ComTech: Computer, Mathematics and Engineering Applications, 13(1), June 2022, 53−62
DOI: 10.21512/comtech.v13i1.7388

P-ISSN: 2087-1244
E-ISSN: 2476-907X

Prediction of Undergraduate Student's Study Completion Status Using MissForest Imputation in Random Forest and XGBoost Models

Intan Nirmala[1]*, Hari Wijayanto[2], and Khairil Anwar Notodiputro[3]

[1-3]Department of Statistics, Institut Pertanian Bogor
Jln. Raya Dramaga, Jawa Barat 16680, Indonesia
[1]inirmalanirmala@apps.ipb.ac.id; [2]hari@apps.ipb.ac.id; [3]khairil@apps.ipb.ac.id

Received: 31st May 2021/ **Revised:** 06th August 2021/ **Accepted:** 09th August 2021

How to Cite: Nirmala, I., Wijayanto, H., & Notodiputro, K. A. (2022). Prediction of Undergraduate Student's Study Completion Status Using MissForest Imputation in Random Forest and XGBoost Models. *ComTech: Computer, Mathematics and Engineering Applications, 13*(1), 53−62. https://doi.org/10.21512/comtech.v13i1.7388

Abstract **-** The number of higher education graduates in Indonesia is calculated based on their completion status. However, many undergraduate students have reached the maximum length of study, but their completion status is unknown. This condition becomes a problem in calculating the actual number of graduates as it is used as an indicator of higher education evaluation and other policy references. Therefore, the unknown completion status of the students who have reached the maximum length of study must be predicted. The research compared the performance of Random Forest and Extreme Gradient Boosting (XGBoost) classification models in predicting the unknown completion status. The research used a dataset containing 13.377 undergraduate students' profiles from the Higher Education Database (PDDikti), Ministry of Education, Culture, Research, and Technology. The dataset was incomplete, and the proportion of missing data was 20,9% of the total data. Because missing data might lead to prediction bias, the research also used MissForest imputation to overcome the missing data in the classification modelling and compared it to Mean/Mode and Median/Mode imputation. The results show that MissForest outperforms the other two imputations in both classifiers but requires the longest computation time. Furthermore, the XGBoost model with MissForest is significantly superior to the Random Forest model with MissForest. Hence, the best model chosen to predict the completion status is XGBoost with MissForest imputation.

Keywords: study completion status; MissForest imputation; Random-Forest model; XGBoost model

I. INTRODUCTION

Every tertiary education institution in Indonesia is obliged to submit its higher education data to PDDikti which is coordinated by the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia. It is a system that collects higher education data from all tertiary education institutions, which are integrated at a national scale. One of the submitted data is about students' study completion status, whether they have graduated or dropped out. The number of higher education graduates is calculated based on this completion status.

According to the 2020 National Higher Education Standards (Menteri Pendidikan dan Kebudayaan Republik Indonesia, 2020), undergraduate programs' maximum length of study is seven years. Those who have reached the maximum period shall have a completion status recorded on PDDikti. However, many students have reached the maximum length of study, but their completion status is unknown. There are two possibilities concerning this case. First, the students have graduated, but the universities do not report their graduation status. Second, students do not continue their studies, and there is no further information. This condition makes it difficult to calculate the actual number of graduates. Moreover, the number of graduates is used as an indicator of higher education evaluation and other policy references. Therefore, the unknown students' study completion status with the maximum length of study must be predicted.

Prediction of student completion status can be carried out with a classification model. One of them is a classification tree. The advantage of this method is that it does not depend on certain assumptions, such

*Corresponding Author

as the normality of data distribution. The problem of instability and high variance of a single tree can be overcome by ensembling the classification tree (Wang & Wu, 2018).

Some previous studies in education have used data mining algorithms with a classification tree as their base learner. Kurniawan, Anggrawan, and Hairani (2020) proposed a graduation prediction system for undergraduate students of Bumigora University using a classification tree algorithm. However, their study was limited to a small dataset and one type of classifier. Then, Yuliansyah, Imaniati, Wirasto, and Wibowo (2021) used a larger dataset to predict students' graduation on time in the engineering faculty of some private universities in Indonesia using a classification tree. They also only used one type of classifier but compared the validation result on a various number of testing data.

In contrast, Hussain, Dahan, Ba-Alwib, and Ribata (2018) compared four classification methods (J48, PART, Random Forest, and Bayes Network) to predict students' performance from three different colleges in India. The result showed that Random Forest outperformed the other classifiers. Similarly, Baruah, Baruah, and Goswami (2020) predicted students' academic performance in an engineering college in India using seven different classifiers (J48, Random Forest, Rap Tree, Logistic Model Tree (LMT), Naïve Bayes, BayesNet, and PART). They found that Random Forest was the most efficient algorithm among all the considered algorithms. Then, Yan (2021) used some machine learning algorithms to predict students' performance in China. Those algorithms were Extreme Gradient Boosting (XGBoost), Random Forest, Lasso, Elastic Net, Support Vector Machine, and Classification Tree. The XGBoost model achieved the best result than five other classic machine learning models.

Based on those previous studies mentioned, the research compares two ensemble tree methods: Random Forest and XGBoost. It can predict the unknown completion status of undergraduate students who have reached the maximum length of study. Moreover, the previous studies use a complete dataset to build a classification model. Meanwhile, the research uses data with missing values. In addition, the data used are also larger than in the previous studies.

Random Forest is an extended Bagging method in which the training data are resampled by repeated bootstrap, and some classification trees are built based on the bootstrapping result (Breiman, 2001). The training procedure for Random Forest is summarized by Ahmad, Mourshed, and Rezgui (2018) in the following steps. First, it performs bootstrap sampling from the original dataset. Second, for each bootstrap drawn in the first step, it grows an unpruned tree by randomly sampling *m* variables from the input variables and selects the best split from among those variables. Third, the first and second steps are repeated in *k* times until a forest consisting of *k* trees is formed. Fourth, it predicts new data by aggregating the prediction of all trees.

XGBoost is an extended Gradient Boosting method with a penalty component on the loss function to prevent overfitting. Compared to the traditional Gradient Boosting, this method has higher speed and performance owing to the parallel nature in which trees are built (Aminu, Abdulkarim, Aliyu, Aliyu, & Turaki, 2019). If the Random Forest method builds trees parallel, XGBoost builds trees sequentially. On XGBoost, every new tree is built to reduce the mistake of the previous tree (Anwar, Winarno, Hadikurniawati, & Novita, 2021).

As previously mentioned, the data used in the research are incomplete, and the proportion of the missing data is 20,9% of the total data. Missing data may cause bias in the parameter estimates of analysis (Blazek, Zwieten, Saglimbene, & Teixeira-Pinto, 2021). Hence, imputation for the missing value is performed in the pre-processing step of classification modelling (Khan & Hoque, 2020).

The simplest imputation replaces missing data with mean or median for numerical data and mode for categorical data. Various imputation methods have been used in many fields, e.g., Hot-Deck Imputation, Principal Component Analysis (PCA), K-Nearest Neighbors (KNN) (Troyanskaya et al., 2001), and MissPALasso (Städler, Stekhoven, & Bühlmann, 2014). However, these methods only work on one type of data, which is only numerical or categorical. For mixed data, imputation is conducted separately according to the type of data that ignores the relationship between numerical and categorical variables.

According to Stekhoven and Bühlmann (2012), MissForest imputation can work on mixed data simultaneously and have a non-parametric character. It does not depend on certain assumptions of data distribution. In MissForest, X_s is the s-th variable containing the missing value in $i_{mis}^{(s)} \subseteq \{1, \dots, n\}$. Furthermore, $y_{obs}^{(s)}$ is the observed value of the X_s , and $y_{mis}^{(s)}$ is the missing value of the X_s. Variables other than X_s with $i_{obs}^{(s)} = \{1, \dots, n\} \setminus i_{mis}^{(s)}$ are denoted by $x_{obs}^{(s)}$. Moreover, variables other than X_s having observation correspond to $i_{mis}^{(s)}$ are denoted by $x_{mis}^{(s)}$. Figure 1 illustrates the partition of the dataset in the MissForest imputation.

MissForest starts by replacing all missing values with initial values, which can be Mean/Median/Mode or other imputation values. Furthermore, the variables of X_s with $s = 1, \dots, p$ are sorted from small to large according to their amount of missing value. The missing value is imputed for each X_s by constructing a Random Forest model using the $y_{obs}^{(s)}$ as the response and the $x_{obs}^{(s)}$ as the predictor. Furthermore, the missing value $y_{mis}^{(s)}$ can be predicted by implementing the model to the $x_{mis}^{(s)}$. The new prediction result replaces the previous imputation value. This procedure is conducted iteratively until the stopping criterion γ is met if the difference between the value of the latest imputation and the previous one increases for the first time.

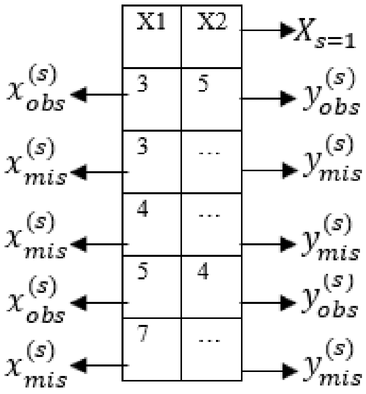

Figure 1 Dataset Partition in MissForest

According to Stekhoven and Bühlmann (2012), MissForest is outperformed by Multivariate Imputation by Chained Equation (MICE) and K-Nearest Neighbors (KNN). Then, the previous research has compared the effect of Mean, Median, KNN, MICE, and MissForest to impute the simulated missing data for Naïve Bayes and Decision Tree Classifier (Cihan, Kalıpsız, & Gökçe, 2019). The most successful imputation in both classifiers is MissForest. Then, according to Alsaber, Pan, and Al-Hurban (2021), several methods can be conducted to impute missing data for the air quality monitoring dataset. MissForest, Bayesian PCA, Predictive Mean Matching (PMM), KNN, and Expectation Maximization imputation are compared. It shows that MissForest is the only method with a consistent and comparatively lower imputation error.

The main objective of the research is to compare the performance of MissForest imputation, Mean/ Mode imputation, and Median/Mode imputation to overcome missing data in predicting the completion status of undergraduate students who have reached the maximum length of study. The next objective is to compare the Random Forest and XGBoost algorithms used as the classifiers. Then, the best model will be chosen to predict the unknown completion status of undergraduate students who have reached the maximum length of study. The goodness of fit employed is accuracy, sensitivity, specificity, G-Mean, and Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC).

II. METHODS

All data in the research are from the PDDikti database from the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia. The data consist of 13.377 samples of undergraduate students who were admitted in 2013 from public and private universities. These samples are taken by stratified random sampling method with proportional allocation from the population of 936.236 undergraduate students admitted in 2013. The determination of strata is based on the type of institution working unit, accreditation of study programs, and field of study. All the samples have reached the maximum length of study in 2020, but some of them have unknown completion status.

The research uses 17 variables, consisting of 10 numerical predictors, 6 categorical predictors, and 1 binary response. The numerical predictors include the average number of credits in each semester (X1), cumulative Grade Point Average (GPA) (X2), average semester GPA (X3), number of courses with an index score of 0−0,99 (X4), number of courses with an index score of 1−1,99 (X5), number of courses with an index score of 2−2,99 (X6), number of courses with an index score of 3−4 (X7), students' age when they are registered for the first time (X8), number of total semesters taken by students (X9), and number of the active semester (X10). All those numerical predictors contain missing values in some observations.

Then, the average number of credits in each semester is assumed to affect the study load for students. Therefore, it can have an impact on their completion status. Furthermore, the index score is used to measure students' performance in a course. The value of this index ranges from 0 to 4. The higher the index score is, the better the students' performance will be in a course. Index scores can be converted to letter scores. For example, an index score of 4 is equivalent to an index letter of A, an index score of 3 is equivalent to an index score of B, and so forth. Cumulative GPA, semester GPA, and index score represent students' academic performance. According to Costa, Bispo, and Pereira (2018), there is evidence that the higher the students' academic performance is, the lower the risk of dropout will be.

Next, the number of total semesters and active semesters is related to students' attendance. Attendance has a significant indirect effect on academic performance (Kim, Shakory, Azad, Popovic, & Park, 2020). Students' age is also one of the features that most researchers agree on when predicting students' academic achievement in higher education (Alturki, Hulpuş, & Stuckenschmidt, 2020).

The categorical variables include accreditation of higher education institution (X11), accreditation of study program (X12), type of institution working unit (X13), the field of study (X14), inactive history (X15), and gender (X16). The response (Y) is the students' completion status, whether they have graduated or dropped out. All categorical variables are complete, except for inactive history and gender, which need to be imputed.

Moreover, the accreditation of higher education institutions and study programs represents the quality of education. They are assumed to have an impact on students' performance. Based on both accreditation types, the research groups each education institution into four categories: not accredited, good, very good, and excellent. Meanwhile, the type of institution working unit divides education institutions based on their finance and human resources management. There are four categories of institutions based on it: Perguruan Tinggi Negeri (PTN), Perguruan Tinggi Negeri Berbadan Hukum (PTN-BH), Perguruan Tinggi Swasta (PTS), and Badan Layanan Umum (BLU). It is also assumed to have an impact on students' performance.

Furthermore, the field of study is related to the level of difficulty experienced by students. The inactive history categorizes students into two groups: those who have ever taken an inactive semester and have never taken an inactive semester. This variable is related to students' attendance, which affects academic performance. Meanwhile, like the age variable, gender is also one of the features that most researchers agree on when predicting students' academic achievement (Alturki et al., 2020).

The data consist of 13.377 rows of observations and 17 columns of variables that produce 227.409 matrix cells. The proportion of the missing data is 20,9% of the total cells. Table 1 shows the proportion of missing data for each variable. The cumulative GPA, the semester GPA, and the average number of semester credits are the variables with the highest proportion of missing data. In contrast, the accreditation of higher education institutions and study programs, the type of institution working unit, and the field of study do not contain any missing data in their observations.

The entire procedure of the research is completed using R software. The research uses some packages, such as MissForest, Caret, and Random Forest. The procedure involves splitting the data into two datasets: data with complete response variables (11.875 observations) and data with unknown response variables (1.502 observations). The data with complete responses are split again into two parts: 80% as the training data to build the model and 20% as the testing data to evaluate the model.

Moreover, the missing data should be dealt with carefully before analysis. Otherwise, the information extracted from the dataset containing missing values will lead to the wrong decision-making (Manimekalai & Kavitha, 2018). The imputation is also performed separately for training and testing data to avoid data leakage (Marcinkevics, Reis Wolfertstetter, Wellmann, Knorr, & Vogt, 2021). Then, three approaches are made to overcome the missing training data: MissForest imputation, Mean/Mode imputation, and Median/Mode imputation. The MissForest imputation uses mtry = $\sqrt{}$the number of predictors = 4 and ntree = 50. The same approaches also impute missing data in the testing data. From the three imputation approaches, there are three completed datasets. Each of them consists of completed training and testing data.

Table 1 Proportion of Missing Data

No.	Variable	Number of Missing Data in Each Variable	Missing Data Proportion (%) for Each Variable in the Total Data
1	X1	4.311	9,06
2	X2	5.926	12,45
3	X3	4.327	9,09
4	X4	3.939	8,27
5	X5	3.958	8,31
6	X6	3.893	8,18
7	X7	3.996	8,39
8	X8	3.226	6,78
9	X9	3.450	7,25
10	X10	3.365	7,07
11	X11	0	0,00
12	X12	0	0,00
13	X13	0	0,00
14	X14	0	0,00
15	X15	3.051	6,41
16	X16	2.663	5,59
17	Y	1.502	3,15
TOTAL		47.607	100

Random Forest and XGBoost classification models are built from the completed training data in the research. The research uses the default values for all hyperparameters in both classifiers. Then, the classification model and imputation performance are evaluated based on the completed testing data. The research repeats all processes ten times, from the splitting data into training and testing data until the evaluation of the model to see the stability of the models. The average performance of each classification model is calculated from the entire repetition. Then, the model with the best performance is chosen to predict the unknown completion status of the 1.502 undergraduate students.

III. RESULTS AND DISCUSSIONS

The research uses 13.377 samples of admitted undergraduate students in 2013. About 88,77% of the observations are employed for modelling. Then, the data are split into training and testing data with a ratio of 80:20. The research has split data into the training and testing data ten times. So, it has produced ten different sets of training and testing data. Furthermore, the unknown responses of the other 11,23% observations are predicted using the classification model that has been constructed.

Missing data can affect the performance of a classification model. However, the missing data in the research happen randomly. The observed variables or the missing values do not influence the occurrence of missing data. Therefore, the missing data are assumed to be Missing Completely at Random (MCAR) and can be completed by imputation technique. Imputation is a technique to replace missing data with certain values obtained based on information from the available dataset (Kokla, Virtanen, Kolehmainen, Paananen, & Hanhineva, 2019). In classification modelling, imputation is carried out in the pre-processing stage. The research separates the imputation of training data and testing data separately to avoid information leakage. However, the imputation of the testing data utilizes the information from the imputation result of training data.

The imputation method used in the research is MissForest imputation. The first step is to replace all the missing data with the Mean or Mode of the observed variables. Furthermore, the variables containing missing data are sorted based on the amount of missing data from small to large. The Random Forest model is built using the observed variables for each variable. Then, the model predicts the missing data. This procedure is repeated iteratively until a stopping criterion is found. It is when the difference between the new imputed value and the previous value increases for the first time in numerical and categorical data.

As previously mentioned, MissForest imputation is an imputation that works based on the Random Forest algorithm. In MissForest, each tree is built using the sample obtained from the bootstrap process. Each bootstrap sample randomly leaves out about one-third of the observations. These left-out observations for a given tree are called Out of Bag (OOB) (Schonlau & Zou, 2020). OOB observations are not included in the tree-building process. MissForest performance can be measured based on predicted and assumed OOB as testing data. Imputation performance on numerical data is measured by Normalized Root Mean Square Error (NRMSE) and categorical data by Proportion of Falsely Classification (PFC). Based on Table 2, MissForest imputation in the research yields an average NRMSE of 0,451 and PFC of 0,058, calculated based on OOB. Imputation performance is categorized as good if NRMSE and PFC are close to 0. On the contrary, it is considered not good if it is close to 1 (Stekhoven & Bühlmann, 2012). NRMSE and PFC in the research appear to be close to 0, so the performance of MissForest is relatively good.

Table 2 Performance of MissForest Based on OOB

***n*-th train data**	**NRMSE**	**PFC**
1	0,452	0,057
2	0,453	0,057
3	0,449	0,057
4	0,450	0,058
5	0,455	0,058
6	0,451	0,058
7	0,451	0,058
8	0,451	0,058
9	0,452	0,059
10	0,450	0,058

As a comparison, imputation is also conducted using Mean/Mode imputation and Median/Mode imputation. In the Mean/Mode imputation, the mean of all values within the same attribute is calculated and imputed in the missing data cells (Khan, Khan, & Singh, 2018). Meanwhile, Median/Mode imputation replaces the numerical missing value with the median of all values within the same attribute. In both methods, mode substitution can be used instead if the attribute is categorical (Acuña & Rodriguez, 2004).

Neither Mean/Mode imputation nor Median/Mode imputation produces OOB observations as in MissForest. Consequently, the NRMSE and PFC in both methods can only be measured if the complete observations are available before. So, they can be compared to the imputation result. The data are a real case that contains the missing value from the beginning. It causes NRMSE and PFC of Mean/Median/Mode imputation not to be measured. Therefore, a performance comparison of MissForest, Mean/Mode, and Median/Mode imputation is carried out after building the classification model, using the goodness of fit of the classification models.

Based on Table 3, MissForest computation time is much longer than the other two methods. It occurs due to the MissForest algorithm complexity. It is influenced by the proportion of missing values, number of variables, and number of observations. Conversely, the Mean/Mode and Median/Mode imputation procedures are less complicated. They only replace the missing value with Mean/Median/Mode without complex algorithms. This procedure requires a shorter computation time. However, Mean imputation for missing values leads to large errors in variance estimates when variables have linear relationships (Köse, Özgür, Coşgun, Keskinoğlu, & Keskinoğlu, 2020). This condition also applies to the use of Median and Mode to impute the missing data.

Tables 4 and 5 present the average performance of Random Forest and XGBoost models with the three different imputation methods. In Random Forest and XGBoost, at the significance level of 5%, models with MissForest appear to be significantly better than the other two methods. Column P1 in Tables 4 and 5 is the p-value of paired t-test between MissForest and Mean/Mode imputation performance. Meanwhile, column P2 presents the p-value of paired t-test between MissForest and Median/Mode imputation performance. The alternative hypothesis in these tests is that the average of MissForest performance is better than the average of the other imputation performances. MissForest excels in all measures, regardless of the only slight difference from the Mean/Median/Mode imputation performance. It is simultaneously affirmed by the performance distribution of each model in Figure 2. Mean/Mode imputation has a slightly better average performance than Median/Mode imputation.

Table 6 shows the average classification performance in both models. Regardless of the imputation method used, performances of Random Forest and XGBoost differ significantly. It is indicated by the p-value of paired t-test of both models in Table 6. Most of the p-values are significant at a significance level of 5%. The alternative hypothesis carried out is that the performance of Random Forest is different from XGBoost. According to Table 6, Random Forest and XGBoost with MissForest imputation are the only models with an overall average performance of more than 90. Those two models are significantly different in all measurements, except for G-Mean.

Table 3 Computation Time of MissForest, Mean/Mode, and Median/Mode Imputation

***n*-th Train Data**	**MissForest (In Second)**	**Mean/Mode (In Second)**	**Median/Mode (In Second)**
1	688,320	0,240	0,017
2	626,890	0,015	0,018
3	651,134	0,022	0,023
4	698,493	0,015	0,017
5	674,990	0,018	0,021
6	781,132	0,024	0,031
7	658,769	0,021	0,024
8	621,354	0,015	0,020
9	614,350	0,020	0,028
10	619,922	0,015	0,018

Table 4 Comparison of Random Forest Performance with Three Different Imputations

Goodness of Fit	Random Forest				
	MissForest (1)	Mean/Mode (2)	Median/Mode (3)	P1	P2
Accuracy	93,98	92,33	92,00	2E-08	9E-11
Sensitivity	90,72	87,17	86,54	3E-06	8E-07
Specificity	94,90	93,77	93,52	2E-08	2E-09
G-Mean	92,78	90,41	89,96	2E-07	8E-09
AUC	97,62	96,17	95,88	4E-08	1E-08

Table 5 Comparison of XGBoost Performance with Three Different Imputations

Goodness of Fit	XGBoost				
	MissForest (1)	Mean/Mode (2)	Median/Mode (3)	P1	P2
Accuracy	94,42	92,18	91,83	5E-09	5E-09
Sensitivity	90,21	86,38	85,78	5E-06	8E-08
Specificity	95,67	93,83	93,54	1E-08	8E-08
G-Mean	92,90	90,02	89,58	2E-07	8E-09
AUC	97,77	95,85	94,00	2E-09	5E-11

Note:
P1 = p-value of paired t-test of model performance with imputation 1 and 2 (Ha: imputation 1 > imputation 2),
P2 = p-value of paired t-test of model performance with imputation 1 and 3 (Ha: imputation 1 > imputation 3).

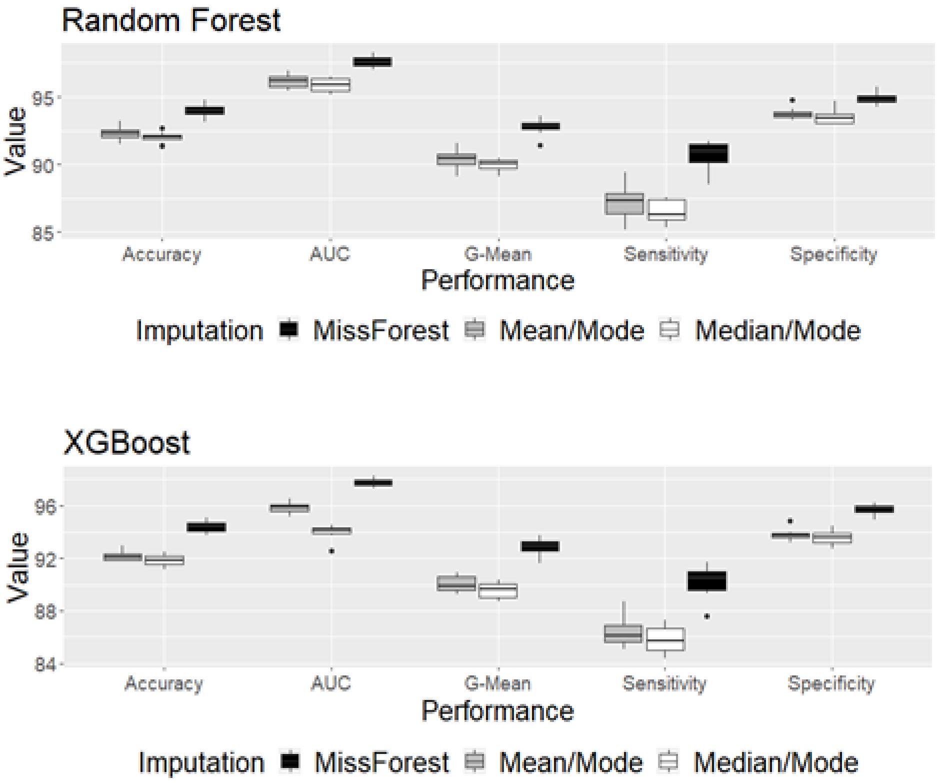

Figure 2 Boxplot of Classification Performance with Three Different Imputations

Table 6 Comparison of Random Forest and XGBoost Classification Models

Imputation Method	Goodness of Fit	Random Forest		XGBoost		P-Value of Paired t-test
		Average	Standard Deviation	Average	Standard Deviation	
MissForest	Accuracy	93,98	0,45	94,42	0,42	0,00
	Sensitivity	90,72	1,01	90,21	1,21	0,02
	Specificity	94,90	0,47	95,67	0,40	0,00
	G-Mean	92,78	0,60	92,90	0,66	0,34
	AUC	97,62	0,43	97,77	0,32	0,02
Mean/Mode	Accuracy	92,33	0,48	92,18	0,38	0,08
	Sensitivity	87,17	1,23	86,38	1,14	0,02
	Specificity	93,77	0,44	93,83	0,42	0,50
	G-Mean	90,41	0,73	90,02	0,60	0,02
	AUC	96,17	0,53	95,85	0,43	0,01
Median/Mode	Accuracy	92,00	0,39	91,83	0,46	0,12
	Sensitivity	86,54	0,84	85,78	1,08	0,00
	Specificity	93,52	0,53	93,54	0,54	0,87
	G-Mean	89,96	0,47	89,58	0,62	0,01
	AUC	95,88	0,49	94,00	0,56	0,00

Note:
P-value = p-value of paired t-test between the average of Random Forest and XGBoost performances
(Ha: the performance of Random Forest model ≠ the performance of XGBoost model).

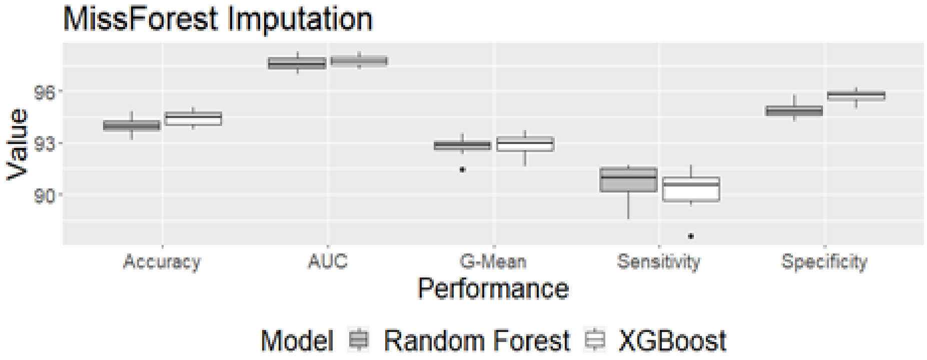

Figure 3 Boxplot of MissForest Performance on Random Forest and XGBoost Classification Models

Then, Figure 3 shows the distribution of Random Forest and XGBoost performances on the data imputed by MissForest. The median of XGBoost performance tends to be better than Random Forest, except for sensitivity. It is also confirmed by paired t-test between the models in Table 7. The alternative hypothesis used is that the performance of XGBoost with MissForest is superior to Random Forest with MissForest. At a significance level of 5%, XGBoost is significantly superior to Random Forest except for G-Mean and sensitivity. Hence, the best model chosen is XGBoost with MissForest imputation, with an average accuracy of 94,42%, sensitivity of 90,21%, specificity of 95,67%, G-Mean of 92,90%, and AUC of 97,77%.

The results show that 1.502 undergraduate students have reached the maximum length of study with unknown completion status. The best model is used to predict this unknown completion status. Before the prediction, the missing value of the data having unknown responses is imputed with MissForest. The imputation obtains NRMSE of 0,412 and PFC of 0,060. It is categorized as performing well. Furthermore, responses of those imputed data are predicted using the best model.

According to the best model, 1.502 students with unknown completion status are predicted. About 62,12% of the students are dropouts, and 37,88% have graduated. This prediction can estimate the actual number of graduates from undergraduate programs. It also helps the government to evaluate higher education performance in Indonesia and make certain decisions.

Table 7 Paired T-Test between MissForest Performance in Random Forest and XGBoost Classification Models

Goodness of Fit	P-Value
Accuracy	0,00
Sensitivity	0,99
Specificity	0,00
G-Mean	0,17
AUC	0,01

Note:
P-value = p-value of paired t-test between average Random Forest and XGBoost performance on the data imputed by MissForest (Ha: performance of XGBoost model > performance of Random Forest model).

IV. CONCLUSIONS

Many undergraduate students have reached the maximum length of study, but some have unknown completion status. The research compares the performance of Random Forest and XGBoost models in predicting the unknown completion status. A dataset containing 13.377 undergraduate students' profiles from the PDDikti is used. However, the dataset is incomplete, and the proportion of missing data is 20,9% of the total data. The research also compares MissForest, Mean/Mode, and Median/Mode imputation to cope with missing data.

The results show that MissForest imputation on Random Forest and XGBoost models outperform Mean/Mode and Median/Mode imputations. Meanwhile, the XGBoost model with MissForest is significantly superior to the Random Forest model with MissForest, except for sensitivity and G-Mean. The best model chosen is XGBoost with MissForest with an average accuracy of 94,42%, sensitivity of 90,21%, specificity of 95,67%, G-Mean of 92,90%, and AUC of 97,77%. According to the best model, 1.502 students with unknown completion status are predicted. It shows 62,12% of the data are dropouts, and 37,88% have graduated. This prediction can estimate the actual number of graduates from undergraduate programs. It also helps the government to evaluate higher education performance in Indonesia and to make certain decisions. However, the scope of the research is limited to predicting the completion status of undergraduate students. It can be conducted in more levels of higher education in future research.

Despite its superior performance, MissForest imputation has a drawback in computational efficiency. For future research, selecting a smaller number of trees and mtry can be used to reduce the computation time. It does not significantly reduce the accuracy, but it must also be adjusted to the size and complexity of the dataset.

REFERENCES

Acuña, E., & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. In *Classification, clustering, and data mining applications* (pp. 639−647). Berlin: Springer.

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2018). Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression. *Energy*, *164*, 465–474. https://doi.org/10.1016/j.energy.2018.08.207

Alsaber, A. R., Pan, J., & Al-Hurban, A. (2021). Handling complex missing data using random forest approach for an air quality monitoring dataset: A case study of Kuwait environmental data (2012 to 2018). *International Journal of Environmental Research and Public Health*, *18*(3), 1–25. https://doi.org/10.3390/ijerph18031333

Alturki, S., Hulpuș, I., & Stuckenschmidt, H. (2020). Predicting academic outcomes: A survey from 2007 till 2018. *Technology, Knowledge and Learning, 27*, 275–307. https://doi.org/10.1007/s10758-020-09476-0

Aminu, A. A., Abdulkarim, A., Aliyu, A. Y., Aliyu, M., & Turaki, A. M. (2019). Detection of phishing websites using Random Forest and XGBoost algorithms. *International Journal of Pure and Applied Sciences, 2*(3), 1–14.

Anwar, M. T., Winarno, E., Hadikurniawati, W., & Novita, M. (2021). Rainfall prediction using Extreme Gradient Boosting. *Journal of Physics: Conference Series*, *1869*, 1–5. https://doi.org/10.1088/1742-6596/1869/1/012078

Baruah, E. A., Baruah, S., & Goswami, J. A. (2020). Comparative analysis of different classification algorithms based on students' academic performance using WEKA. *IOSR Journal of Computer Engineering (IOSR-JCE), 22*(1), 49–56.

Blazek, K., Zwieten, A. V., Saglimbene, V., & Teixeira-Pinto, A. (2021). A practical guide to multiple imputation of missing data in nephrology. *Kidney International*, *99*(1), 68–74. https://doi.org/10.1016/j.kint.2020.07.035

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5−32. https://doi.org/10.1023/A:1010933404324

Cihan, P., Kalıpsız, O., & Gökçe, E. (2019). Effect of imputation methods in the classifier performance. *Sakarya University Journal of Science, 23*(6), 1225–1236.

Costa, F. J. D., Bispo, M. D. S., & Pereira, R. D. C. D. F. (2018). Dropout and retention of undergraduate students in management: A study at a Brazilian Federal University. *RAUSP Management Journal*, *53*(1), 74–85 https://doi.org/10.1016/j.rauspm.2017.12.007

Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science, 9*(2), 447–459.

Khan, F. U. F., Khan, K. U. Z., & Singh, S. K. (2018). Is Group Means imputation any better than Mean imputation: A study using C5.0 classifier. *Journal of Physics: Conference Series*, *1060*, 1–5. https://doi.org/10.1088/1742-6596/1060/1/012014

Khan, S. I., & Hoque, A. S. M. L. (2020). SICE: An improved missing data imputation technique. *Journal of Big Data*, *7*(1), 1–21.

Kim, A. S. N., Shakory, S., Azad, A., Popovic, C., & Park, L. (2020). Understanding the impact of attendance and participation on academic achievement. *Scholarship of Teaching and Learning in Psychology*, *6*(4), 272–284. https://doi.org/10.1037/STL0000151

Kokla, M., Virtanen, J., Kolehmainen, M., Paananen, J., & Hanhineva, K. (2019). Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: A comparative study. *BMC Bioinformatics, 20*, 1−10. https://doi.org/10.1186/s12859-019-3110-0

Köse, T., Özgür, S., Coşgun, E., Keskinoğlu, A., & Keskinoğlu, P. (2020). Effect of missing data imputation on deep learning prediction performance for vesicoureteral reflux and recurrent urinary tract infection clinical study. *BioMed Research International*, *2020,* 1–15. https://doi.org/10.1155/2020/1895076

Kurniawan, D., Anggrawan, A., & Hairani. (2020). Graduation prediction system on students using C4.5 algorithm. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, *19*(2), 358–366. https://doi.org/10.30812/matrik.v19i2.685

Manimekalai, K., & Kavitha, A. (2018). Missing value imputation and normalization techniques in myocardial infarction. *ICTACT Journal on Soft Computing, 8*(03), 1655–1662.

Marcinkevics, R., Reis Wolfertstetter, P., Wellmann, S., Knorr, C., & Vogt, J. E. (2021). Using machine learning to predict the diagnosis, management and severity of pediatric appendicitis. *Frontiers in Pediatrics*, *9*, 1–12. https://doi.org/10.3389/fped.2021.662183

Menteri Pendidikan dan Kebudayaan Republik Indonesia. (2020). *Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 3 Tahun 2020 Tentang Standar Nasional Pendidikan Tinggi*. Retrieved from https://jdih.kemdikbud.go.id/arsip/Salinan%20PERMENDIKBUD%203%20TAHUN%202020%20FIX%20GAB.pdf

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal, 20*(1), 3–29. https://doi.org/10.1177/1536867X20909688

Städler, N., Stekhoven, D. J., & Bühlmann, P. (2014). Pattern alternating maximization algorithm for missing data in high-dimensional problems. *Journal of Machine Learning Research*, *15*(1), 1903–1928.

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ... & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics, 17*(6), 520-525.

Wang, C., & Wu, H. (2018). A new machine learning approach to house price estimation. *New Trends in Mathematical Sciences*, *6*(4), 165–171. https://doi.org/10.20852/ntmsci.2018.327

Yan, K. (2021). Student performance prediction using XGBoost method from a macro perspective. In *2021 2nd International Conference on Computing and Data Science (CDS)* (pp. 453–459). IEEE. https://doi.org/10.1109/CDS52072.2021.00084

Yuliansyah, H., Imaniati, R. A. P., Wirasto, A., & Wibowo, M. (2021). Predicting students graduate on time using C4. 5 algorithm. *Journal of Information Systems Engineering and Business Intelligence, 7*(1), 67–73.