

# Prediction of student performance at polytechnic using machine learning approach

Kristina Hutajulu, Lili Ayu Wulandhari

Computer Science Department, BINUS Graduate Program, Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

---

## Article Info

### Article history:

Received Dec 28, 2023

Revised Apr 9, 2024

Accepted Jun 9, 2024

---

### Keywords:

Educational data mining

Machine learning

Regression model

Student performance

Waiting period for graduate employment

---

## ABSTRACT

Educational data mining (EDM) is a strategic technique for exploring data in educational environments to gain a deeper understanding of education. One of the goals of EDM is to predict things related to students in the future which can be done using a machine learning approach. In this paper, a regression model is developed to predict student performance in the first semester and the waiting period for graduate employment using machine learning approach based on informatics management (MI) and non-informatics management (non-MI) student data. Four regression models are compared for predicting student performance in the first semester and waiting period for graduate employment, including support vector regression (SVR), random forest regression (RFR), AdaBoost regression (ABR), and XGBoost regression. Based on the experiment, prediction of students' performance in the first semester, the highest R2 result produced by SVR model by value of 0.58 for MI and by RFR by value of 0.34 for non-MI. While, waiting period for graduate employment prediction, the highest R2 result produced by AdaBoost regression by value of 0.44 for MI and SVR by value of 0.39 for non-MI.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



---

## Corresponding Author:

Kristina Hutajulu

Computer Science Department, BINUS Graduate Program, Master of Computer Science, Bina Nusantara University

Jakarta, 11480, Indonesia

Email: kristina.hutajulu@binus.ac.id

---

## 1. INTRODUCTION

The quality of an educational institution is measured in part by the performance of its students. Polytechnics are obliged to prepare and produce competent graduates so they can compete in the world of work. One indicator of the quality of graduates who can compete in the world of work can be represented by the waiting period for graduates to get their first job. To produce skilled graduates, it is the responsibility of polytechnics to ensure that students maintain the expected level of performance from the first semester until the completion of their studies. The first semester is a significant adaptation period for students. The grade point in this semester can impact students' self-perception of their learning abilities and provide an early indication of potential academic success in the future. The grade point of students in their first semester can act as an indicator and early warning for polytechnics to provide additional support to graduates who exhibit poor performance. Essentially, student performance, as evidenced by their academic achievements, serves as the initial foundation for students to compete in the professional world. Success in the first semester can be influenced by the academic and social support received by students. The Polytechnic can do a selective admission process as an effort to attract potential students, considering factors such as their social background, educational history, and admission result.

Achieving positive results in the early semesters has a beneficial impact on students, motivating students to take a proactive approach to their studies. In addition to academic performance, students' involvement in non-academic activities in college, such as proficiency in foreign languages, character education, organizational experience, internship experience, and other activities related to soft skills, can significantly influence their ability to face the world of work. Polytechnics can enhance their understanding of their graduates' potential in the world of work by analyzing tracer study data and diploma supplements. Tracer study data provides insights into the waiting period for employment of graduates. Meanwhile, the diploma supplement offers a comprehensive overview of graduates' academic track record, encompassing proficiency in foreign languages, character education, organizational experience, internship experience, and other activities during their academic tenure.

Based on the explanation above, there is a need-to-know student performance at the beginning of the semester and the student's waiting period for their first job. This need can be met by conducting research with an educational data mining approach. Data mining can be utilized to uncover hidden patterns, thereby extracting crucial information from data. Educational data mining (EDM) is a strategic technique for exploring data in an educational environment to gain a deeper understanding of education [1]. The aim of EDM is to identify patterns from key factors influencing learning, enhance the scholarly knowledge of educators and learners, and predict learning patterns of students in the future. Prediction models are developed using machine learning models [2] based on data within the educational relevant to the issues at hand.

Previous research on predicting student performance by analyzing student data has been done [3], [4]. Zulfiker *et al.* [5] revealed that machine learning can be applied to predict student grades, enabling crucial actions for grade improvement using a classification approach. Rai *et al.* [6] used machine learning to classify student performance, finding that the random forest algorithm outperformed the support vector machine (SVM). Hashim *et al.* [7] and Bilal *et al.* [8] demonstrated that student characteristics, including demographic information, academic background, and behavioral features, can be utilized as training data for machine learning algorithms. Kumar *et al.* [9] did research on predicting student performance using a machine learning regression model based on previous academic data and family background.

Previous research on predicting waiting period for graduate employment has been executed to evaluate the employability of graduates in the workplace [10], [11]. Casuat and Festijo [12] applied three machine learning methods, namely decision trees (DT), random forest (RF), and support vector machine (SVM), to predict the employability of students. Amalia and Wibowo [13] conducted research on creating a predictive model for the waiting period for graduates' employment when obtaining their first job using the Naïve Bayes data mining classification algorithm. Abdulloh *et al.* [14] did a study comparing SMOTE, SMOTE-ENN, and SMOTE-Tomek combined with SVM to detect the employability of graduates using tracer study datasets.

Based on a review of previous research, it is necessary to conduct data mining of academic records in the polytechnic to discover patterns related to student performance and the quality of graduates. This research represents the polytechnic's effort to analyze student performance patterns based on educational background, social background, and admission result. The aim is to early identify the performance of accepted prospective students. Furthermore, early detection is required regarding the waiting period for graduates entering the workplace based on academic achievements and activities during their academic tenure.

The novelty of this research lies in its ability to predict student performance outcomes, which includes first-semester grade point and waiting period for employment of graduates, using a regression model. This research aims to develop regression models for the first-semester student performance and the waiting period for graduates' employment. In the regression modelling of student performance, it is developed using enrolment data and admission result to the polytechnic. In the regression modelling for the waiting period for graduates' employment, it is developed using data from tracer studies [15], diploma supplements, and academic performance indices over the three years of study. The novelty in this research also involves comparing four regression models for both prediction models. Four regression models are support vector regression (SVR), random forest regression (RFR), AdaBoost regression (ABR), and XGBoost regression [16]. The goal is to identify the regression model with the best performance for both cases. Both regression models can provide early warnings to polytechnics regarding student performance in the first semester, enabling them to anticipate performance in the next semester. This anticipation can impact the waiting period for students to secure their first job upon graduation.

## 2. METHOD

There are various methods that can be utilized to understand how data mining aids improvement in the education sector. In data mining, there are four primary methods: classification, prediction, association, and clustering. The processes involved in applying data mining to the education sector can vary, with the

number of stages ranging from four to eleven [17]. However, generally, the stages commonly used in the education sector are referred to as the cross-industry standard process for data mining (CRISP-DM). CRISP-DM is a systematic guide for implementing data mining across sectors, including education as shown in Figure 1 [18], [19]. It comprises six stages: business understanding, data understanding, data preparation, modelling, evaluation, and implementation.

In this research, the research methodology is designed based on the CRISP-DM framework. The CRISP-DM framework is modified according to the research objectives as shown in Figure 2. It begins with collecting the necessary data for developing a regression model for student performance at first semester and a regression model for the waiting period for graduates' employment at the polytechnic. Subsequently, data pre-processing is applied to the collected data to obtain a format suitable for the model. The results of data pre-processing are then used to create regression models. All models used in the experiments are evaluated to obtain the model with the best performance.

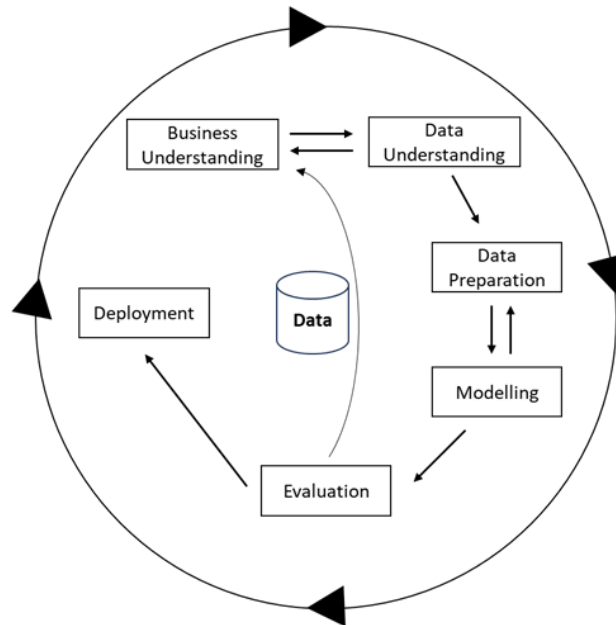


Figure 1. Cross-industry standard process methodology [20], [21]

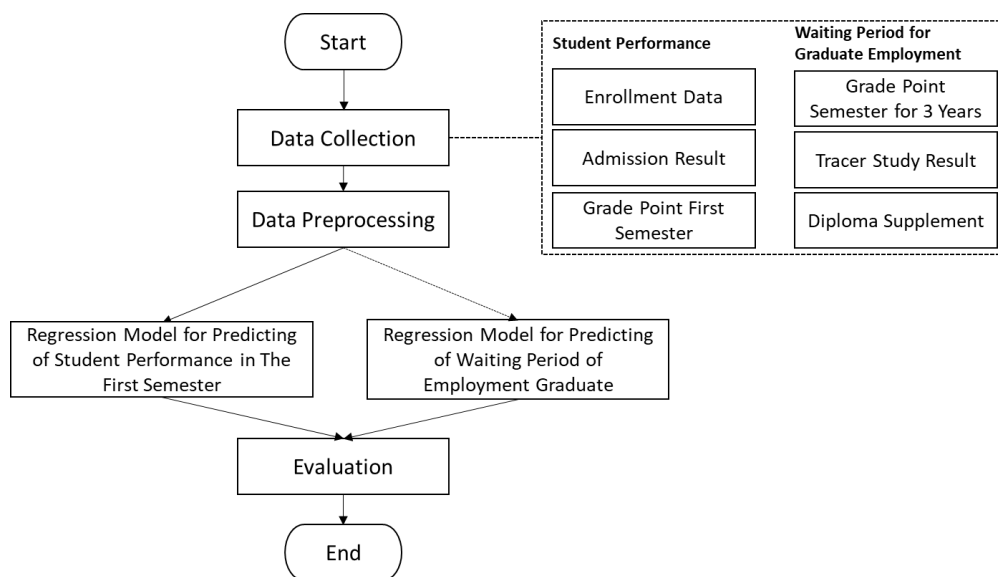


Figure 2. Research methodology

## 2.1. Data collection

To develop a regression model for predicting student performance, we utilize enrollment data, admission results data, and grade points at the first semester in Table 1. The target variable is the grade points at the first semester. The dataset for developing the regression model to predict student performance consists of 114 rows for informatics management (MI) majors and 617 rows for non-informatics management (non-MI) majors, with each dataset containing 29 features. The datasets are differentiated by unique features, specifically TPA test results for MI and Physics test results for non-MI.

To predict the waiting period for graduate employment, we utilize tracer study data, diploma supplement and grade points within the three years of study. The tracer study data include the waiting period for graduate employment, as a dependent variable. Diploma supplement encompasses internship experience, international language, character education, organizational experience, and other activities as shown in Table 2. The dataset consists of 47 rows for MI majors with 11 features in it (except international language because it only has one value) and 135 rows for non-MI majors with 12 features in it.

Table 1. Dataset for predicting student performance

No	Data source	Feature
1	Enrollment data	Registration year
2		Year of birth
3		Gender
4		Graduation year
5		Height
6		Weight
7		What order do you come in your family
8		Number of siblings
9		Home city
10		Home province
11		Father's education
12		Father's job
13		Mother's education
14		Mother's job
15		Parent average income
16		Parent city
17		Parent province
18		School origin
19		School city
20		School province
21		Accreditation
22		School majors
23		Average mathematics grade
24		Average physics grade
25		Average English grade
26	Admission result	Mathematics test results
27		Physics/TPA test results
28		English test results
29	Target variable	GPA semester 1

Table 2. Dataset for predicting waiting period for graduate employment

No	Data source	Feature
1	GP semester for three years	GP Semester 1
2		GP Semester 2
3		GP Semester 3
4		GP Semester 4
5		GP Semester 5
6		GP Semester 6
7	Diploma supplement	Internship experience
8		International language
9		Character education
10		Organizational experience
11		Other activities
12	Tracer study	Waiting period for graduate employment

## 2.2. Data pre-processing

Data preprocessing aims to clean, transform, and prepare raw data before the data is used in data analysis or statistical modeling [22]. In this research, the first step at this stage is to handle missing values.

Missing values are handled by filling in the mean, median or mode value. The second step involved data transformation by changing categorical variables into numerical format.

In the student performance dataset for MI major, there are several features that have missing values, including father's job, mother's job, average physics grade, and gap year. In the student performance dataset for non-MI majors, features with missing values include mother's job, father's job, mother's education, father's education, accreditation, average physics grades, and gap year. Then, after handling missing values, we encode features that have categorical values. Two transformation techniques are utilized, namely label encoder and one-hot encoder. The label encoder is utilized to convert ordinal categorical data, while the one-hot encoder is used for converting categorical data without a specific order. The label encoder is used for gender, father's education, mother's education, parent average income, accreditation, average mathematics grade, average physics grade, and average English grade. One-hot encoder is used to registration year, home city, parent city, home city, school city, school city, father's job, mother's job, and school major. After the data transformation, the dataset for predicting student performance comprises 76 features for the MI dataset and 98 features for the non-MI dataset.

The dataset for predicting the waiting period for graduate employment consists of 11 features for MI and 12 features for non-MI. Dataset for MI major, international language has only one value, so the feature is eliminated. In this case, there are features that have missing values, including international languages, character education, organizational experience, and other activities. The label encoder is utilized, resulting in a dataset for predicting the waiting period for graduate employment consisting of all features for both MI and non-MI.

Thorough dataset analysis is carried out to identify features that have an impact on the target variable. In the student performance prediction dataset, there are many features, so it needs to analyze and ignore features that have no impact for predicting student performance. Dataset for predicting of waiting period for graduate employment, analysis was also carried out to see which features had impact on this case. The feature selection in this research uses the feature importance permutation algorithm.

### 2.3. Regression model

To predict student performance and the waiting period for graduate employment, four regression techniques are employed: SVR, RFR, ABR, and XGBoost regression [23]–[25]. All four regression models are implemented with hyperparameter tuning to achieve optimal performance in Table 3. The selection of regression techniques is based on recent research in data mining literature and their strengths in prediction.

Each dataset will be split into three parts: training set, test set, and validation set. This data split is aimed at avoiding biased prediction results. In both cases, predicting student performance and predicting the waiting period for graduate employment, the data will be divided with an 80:10:10 ratio, meaning 80% for training set, 10% for test set, and 10% for validation set.

Table 3. Hyperparameters for each regression model

Regression model	Hyperparameter
Support vector regression	<i>Kernel, C, epsilon</i>
Random forest	<i>n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features, bootstrap</i>
AdaBoost regression	<i>max_depth, loss, n_estimators, learning_rate</i>
XGBoost regression	<i>max_depth, subsample, n_estimators, learning_rate</i>

### 2.4. Evaluation

During the experiment all the regression algorithms were executed using 5-fold cross validation to train the model. Each dataset was divided into five corresponding subsets, four were used for training the model, and one subset was used for model testing and validation. The performance of the four regression models will be evaluated using three metrics, namely mean absolute error (MAE), mean square error (MSE), and R square.

## 3. RESULTS AND DISCUSSION

Four regression models are compared, tested, and analyzed. The regression models are SVR, RFR, AdaBoost regression and XGBoost regression. In the experiment, the model was trained with a baseline dataset, with a normalized dataset and with feature selection. Feature selection use permutation feature importance algorithm.

**3.1. Student performance at first semester**

In the regression model for predicting student performance, the model is trained with a dataset that consists of 76 features for the MI dataset and 98 features for the non-MI dataset. Each model is trained through various experiments, including baseline, best parameters, and best features. The most optimal results from each regression model experiment are presented in Table 4. The model with the best performance for predicting the performance of students at the first semester in MI dataset is SVR while in non-MI dataset is RFR. Both SVR and RFR achieved their best performance by performing hyperparameter tuning, feature normalization, and feature selection. Feature selection used the permutation feature selection technique, and normalization was done using StandardScaler. In the MI dataset, SVR used hyperparameters  $C = 1$ ,  $\epsilon = 0.3$ ,  $kernel = rbf$ . In the non-MI dataset, RFR used hyperparameters  $bootstrap = True$ ,  $max\_depth = 30$ ,  $max\_features = 'sqrt'$ ,  $min\_samples\_leaf = 2$ ,  $min\_samples\_split = 2$ , and  $n\_estimators = 500$ . The best features in the best regression models for both datasets are presented in Table 5.

Table 4. Comparison of regression models to predict student performance at the first semester.

Regression model	MI dataset			Non-MI dataset		
	MAE	MSE	R <sup>2</sup>	MAE	MSE	R <sup>2</sup>
<b>Support vector regression</b>	<b>0.14</b>	<b>0.03</b>	<b>0.58</b>	0.27	0.21	0.22
<b>Random forest regression</b>	0.23	0.06	0.13	<b>0.27</b>	<b>0.18</b>	<b>0.34</b>
AdaBoost regression	0.16	0.04	0.41	0.30	0.23	0.13
XGBoost regression	0.24	0.08	-0.07	0.29	0.22	0.16

Table 5. Best features of the best regression models to predict student performance at the first semester

Dataset	Best feature
MI	Weight, number of younger siblings, home city = school city, parent average income, mathematics test results, English test results, PAT test result, registration year (2018 and 2019), father’s job (Teacher and Indonesian National Armed Forces), mother’s job (Housewife), average mathematics grade, average physics grade, average english grade, school city (Kab. Boyolali dan Kab. Pemalang), home city (Kab. Pemalang), parent city (Kab. Pemalang and Others).
Non-MI	Gender, home city = school city, accreditation, father’s job (Laborer, Employee, and Entrepreneur), mother’s job (Civil Servant, Teacher, Farmer, and Entrepreneur), category of senior high school (SMA and MA), school majors (Engineering), parent city (Kab. Bekasi, Kab. Karawang, Kab. Ponorogo, Kab. Purworejo, Kota Bekasi, Jakarta Timur, Jakarta Utara, and Other), Father’s Education, Parent Average Income, registration year (2016, 2018 and 2020), school city (Kab. Bekasi, Kab. Bogor, Kab. Boyolali, Kab. Karawang, Kab. Malang, Kab. Purworejo, Semarang, and Other), home city (Kab. Bekasi, Kab. Boyolali, Kab. Karawang, Kab. Kebumen, Kab. Malang, Kab. Ponorogo, Kab. Purworejo, Kota Bekasi, Jakarta Utara, and Other), average mathematics grade, average English grade, mathematics test results, physics test results, English test result.

Model performance on non-MI datasets is lower than on MI datasets. The MI dataset comprises a smaller dataset size, resulting in the RFR not exhibiting superior performance compared to SVR. Conversely, in the non-MI dataset, RFR demonstrates a slight superiority, albeit not significantly, attributable to the greater variation observed in the non-MI dataset compared to the MI dataset as shown in Figures 3 to 8. Furthermore, both XGBoost regression and AdaBoost regression have not yielded optimal results, highlighting the necessity for heightened attention towards appropriately tuning hyperparameters and also add more data to enhance model performance. The distribution can be observed that the distribution of parent city, home city and school city in non-MI has higher variations.

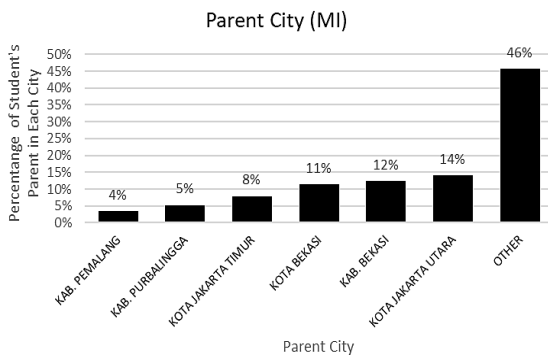


Figure 3. Distribution of parent city in MI dataset

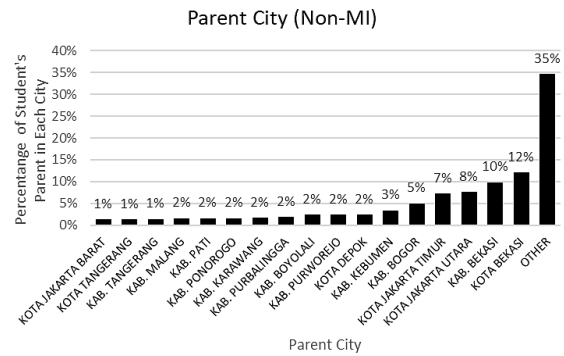


Figure 4. Distribution of parent city in non-MI dataset

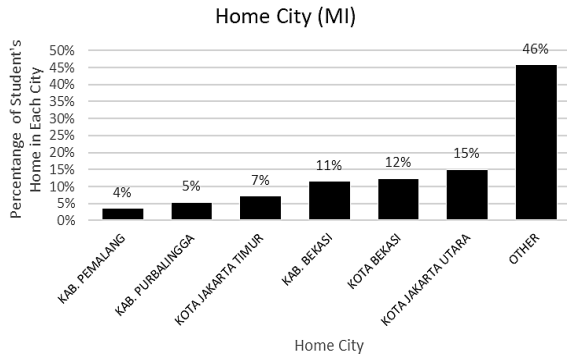


Figure 5. Distribution of home city in MI dataset

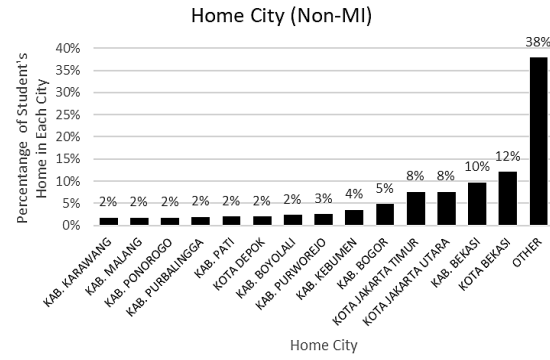


Figure 6. Distribution of home city in non-MI dataset

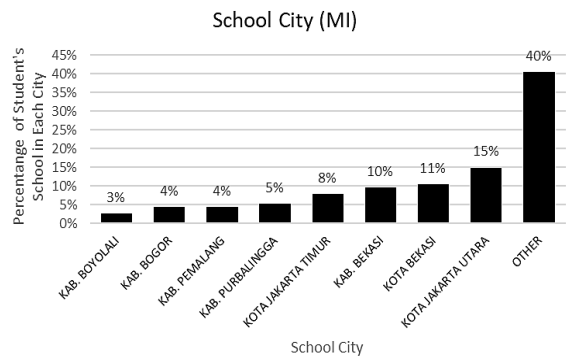


Figure 7. Distribution of school city in MI dataset

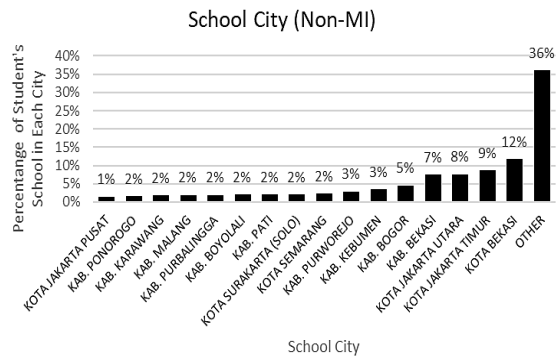


Figure 8. Distribution of school city in non-MI dataset

The weight feature in the non-MI dataset has a skewness value of 1.81 while the MI dataset has a skewness value of 1.4 so that the weight feature in the non-MI dataset has more skewed data compared to the MI dataset. In Figures 9 and 10, the feature weight in the non-MI dataset has more outliers than in the MI dataset. The English test result feature in the non-MI dataset has a skewness value of 1.16 while the MI dataset has a skewness value of 1.4 so that the English test result feature in the non-MI dataset has more skewed data compared to the MI dataset. In Figures 11 and 12, the feature English test result in the non-MI dataset has more outliers than in the MI dataset. Mathematics test result features as shown in Figures 13 and 14 have data distribution that tends to be the same but in the non-MI dataset there are outliers. The difference between the two datasets is the PAT test result feature on the MI dataset and the physics test result feature in the non-MI dataset. The PAT test result feature has a skewness value of 0.4 while the physics test result feature has a skewness value of -0.19. Considering the skewness value of both, the PAT test result feature in Figure 15 tends to have higher skewness, but the physics test result feature has many outliers in Figure 16. Outliers and skewed data have an unstable impact on model performance. This also means that the regression coefficient values can be very sensitive to the outlier values and skewness data, leading to high variability in the model results.

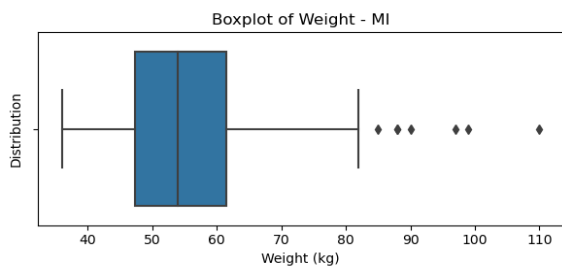


Figure 9. Boxplot of weight MI dataset

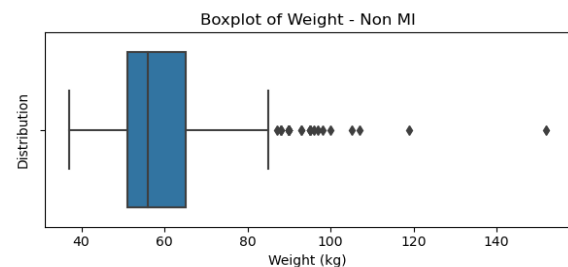


Figure 10. Boxplot of weight non-MI dataset

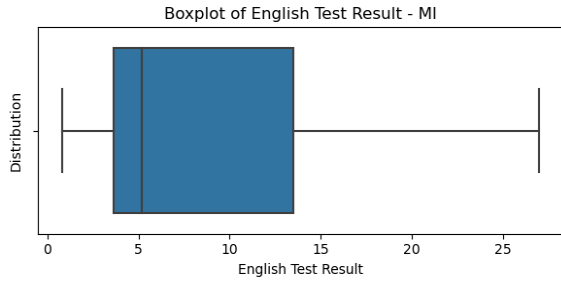


Figure 11. Boxplot of English test results MI dataset

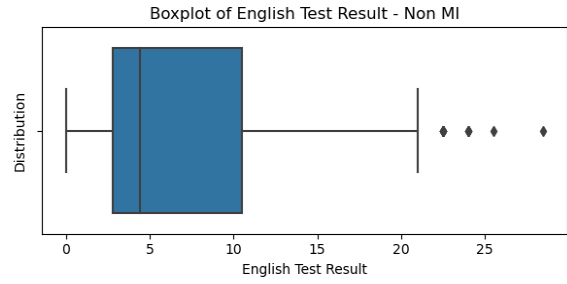


Figure 12. Boxplot of English test results non-MI dataset

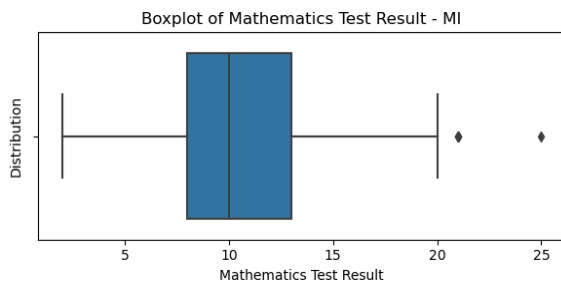


Figure 13. Boxplot of mathematics test results MI dataset

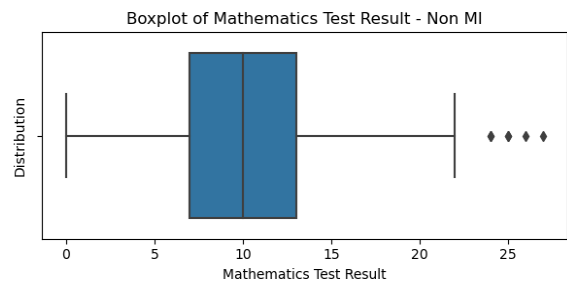


Figure 14. Boxplot of mathematics test results non-MI dataset

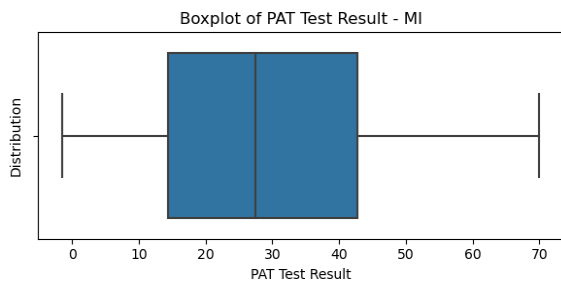


Figure 15. Boxplot of PAT test results

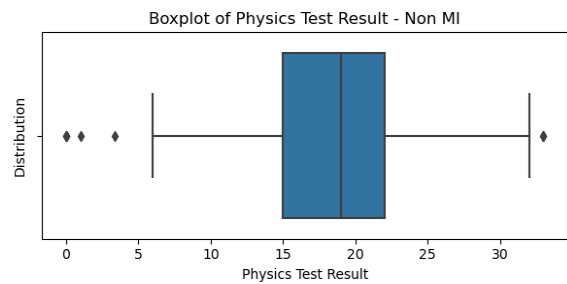


Figure 16. Boxplot of physics test results

### 3.1. Waiting period for graduate employment

In the regression model for predicting the waiting period for graduate employment, the model is trained with a dataset that consists of 10 features for MI dataset and 11 features non-MI datasets. Each model is trained through various experiments, including baseline, best parameters, and best features. The most optimal results from each regression model experiment are presented in Table 6. The model with the best performance for predicting waiting period for graduate employment in MI dataset is ABR while in non-MI dataset is SVR. Both SVR and ABR achieved their best performance by performing hyperparameter tuning and feature normalization. Normalization was done using Standard Scaler. In the non-MI dataset, SVR used hyperparameters  $C = 1, \epsilon = 0.1, \text{kernel} = \text{linear}$ . In the MI dataset, ABR used hyperparameters  $\text{max\_depth} = 2, \text{learning\_rate} = 0.001, n\_estimators = 400$ . The best features in the best regression models for both datasets are presented in Table 6.

Based on the evaluation results, the MAE does not exhibit significant differences among the various models. Specifically, in the MI dataset, AdaBoost regression outperforms the others, suggesting its effectiveness in managing small datasets with minimal outlier data. Conversely, in non-MI datasets, support vector machines demonstrate superior performance compared to other models in handling small datasets with numerous outlier data points.



Table 6. Comparison of regression models to predict waiting period for graduate employment

Model	MI dataset			Non-MI dataset		
	MAE	MSE	R <sup>2</sup>	MAE	MSE	R <sup>2</sup>
<b>Support vector regression</b>	0.74	0.59	0.34	<b>0.53</b>	<b>0.56</b>	<b>0.39</b>
Random forest regression	0.77	0.77	0.07	0.63	0.87	0.25
<b>AdaBoost regression</b>	<b>0.60</b>	<b>0.46</b>	<b>0.44</b>	0.69	0.99	0.15
XGBoost regression	0.85	0.79	0.04	0.76	1.08	0.06

#### 4. CONCLUSION

The best regression model for predicting students' performance in the first semester is SVR for MI dataset and RFR for non-MI dataset. For the MI dataset, SVR has an R<sup>2</sup> value of 0.58, indicating that the model can explain approximately 58% of the variation in the target data. This means that a significant portion of the variation remains unexplained, suggesting the presence of other factors not considered in the model or uncertainties in the data contributing to the remaining variability. For the non-MI dataset, RFR has an R<sup>2</sup> value of 0.34. Within the best model for student performance on the MI dataset, the MAE was found to be 0.14. Similarly, within the best model for student performance in the non-MI dataset, the MAE was found to be 0.27. For example, if a student's actual score is 3.0, the possible error in the predicted value could range from 2.86 to 3.14 for the MI dataset and from 2.73 to 3.27 for the non-MI dataset. This error range is still acceptable because within these values, the students' characteristics remain consistent, allowing for the same treatment to be applied.

The best regression model for predicting the waiting time of graduate employment is AdaBoost Regression for the MI dataset and SVR for the non-MI dataset. SVR performs with R<sup>2</sup> value of 0.39, while AdaBoost has a performance with R<sup>2</sup> value of 0.44. In this case, considering the performance of the generated models, the availability of the data and high variance become one of the problems to get more improvement of the model performance. Further research with more data can be applied to enhance it. Within the best model for the waiting period for graduate employment, the MI and non-MI dataset, the MAE was found to be 0.6 and 0.53. In this case, for example, the actual waiting time is 1 month, then the possible error in the prediction value that can occur is 0.4 months or 1.4 months for MI and 0.47 months or 1.53 months for non-MI. This error range is still acceptable because the only error value is no more than 1 month.

The predicted outcomes for student performance in the first semester and the waiting period for graduate employment exhibit an R<sup>2</sup> value below 0.5, falling within the low category within the context of regression analysis. These findings indicate that the model is unable to elucidate more than half of the variability in the target data. Consequently, considering the current R<sup>2</sup> results, additional research is warranted to investigate other factors influencing student performance in the first semester and the waiting period for graduates at polytechnic institutions.

#### ACKNOWLEDGEMENTS

The authors thank Bina Nusantara University for the research grant and supporting this research.




#### REFERENCES

- [1] S. M. Dol and D. P. M. Jawandhiya, "A review of data mining in education sector," *Journal of Engineering Education Transformations*, vol. 36, no. S2, pp. 13–22, Jan. 2023, doi: 10.16920/jeet/2023/v36is2/23003.
- [2] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381–386, 2020.
- [3] E. Alyahyan and D. Düşteğör, "Predicting academic success in higher education: literature review and best practices," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, Dec. 2020, doi: 10.1186/s41239-020-0177-7.
- [4] P. Dabhade, R. Agarwal, K. P. Alameen, A. T. Fathima, R. Sridharan, and G. Gopakumar, "Educational data mining for predicting students' academic performance using machine learning algorithms," *Materials Today: Proceedings*, vol. 47, pp. 5260–5267, 2021, doi: 10.1016/j.matpr.2021.05.646.
- [5] M. S. Zulfiker, N. Kabir, A. A. Biswas, P. Chakraborty, and M. M. Rahman, "Predicting students' performance of the private universities of Bangladesh using machine learning approaches," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 3, pp. 672–679, 2020.
- [6] S. Rai, K. A. Shastry, S. Pratap, S. Kishore, P. Mishra, and H. A. Sanjay, "Machine learning approach for student academic performance prediction," in *Advances in Intelligent Systems and Computing*, 2021, pp. 611–618.
- [7] A. Salah Hashim, W. Akeel Awadh, and A. Khalaf Hamoud, "Student performance prediction model based on supervised machine learning algorithms," *IOP Conference Series: Materials Science and Engineering*, vol. 928, no. 3, Nov. 2020, doi: 10.1088/1757-899X/928/3/032019.
- [8] M. Bilal, M. Omar, W. Anwar, R. H. Bokhari, and G. S. Choi, "The role of demographic and academic features in a student performance prediction," *Scientific Reports*, vol. 12, no. 1, Jul. 2022, doi: 10.1038/s41598-022-15880-6.
- [9] A. Kumar, K. K. Eldhose, R. Sridharan, and V. V. Panicker, "Students' academic performance prediction using regression: a case study," in *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, Jul. 2020, pp. 1–6.




- doi: 10.1109/ICSCAN49426.2020.9262346.
- [10] M. M. Usita, "Graduates employability analysis using classification model: a data mining approach," *Journal of Positive School Psychology*, vol. 6, no. 3, pp. 2788–2796, 2022.
- [11] A. Miranda and K. M. Lhaksamana, "Classification analysis of waiting period for Telkom University alumni to get jobs using decision tree and support vector machine," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 2, Sep. 2022, doi: 10.47065/bits.v4i2.1963.
- [12] C. D. Casuat and E. D. Festijo, "Predicting students' employability using machine learning approach," in *2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, Dec. 2019, pp. 1–5, doi: 10.1109/ICETAS48360.2019.9117338.
- [13] R. Amalia and A. Wibowo, "Prediction of the waiting time period for getting a job using the naive Bayes algorithm," *International research journal of advanced engineering and science*, vol. 5, no. 2, pp. 225–229, 2020.
- [14] F. F. Abdulloh, M. Rahardi, A. Aminuddin, S. D. Anggita, and A. Y. A. Nugraha, "Observation of imbalance tracer study data for graduates employability prediction in Indonesia," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 8, 2022, doi: 10.14569/IJACSA.2022.0130820.
- [15] D. Megasari, A. Puspitorini, and D. Lutfiati, "Employability tracer study of cosmetology education graduates at the Universitas Negeri Surabaya," *International Joint Conference on Arts and Humanities 2021 (IJCAH 2021)*, 2021, pp. 937–940, doi: 10.2991/assehr.k.211223.162.
- [16] L. H. Alamri, R. S. Almuslim, M. S. Alotibi, D. K. Alkadi, I. Ullah Khan, and N. Aslam, "Predicting student academic performance using support vector machine and random forest," in *2020 3rd International Conference on Education Technology Management*, Dec. 2020, pp. 100–107, doi: 10.1145/3446590.3446607.
- [17] A. Al-Rawahnaa, "Data mining for education sector, a proposed concept," *Journal of Applied Data Sciences*, vol. 1, no. 1, pp. 1–10, Sep. 2020, doi: 10.47738/jads.v1i1.6.
- [18] C. Schröder, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.
- [19] N. S. Sapare and S. M. Beelagi, "Comparison study of regression models for the prediction of post-graduation admissions using machine learning techniques," in *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Jan. 2021, pp. 822–828, doi: 10.1109/Confluence51648.2021.9377162.
- [20] P. Chapman, *CRISP-DM 1.0: step-by-step data mining guide*, SPSS, 2000.
- [21] R. Wirth and J. Hipp, "CRISP-DM: towards a standard process model for data mining," *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, vol. 1, pp. 29–39, 2000.
- [22] S.-A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, "Data preprocessing in predictive data mining," *The Knowledge Engineering Review*, vol. 34, Jan. 2019, doi: 10.1017/S026988891800036X.
- [23] G. Shanmugasundar, M. Vanitha, R. Čep, V. Kumar, K. Kalita, and M. Ramachandran, "A comparative study of linear, random forest and Adaboost regressions for modeling non-traditional machining," *Processes*, vol. 9, no. 11, Nov. 2021, doi: 10.3390/pr9112015.
- [24] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," *Interactive Learning Environments*, vol. 31, no. 6, pp. 3360–3379, Aug. 2023, doi: 10.1080/10494820.2021.1928235.
- [25] S. Amjad, M. Younas, M. Anwar, Q. Shaheen, M. Shiraz, and A. Gani, "Data mining techniques to analyze the impact of social media on academic performance of high school students," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–11, Mar. 2022, doi: 10.1155/2022/9299115.

## BIOGRAPHIES OF AUTHORS



**Kristina Hutajulu**    is a laboratory assistant in the Informatics Management Study Program. She completed her diploma education at Politeknik Astra in 2019. Subsequently, she pursued her bachelor's degree in information system starting in 2019 at Universitas Bina Nusantara and successfully completed it in 2021. She can be contacted at email: kristina.hutajulu@binus.ac.id.



**Lili Ayu Wulandhari**    is a data scientist and computer science lecturer. Experienced in machine learning for research scale and real implementation. Involving in data acquisition, visualization, analysis and modelling to accelerate business strategy using qualified knowledge of data science and machine learning. She can be contacted at email: lili.wulandhari@binus.ac.id.