguran. *Jurnal Ilmu Ekonomi & Sosial Unmus*, *11*(2). https://doi.org/10.35724/jies.v11i2.2965

Alam, S. (2016). Tingkat Pendidikan dan Pengangguran di Indonesia (Telaah Serapan Tenaga Kerja SMA/SMK dan Sarjana). *Jurnal Ilmiah Bongaya*, *1*(1), 250–257. https://ojs.stiem-bongaya.ac.id/JIB/article/view/19

Alharis, F., & Yuniasih, A. (2022). Determinan Pengangguran Usia Muda Terdidik di Provinsi Banten Tahun 2020. *Seminar Nasional Official Statistics*, *2022*, 53–62. https://doi.org/10.34123/semnasoffstat.v2022i1.1153

Badan Pusat Statistik. (2024a). *Tingkat pengangguran terbuka berdasarkan kelompok umur*. https://www.bps.go.id/id/statistics-table/2/MTE4MCMy/tingkat-pengangguran-terbuka-berdasarkan-kelompok-umur.html

Badan Pusat Statistik. (2024b). Statistik Indonesia 2024.

Febryanna, S. (2022). Pola Karakteristik NEET (Not In Employment, Education, Or Training) Dan Pengaruh Pengetahuan Pemuda Tentang Program Kartu Prakerja Terhadap Status NEET Di Masa Pandemi. *Seminar Nasional Official Statistics*, *2022*(1), 11–20. https://doi.org/10.34123/semnasoffstat.v2022i1.1113

Ihfa, R., & Harsanti, T. (2021). Komparasi Teknik Resampling pada Pemodelan Regresi Logistik Biner. *Seminar Nasional Official Statistics*, *2020*(1). https://doi.org/10.34123/semnasoffstat.v2020i1.540

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2nd ed.). Springer New York, NY. https://doi.org/https://doi.org/10.1007/978-1-0716-1418-1

Maulana, A., & Suryaningrum, N. (2023). Pendidikan vokasi, pelatihan dan pengangguran usia muda di Indonesia pada masa pandemi Covid-19. *Jurnal Kependudukan Indonesia*, *18*(1), 93–108. https://doi.org/10.55981/jki.2023.1697

Mitsi, D. (2023). Unemployment and Economic Growth: An In-depth Analysis. *International Journal of Science and Management Studies*. https://doi.org/10.51386/25815946/ijsms-v6i4p115

Ohara, E., Harto, S. P., & Maruanaya, R. F. (2020). Policy Shift to Reduce Unemployment of Vocational School Graduates in Indonesia (A National Study). *Jurnal Pendidikan Teknologi Dan Kejuruan*, *26*(2), 129–139. https://doi.org/10.21831/jptk.v26i2.33144

Pemerintah Republik Indonesia. (2022). *Peraturan Presiden Republik Indonesia Nomor 68 Tahun 2022 tentang Revitalisasi Pendidikan Vokasi dan Pelatihan Vokasi*. Lembaran Negara Republik Indonesia Tahun 2022 Nomor 108.

Pramesti, K. D., Meisya, N. I., & Amrillah, R. (2024). Relevansi Lulusan Perguruan Tinggi dengan Dunia Kerja. *An Najah (Jurnal Pendidikann Islam Dan Sosial Keagamaan)*, *3*(4), 236–243. https://journal.nabest.id/index.php/annajah

Rodríguez-Pérez, R., & Bajorath, J. (2020). Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design*, *34*. https://doi.org/10.1007/s10822-020-00314-0

Sari, R., Kharis Al Basyar, A., Rahman, A., & Wardoyo, S. (2024). Edukatif: Jurnal Ilmu Pendidikan Peran Pendidikan Vokasi dalam Meningkatkan Keterampilan Kerja di Era Industri 4.0. *Jurnal Ilmu Pendidikan*, *6*. https://doi.org/10.31004/edukatif.v6i6.7849

Subiyantoro, H., Tarziraf, A., & Asmara, A. Q. (2023, June). The Role of Vocational Education as the Key to Economic Development in Indonesia. *Proceedings of the 3rd Multidisciplinary International Conference*. https://doi.org/10.4108/eai.28-10-2023.2341745

Yanindah, A. (2021). An insight into Youth Unemployment in Indonesia. *Proceedings of The International Conference on Data Science and Official Statistics, 2021*(1), 666–682. https://doi.org/10.34123/icdsos.v2021i1.229

Yoana, Ilmiawan, A., & Rumayya, and. (2024). The role of vocational education on unemployment in Indonesia. *Cogent Education*, *11*(1), 2340858. https://doi.org/10.1080/2331186X.2024.2340858

World Bank. (2025). *Unemployment, total (% of total labor force) (national estimate) [SL.UEM.TOTL.NE.ZS]*. World Development Indicators. https://databank.worldbank.org/metadataglossary/world-development-indicators/series/SL.UEM.TOTL.NE.ZS