PREDICTING POST-SECONDARY EDUCATION ENROLMENT IN INDONESIA

A COMPARISON OF LOGISTIC REGRESSION & MACHINE LEARNING MODELS

ILSE KLEVERLAAN

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY

DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

STUDENT NUMBER

2075147

WORD COUNT

7772

THESIS COMMITTEE

Supervisor: dr. B. Nicenboim

Second reader: dr. F. Blain

LOCATION

Tilburg University

School of Humanities & Digital Sciences

Department of Cognitive Science & Artificial Intelligence

Tilburg, The Netherlands

DATE

December 2, 2022

Abstract

High rates of post-secondary education enrolment are beneficial on both an individual and societal level. Post-secondary education enrolment is increasing worldwide. However, Indonesia's post-secondary enrolment is not rising as strongly as in surrounding countries. To increase the post-secondary education enrolment rates, it is important to identify the factors that contribute to enrolment. Previous research has focused on explaining these factors using statistical analysis. The present study attempts to extend the existing literature by looking at how post-secondary education can best be predicted with machine learning models. It does this by answering the following research question: To what extent can post-secondary education enrolment be predicted by socioeconomic factors, demographics and gender using machine learning techniques? To answer this question the machine learning algorithms K-nearest neighbours, random forest and AdaBoost are compared against a statistical model, logistic regression, as evaluated by AUC-ROC. The data that is used in this study is the Demographics and Health Survey, specifically the 2017 Indonesia version. Results indicate that logistic regression is outperformed by machine learning models. Specifically, AdaBoost showed the best performance, but the differences between the models is small. This is in line with the existing literature. Only small differences in feature importance between AdaBoost and logistic regression, and between genders were found. The present study has some limitations, including the lack of a strict method for feature selection and the large number of dummies that has to be created in order for KNN to work with categorical data.

Data Source/Code/Ethics Statement

Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. The author of this thesis acknowledges that

they do not have any legal claim to this data. The code used in this thesis is available at https://gitfront.io/r/Ilse-k/J4hCuLtfpGEF/thesis/. All figures and tables in this thesis are made by the author.

1 Problem Statement & Research Goal

Attending post-secondary education is an important factor for succeeding in several areas of life (Ma et al., 2020). On an individual level, it leads to better employability, higher income and social mobility. It is also beneficial on a societal level. In a population with a high rate of post-secondary education enrolment, there are lower poverty rates, higher tax payments, lower unemployment rates, and a more active participation in society (Ma et al., 2020). In the past decade, governments all over the world have recognized the importance of post-secondary education, and as a result, enrolment rates have been increasing all around the world (OECD, 2022). This is also the case in Indonesia. However, Indonesia's enrolment rates cannot keep up with the enrolment rates of surrounding countries (Digdowiseiso, 2020). Therefore, it is important for Indonesia to look at how they can encourage more students to enrol in post-secondary education.

To achieve higher rates of post-secondary education enrolment, it is important to identify the factors that underlie the decision to continue education after secondary school. Various factors have been found to be associated with post-secondary education enrolment, such as location, family attitudes and socioeconomic status (Agger et al., 2018; Digdowiseiso, 2020; Hardy & Marcotte, 2020; Vandelannote & Demanet, 2021). Gender has also been identified as an influential factor in post-secondary education enrolment. However, the results of research that looked into this have been inconsistent (Agger et al., 2018; Indrahadi & Wardana, 2020; Richardson et al., 2020; Wenang et al., 2022).

While it is important to identify these factors, this is not sufficient for predicting post-secondary education enrolment. Being able to predict post-secondary education enrolment could be useful to both secondary schools as well as policy makers at a higher level, such as the government. Secondary schools would be able to identify students at risk of not continuing their education after graduation, so they could encourage them to continue. For policy makers, being able to predict enrolment may help them develop more targeted interventions to encourage students to enrol in post-secondary education. Scientifically, the primary focus of the existing literature has been explanatory, using statistical analyses, but the predictive element is still missing in the literature. This presents a gap in the literature, which the present study aims to address.

Taking all this together, the research goal of this study is to investigate to what extent socioeconomic factors, demographics and gender can predict who will enrol in post-secondary education. To achieve the goal of the study and attempt to fill the identified research gaps, the following research question has been formulated:

To what extent can post-secondary education enrolment be predicted by socioeconomic factors, demographics and gender using machine learning techniques?

To answer this main question, three sub research questions have been defined:

SRQ1: *Comparing AdaBoost, random forest and K-nearest neighbours, which model performs best evaluated by AUC-ROC as compared to logistic regression?*

SRQ 2: *What is the difference in feature importance between the best performing machine learning model and logistic regression?*

SRQ 3: *Does feature importance vary between boys and girls?*

The main findings of the present study show that machine learning models perform slightly better than the statistical model, that is logistic regression. More specifically, AdaBoost performed best, but the differences between all models were small. When looking at the

difference in the top five most important features between the best performing machine learning model and logistic regression, only small differences were found. The three most important features were the same for both models, while the fourth and fifth feature differed. Additionally, no major differences in feature importance were detected when comparing the five most important features for each gender. The same five predictors appear in the top five of both genders, but in a different order. Having a bank account appeared to be most important in both models.

2 Literature Review

In this section the theoretical background of this study is described. It starts with an overview of the existing literature about the predictors of post-secondary education enrolment. Subsequently, literature from related areas that form the background for the methodology of the current study is reviewed.

Previous research has identified various factors that influence post-secondary education enrolment, including socioeconomic status (SES), demographics and gender (Agger et al., 2018; Digdowiseiso, 2020; Hardy & Marcotte, 2020; Vandelannote & Demanet, 2021). Regarding socioeconomic status, prior studies have established that students with lower SES are less likely to enrol in post-secondary education (Digdowiseiso, 2020; Hardy & Marcotte, 2020; Vandelannote & Demanet, 2021). When examining this relationship in Indonesia, it appeared that low SES students are restricted in their choice of post-secondary institutions. This is because institutions with high tuition fees are not affordable for those students due to their family having a low income (Digdowiseiso, 2020). The same study also found that the choice of post-secondary institutions is further restricted for these students due to their families not being able to pay for training for, for example, entrance exams.

Additionally, demographics (e.g. location, ethnicity, civil status) have been found to influence post-secondary education enrolment (Batyra & Pesando, 2020; Chankseliani et al., 2020; Digdowiseiso, 2020; Hyseni Duraku et al., 2020; Khattab, 2018; Rumble et al., 2018). Students from rural areas are less likely to attend post-secondary education than students from urban areas for several reasons, with the main reason in Indonesia being that there are fewer post-secondary education institutions in the area (Chankseliani et al., 2020; Digdowiseiso, 2020). In a similar way, students from minority groups, married students, and students with children have relatively lower post-secondary education rates than students from the majority group, unmarried students, and students without children (Batyra & Pesando, 2020; Hyseni Duraku et al., 2020; Indrahadi & Wardana, 2020; Khattab, 2018; Rumble et al., 2018).

Furthermore, the existing literature suggests that gender plays a role in post-secondary education enrolment (Agger et al., 2018; Indrahadi & Wardana, 2020; Richardson et al., 2020; Wenang et al., 2022). However, the findings of studies about this relationship have been inconsistent. Some studies suggest that girls achieve a higher education level than boys (Agger et al., 2018; Richardson et al., 2020), while other studies found the opposite (Indrahadi & Wardana, 2020; Wenang et al., 2022).

In short, looking at the existing literature, it becomes apparent that socioeconomic status, demographics and gender are important factors to consider when predicting post-secondary education enrolment, which forms the base for the main research question of this study. Gender is specifically interesting to include due to the inconsistent findings of previous studies. This also relates to our third sub research question.

However, the existing literature focuses on explaining a relationship between a small number of factors and post-secondary education enrolment. While they have been successful in doing so, they are missing a predictive element. To my knowledge, no studies up-to-date have looked into predicting post-secondary education enrolment using machine learning models. As

such, to determine which models are best to use in the current study, we will look into research that uses machine learning models in a related area, namely drop out prediction. In the domain of drop out prediction, there is no unanimous agreement about which model performs best (Berens et al., 2018; Kemper et al., 2020; Wan Yaacob et al., 2020). Previous research has compared several models with each other, such as decision trees, random forest, K-nearest neighbours (KNN), AdaBoost and neural networks, and all models showed a relatively good performance between 79% and 95%. AdaBoost, decision trees and random forest have shown to outperform the other models, while KNN was used most often. Logistic regression was used as a baseline in all studies. Interestingly, it was found to perform similar or even better than the machine learning models (Berens et al., 2018; Kemper et al., 2020; Wan Yaacob et al., 2020).

Although drop out prediction is related to predicting post-secondary education enrolment, there is a limitation in comparing the two. This has to do with the fact that there is a class imbalance in drop out prediction (Kemper et al., 2020), while this is not an issue in the present study. Therefore, to create a more comprehensive background for the current study, it is important to consider studies that may be less related to the topic of post-secondary education enrolment, but use demographic information and socioeconomic factors as predictors in a classification problem with balanced classes. Among these studies, there is also no unanimous agreement about the best performing model (Dwi Fajar Maulana et al., 2020; Li et al., 2022; Zulfiker et al., 2021). Similar machine learning methods were used and similar results were found in these studies compared to the ones about drop out prediction. AdaBoost and random forest also outperformed other models in studies with balanced classes with a performance between 85% and 96% (Dwi Fajar Maulana et al., 2020; Zulfiker et al., 2021). Additionally, XGBoost demonstrated good results with a performance between 82% and 92% (Li et al., 2022; Zulfiker et al., 2021). However, differences in performance between the models that were compared were small. Other models that were included in these studies (e.g. KNN, gradient

boosting, decision trees and support vector machines) also performed relatively well with a performance between 64% and 89%. Similar to research about drop out prediction, logistic regression was used as a baseline in all studies (Dwi Fajar Maulana et al., 2020; Li et al., 2022; Zulfiker et al., 2021). In summary, among studies with balanced classes using socioeconomic status, demographics and/or gender as predictors, AdaBoost and random forest most often performed best, while KNN was used most often and logistic regression was used as a baseline.

In conclusion, based on these previous studies, AdaBoost and random forest appear to be most likely to perform best, which is why they are included in the present study. Additionally, KNN is used in the current study, as it is was the most used model and performed only slightly worse than the best performing model in related studies. All previous research used logistic regression as a baseline model, which why this is also done in the current study. Each model will be explained in more detail in section 3.2.

3 Methodology & Experimental Set-up

This chapter describes the data science pipeline that is gone through in this study. A graphical representation of all steps can be found at the end of the chapter (see figure 2).

3.1 Dataset Description

This section details the dataset used in this study. The dataset that is used is part of the Demographics and Health Survey Program (DHS). This is a survey conducted in developing countries, funded by the U.S. Agency for International Development, that measures a large number of demographic variables and health variables. Most variables are measured in all countries, but there are a few variables that are country-specific. The data collection is executed by a local organisation and the data is owned by the country in which the data was collected. The survey has consistently proven to include accurate and representative data (*Who We Are*,

n.d.). In this study, the 2017 Indonesia version of the DHS is used. The data is not publicly available. Access has to be requested via the DHS website (https://dhsprogram.com/).

The original dataset contains 49,627 rows and 5,494 columns. However, only a subset of the complete dataset is used, because only specific rows and columns are relevant to this study. This will be explained further in the next section. The target classes are roughly balanced with 54.1% having completed secondary education and 45.9% attending or have attended post-secondary education.

3.2 Train Test Split

Before data pre-processing, the data is first split into a training and a test set to avoid data leakage. 70% of the data is used as the training set and 30% is used as the test set. Previous studies have shown that this split optimizes performance (Vrigazova, 2021; Nguyen et al., 2021). The train test split is performed in a stratified way. A stratified split ensures that the class proportions found in the whole dataset remain the same in the test and training set (Bhagat & Bakariya, 2022). This is preferable in the present study, because we are working with a balanced dataset and we want to prevent class imbalances in the training and test set caused by the splitting of the data. By performing a stratified split, we ensure that the classes in both the train and test set remain balanced.

3.3 Target Group Selection

After splitting the data, the cleaning and pre-processing of the data starts. The first step in the data pre-processing pipeline is to select only those rows that contain data from participants that have at least completed secondary education. All participants who indicated they had no education, incomplete primary education, complete primary education or incomplete secondary education are discarded. Moreover, only participants under the age of

30 are considered. Because post-secondary education enrolment in Indonesia has strongly increased over the past few decades and the education system has undergone major changes (Digdowiseiso, 2020), data from these participants is not relevant for predicting new enrolments. Additionally, participants that are not a resident of the house they were interviewed in are also excluded as these participants have many features with the value *not a dejure resident* or *visitor,* which is not informative for prediction. This leaves us with 10,237 participants. To avoid data leakage, this selection is done for the train and test set separately.

3.4 Feature Selection, Construction, and Transformation

The next step in the pre-processing pipeline is manual feature selection. Features that were not measured in the 2017 Indonesia DHS, or are not related to socioeconomic factors, demographics or gender were excluded. To avoid data leakage, this is done for the train and test set separately. The method of manual feature selection has some disadvantages, which are discussed in section 6. The final feature selection can be found in Appendix A.

Although the dataset only includes variables related to demographics, socioeconomic factors and gender at this point, there are still a few variables that are not relevant for prediction. These are variables such as *number of children* or *civil status*. Those variables are not directly relevant as they could have happened after post-secondary education enrolment. To overcome this problem, three new variables are created. From the variables *age of marriage* and *civil status*, a new variable *married* is constructed. This variable indicates whether or not a participant was married at the end of secondary school. Participants that were never married or married after the age of 17 are assigned *no* and participants that were married before the age of 18 are assigned *yes*. Similarly, the new variable *cohabitation* is created from the variables *living with partner* and *age of first cohabitation*. This variable indicates whether or not the person was living together with their partner at the end of

secondary school. To construct this variable, participants that never lived with a partner and whose age was 18 or above when they first lived together with a partner are assigned *no*, while participants that were living with their partner before the age of 18, are assigned *yes*. Lastly, the variable *children* is constructed by subtracting the age of the participant when their first child was born from their current age and assigning those with an age of 18 and above *no*, and those that were under 18 *yes*.

Next to variable construction, some variable transformations are also performed. *Years lived in place of residence*, *time to get to water source*, *age of household head* and *ideal number of children* are transformed from factors to numerical variables. To be able to do so, the value *always* in *years lived in place of residence* is set to the current age of the participant. For *time to get to water source*, the value *on premises* is set to zero, while *don't know* is set to NA. *Age of household head* contains the value *non-numeric response*, which is set to NA. Feature selection, construction and transformation are performed on the training and test set separately to avoid data leakage. After feature selection, construction and transformation, there are 41 predictors, 1 ID variable and 1 outcome variable left in the dataset.

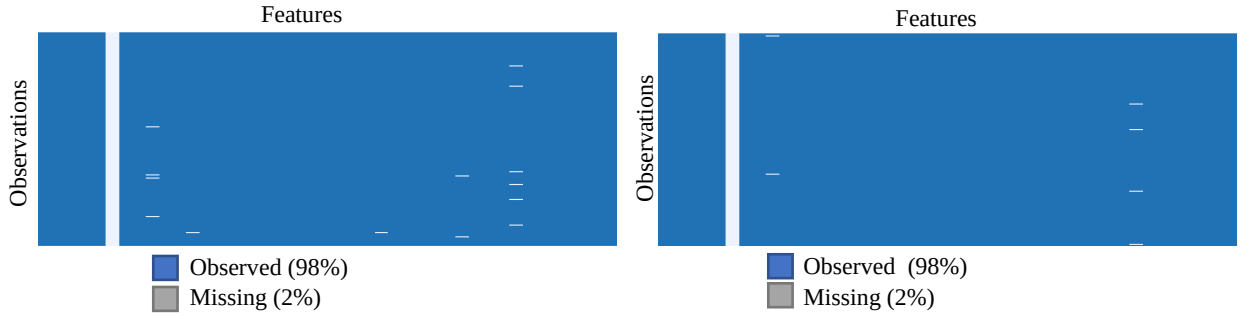
3.5 Missing Data

Now that the dataset only contains variables that are related to the subject and that are relevant for prediction, it is important to look at the missing values in the data. Firstly, we check how much missing data there is in the dataset. Figure 1 displays a missingness map of all the missing data in the dataset. After further investigation, it appears that all values for the

variable *childhood place of residence* are missing. As a consequence of all values missing for this variable, it was discarded.

Figure 1

Missing values in the test and training set.



All other variables contain no or at most five percent missing values, which is filled using missing value imputation. The K-nearest neighbours (KNN) imputation method is used in this study. Despite KNN being a simple imputation method, more complex methods (such as MICE) have not always proven to perform better (Liao et al., 2014). As such, the choosing a more complex model does not necessarily guarantee better imputation. Moreover, there is a very small percentage of missing data in the dataset, meaning that the final results are likely not influenced by the selection of a particular imputation method. The data is standardized as required by the KNNImpute function from the CARET package. In order to standardize the data, first the data is centred, meaning that the mean of the variable is subtracted from the original value, and after that, the data is scaled, which means that the centred value is divided by the standard deviation of that variable. Imputation is done using the default number of neighbours (5) as recommended in the documentation (*preProcess Function - RDocumentation*, n.d.). Both standardizing the data and missing value imputation is done for the test set and training set separately. For the training set, imputation is done within each fold of cross-validation. This is done to prevent data leakage.

3.6 One-Hot Encoding

To make the predictions necessary for answering the research questions, four models will be run: logistic regression, random forest, AdaBoost and KNN. As KNN does not handle categorical variables directly, categorical predictors are transformed into dummy variables, which is also known as one-hot encoding. KNN is the only model that requires categorical variables to be one-hot encoded. Logistic regression, Random Forest and Adaboost can handle categorical variables, so the original data could be used to run those models. However, because we use KNN as an imputation method, the dummy-coded data is used for all models.

3.7 Standardizing Numerical Data

Another pre-processing action that needs to be carried out in order for KNN and logistic regression to work correctly is standardizing the numerical data (Malato, 2022). This includes the dummy variables created in the previous step. Random forest and Adaboost are not sensitive to scale differences, so the original values could be used. However, as was the case for one-hot encoding, we use the standardized data for all models, because KNN is used as an imputation method. The standardizing of the data is the same procedure as was necessary for the missing value imputation. As such, the data are first centred and then scaled. This is done for the test set and training set separately. For the training set, imputation is done within each fold in cross-validation. This is done to prevent data leakage.

3.8 Removing Highly Correlated Variables and Variables with Near-Zero Variance

Now that the data is relevant and in the correct form for the models to handle, it is important to make sure that there are only useful predictors in the model, meaning that all highly correlated variables should be removed from the model and variables with little variance should also be removed from the model. Excluding variables that are highly

correlated can make the model work better (Tološi & Lengauer, 2011; Khondoker et al., 2016). Features with a correlation higher than 0.95 are excluded from the models.

Moreover, features with near-zero variance are removed from the model. Deciding whether a feature has near-zero variance is based on two criteria (Kuhn, 2008):

1. A low percentage of unique values (< 10%).
2. A high proportion of the most frequent value to the second most frequent value (> 20).

Including features with near-zero variance is not informative for prediction and can negatively affect model performance (Kuhn, 2008). Removing both highly correlated predictors and predictors with near-zero variance is done separately for the training and test set. For the training set, removing these variables is done within each fold in cross-validation. This is done to avoid data leakage.

3.9 Algorithms

To help answer the research questions, four algorithms are implemented: logistic regression, random forest, AdaBoost and KNN. As discussed in the literature review, logistic regression was chosen as a baseline, because it was widely used by previous studies (Berens et al., 2018; Kemper et al., 2020; Wan Yaacob et al., 2020; Dwi Fajar Maulana et al., 2020; Li et al., 2022; Zulfiker et al., 2021). Adaboost and random forest showed good results in previous studies using similar classification tasks, which is the motivation to include these models in the current study (Dwi Fajar Maulana et al., 2020; Li et al., 2022; Zulfiker et al., 2021). Likewise, KNN was most often used in similar classification problems, and performed only slightly worse than tree-based models (e.g. AdaBoost and random forest) (Berens et al., 2018; Kemper et al., 2020; Wan Yaacob et al., 2020; Dwi Fajar Maulana et al., 2020; Li et al., 2022; Zulfiker et al., 2021). Below each model will be explained further.

3.9.1 Logistic Regression

Logistic regression is a statistical model used to predict a binary outcome variable. Predictors can be both categorical and numerical (Niu, 2018), which fits well with the current dataset. Logistic regression is a common approach in social sciences, including educational science (Niu, 2018). Many explanatory studies that looked into post-secondary education enrolment used logistic regression in their research gender (Agger et al., 2018; Digdowiseiso, 2020; Hardy & Marcotte, 2020; Vandelannote & Demanet, 2021). Moreover, in machine learning models, it is more often used as a baseline model (Berens et al., 2018; Kemper et al., 2020; Wan Yaacob et al., 2020; Dwi Fajar Maulana et al., 2020; Li et al., 2022; Zulfiker et al., 2021), which is also done in the current study. The logistic regression model is fast and easy to interpret. However, it makes assumptions about the data, such as a linear relationship between the predictors and outcome variable, while this is most often not the case in real life data (GeeksforGeeks, 2022).

In this study, the logistic regression model is trained once and evaluated on the test set to assess its performance. This is because it does not have any parameters to tune and no further decisions regarding this model have to be made, so cross-validation is not necessary.

3.9.2 Random Forest

Random forest is a tree-based machine learning model using a large number of classification trees to predict the correct class in a classification problem. Each classification tree predicts a class and the class that gets predicted most is the final prediction. Within the individual decision trees, the nodes are split based on a random subset of features. That is to say, random forest does not consider all predictors in this split. This is an advantage over a simple decision tree (Breiman, 2001). Another strength of random forest is that the trees

within the random forest are independent and not correlated, which enhances performance and prevents overfitting. Furthermore, it works well and fast with large datasets (Mohapatra et al., 2020).

In this study, we use the CARET package to train and evaluate the random forest model. Five-fold cross-validation is performed to tune the parameters number of trees, number of selected features and minimum node size. Using five-fold cross-validation is based on a previous study by Ogutu (2011). The model with the optimal parameters, as evaluated by ROC-AUC, is used to retrain the model, run the five-fold cross-validation again and the final ROC-AUC value is used to compare the model's performance against the performance of the other machine learning models. The motivation for the choice of evaluation metric will be discussed in section 3.10.

3.9.3 AdaBoost

Similar to random forests, AdaBoost is a tree-based model. It has been applied in various context and generally considered to make highly accurate predictions (Schapire, 2013). To make these predictions, it applies a prediction rule created by a lot of weak and inaccurate rules. A weighted combination of these rules is used to make a final prediction (Schapire, 2013). A benefit of AdaBoost over other algorithms is that overfitting is less likely due to the parameters not being optimized together. A limitation of AdaBoost is that it needs high quality data and is sensitive to outliers and noise in the data (Thailappan, 2022). In this study, that is not a problem as we do not have a noisy dataset.

In this study, the CARET package is used to train and evaluate the AdaBoost model. Five-fold cross-validation is used to tune the parameters (maximum depth and number of iterations). Using five-fold cross-validation to tune the parameters is supported by a study by Ogutu (2011). Based on the AUC-ROC value, the model with the optimal parameters is used

to retrain the model, run the five-fold cross-validation again and the final ROC-AUC value is used to compare the model's performance against the performance of the other machine learning models.

3.9.4 KNN

K-nearest neighbours is a classification algorithm, which predicts classes based on the *K* most similar observations (Zhang et al., 2017; Guo et al., 2003). There are different ways in which this similarity can be assessed, but the KNN function from the CARET package in R uses Euclidean distance. Euclidean distance is the distance between two points, or in other words, the length of the line between two points (Zhang et al., 2017). A benefit of KNN over more complex algorithms is that it is nonparametric, meaning it does not make assumptions about the data (Zhang et al., 2017). However, it is sensitive to outliers, does not work well with high dimensional data and there is no consensus about what the optimal K-value is (Guo et al., 2003). Despite these limitations, it is still one of the most used algorithms due to its simplicity (Zhang et al., 2017).

In this study, the CARET package is used to train and evaluate the KNN model. As there is no standard best number of nearest neighbours (*k*), five-fold cross-validation is used to choose the best *k* (Guo et al., 2003). Based on the AUC-ROC value, the model with the optimal parameters is used to retrain the model, run the five-fold cross-validation again and the final ROC-AUC value is used to compare the model's performance against the performance of the other machine learning models.

3.10 Evaluation Metric

To optimise the models, compare and evaluate them, we need an evaluation metric. In this study, AUC-ROC was chosen as an evaluation metric. AUC (Area Under the Curve) is a

value that provides a numerical representation of the trade-off between recall and precision. The closer the AUC value is to one, the better the performance of the model. When the AUC value is one, it means that recall and precision are both a hundred percent (Fan et al., 2006). The ROC value shows a graphical representation the AUC value. Recall is plotted against 1 – precision, which creates a curve. This curve is the ROC (Receiver Operating Characteristic) curve (Fan et al., 2006).

The reason for the choice of AUC-ROC as an evaluation metric is that both target classes are considered to be equally important in this study. In this situation, accuracy is also an option. However, AUC-ROC gives a more complex and nuanced view of the performance than accuracy. This is supported by a study by Ling et al. (2003), who found that AUC-ROC is statistically more consistent and discriminating than accuracy.

In this study, each model is optimised based on the AUC-ROC value in the cross-validation phase. The model with the best AUC-ROC value in this phase is chosen as the best performing machine learning model. After the best machine learning model has been selected, the best performing machine learning model and the logistic regression model are compared based on their respective AUC-ROC value evaluated on the test set.

3.11 Feature Importance

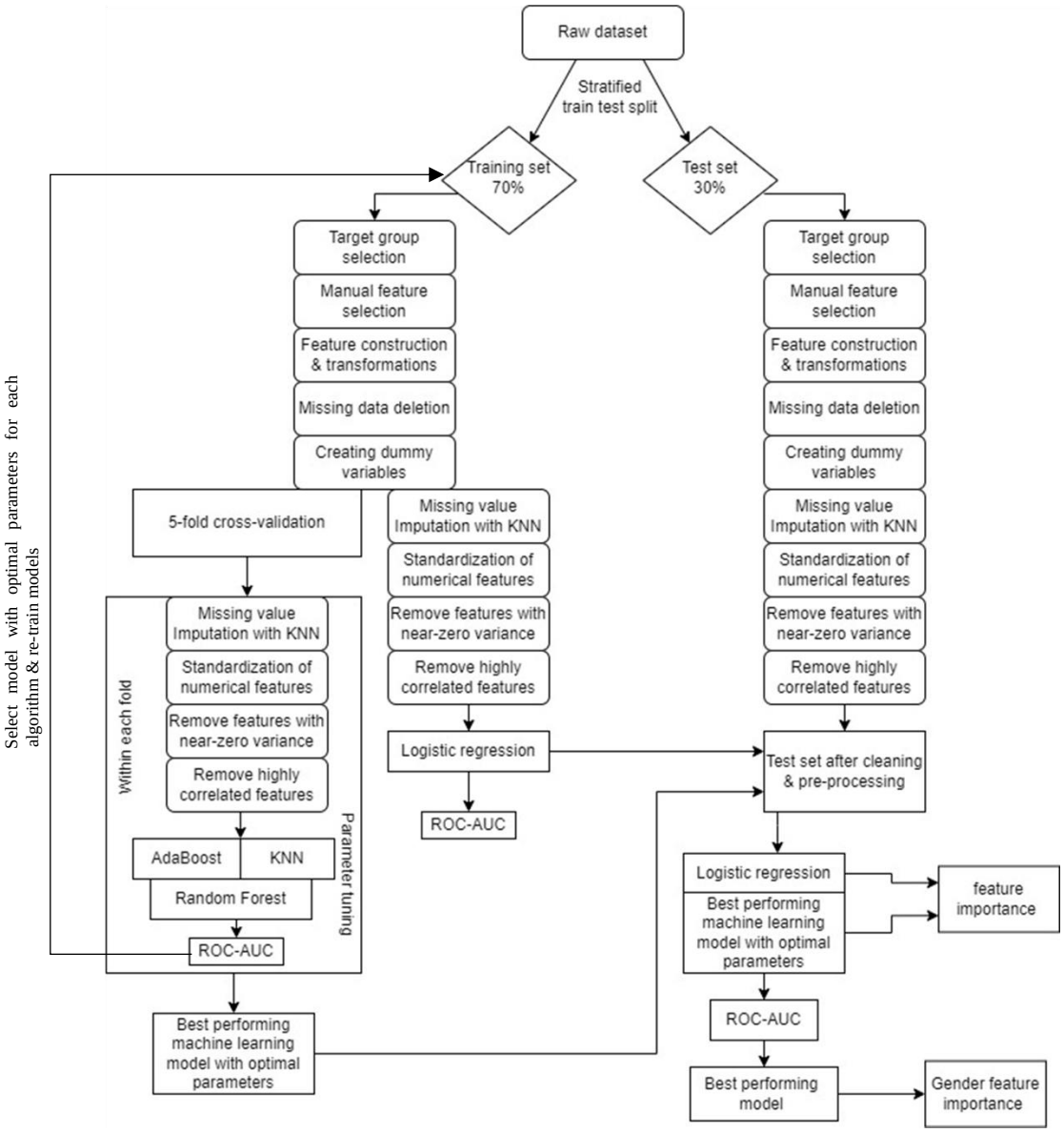
Once it has been decided which machine learning model performs best, we can look at the difference in feature importance of the best performing machine learning model and the baseline model, which is logistic regression. The feature importance is assessed using VarImp function from the CARET package in R.

Similarly, to examine the difference in feature importance between boys and girls, the same function (VarImp) for the CARET package is used. To be able to observe this difference, the best performing model is trained again. This model is trained on two subsets of

the data, one containing only the responses of girls and one containing only the responses of boys. In this way, we obtain a girls-only subset and a boys-only subset. The variable importance is assessed for each of these subsets separately and then compared.

Figure 2

Data science pipeline.



4 Results

This chapter describes the results of the study. It starts with the cross-validation results. After that, the results of each model is described. Lastly, the overall differences in feature importance and gender differences in feature importance are discussed.

4.1 Cross-Validation

The performance represented in the AUC-ROC values across the five-fold cross-validation is presented in table 1. As can be seen in the table, AdaBoost performed best, while logistic regression performed worse. However, the differences between the models are small.

Table 1

Cross-validation performance results.

Model	AUC-ROC
Logistic regression	0.7756
Random forest	0.7999
AdaBoost	0.8217
KNN	0.7930

4.2 Logistic Regression

Logistic regression does not have any parameters to tune and no further decisions regarding this model have to be made. As a result, the model is evaluated on the train and validation data using five-fold cross-validation, and finally evaluated on the test set to assess its out-of-sample performance.

For the five-fold cross-validation, the logistic regression model showed a performance of an AUC-ROC between 0.7423 and 0.7992. The out-of-sample performance on the test set

was slightly worse with an AUC-ROC of 0.7124. Logistic regression performed the worst both in the cross-validation stage and in terms of out-of-sample performance.

4.3 Random Forest

Random forest showed the best performance with 1000 trees, 15 randomly selected predictors, and a minimum node size of 9. The other parameter combinations that were tried did not differ much in performance (between 0.7794 and 0.7980). The model was retrained with the hyperparameters set to the optimal values and the newly trained model showed a AUC-ROC value between 0.7813 and 0.7999. On the test set, the model performed similar with an AUC-ROC value of 0.7951. With this performance, random forest appears to be the second best model. It only performs slightly worse than the best performing model, that is AdaBoost

4.4 AdaBoost

AdaBoost showed the best performance with the maximum depth set to three, the number of iterations set to 150 and the learning rate was kept constant on 0.1. An overview of the combinations of parameters that were tried with their corresponding AUC-ROC value can be found in table 2. The model was retrained with the hyperparameters set to these values and this newly trained model showed a AUC-ROC value between 0.7989 and 0.8421. On the test set, the model performed similar with an AUC-ROC value of 0.8123. As expected, AdaBoost showed the best cross-validated result, as well as the best performance on the test set.

Table 2

Hyperparameter tuning results.

Max depth	Number of iterations	AUC-ROC

1	50	0.7666
1	100	0.7844
1	150	0.7926
2	50	0.7998
2	100	0.8093
2	150	0.8137
3	50	0.8084
3	100	0.8162
3	150	0.8191

4.5 KNN

KNN demonstrated its best performance *k = 10*. The model was retrained with *k* set to 10 and this new model showed a AUC-ROC value between 0.7815 and 0.8030. On the test set, the model performed slightly worse with an AUC-ROC value of 0.7803. Although KNN did not perform best in either the cross-validation stage or in the test stage, it only performed slightly worse than the best performing model, that is AdaBoost.

4.6 Overall Difference in Feature Importance

As per the second sub research question, the difference in feature importance between the best performing machine learning model, that is AdaBoost, and the classic statistical model, that is logistic regression are compared. There are small differences found between the top five most important features. However, the top three most important predictors were the same. These are: *has an account in a bank or other financial institution*, *wealth index* and

frequency of reading a newspaper. An overview of the top five most important predictors for each model can be found in figures 3 and 4.

Figure 3

Relative feature importance for AdaBoost

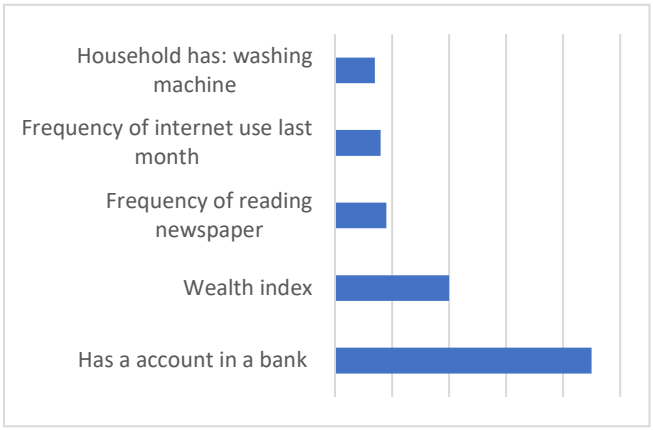
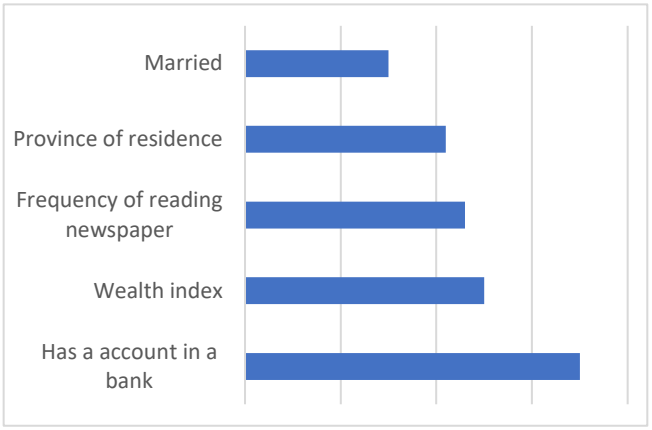


Figure 4

Relative feature importance for logistic regression

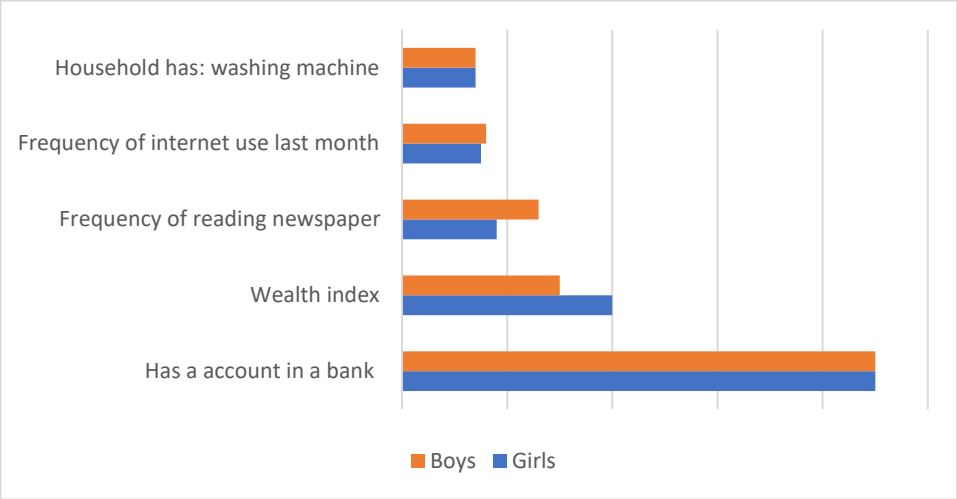


4.7 Difference in Feature Importance for Gender

Regarding differences in feature importance between the two genders, no differences is found. The top five most important predictors were the same for both genders. For both genders the predictor *has an account in a bank or other financial institution* is clearly the most important predictor. *Wealth index* is the second most important, but appears to be relatively more important for girls than for boys, while this is the other way around for the third most important predictor (*frequency of reading a newspaper or magazine).* The top five predictors and their relative importance for each gender can be found in figure 5.

Figure 5

Relative feature importance for boys and girls



5 Discussion

This section discusses the results in relation to its scientific and societal relevance. It starts with an overview of the research goal, research questions and main results with reference to the existing literature. Following this overview, the scientific and societal relevance is discussed. After that, the limitations of the current study are highlighted. Finally, suggestions for future research are given.

5.1 Results & Existing Literature

This study aimed to investigate to what extent demographics, socioeconomic factors and gender are able to predict who will enrol in post-secondary education. To achieve this goal, a main research question and three sub research questions were formulated. With the results of the study, we can answer the main question, which was *To what extent can post-secondary education enrolment be predicted by socioeconomic factors, demographics and gender using machine learning techniques?* This study revealed that it is possible to predict

post-secondary education enrolment using socioeconomic factors, demographics and gender. With the use of three different machine learning models and a classic statistical model, a performance between 0.71 and 0.81 was achieved. A model that can perfectly predict the outcome variable would have a value of 1, meaning that the models we used can predict post-secondary education enrolment relatively well.

Regarding the first research question, *Comparing AdaBoost, random forest and K-nearest neighbours, which model performs best evaluated by AUC-ROC as compared to logistic regression?,* it can be concluded that all machine learning models outperform the baseline model, that is logistic regression. AdaBoost was the overall best performing model, but the difference between the models was small.

Moreover, the study looked into two feature importance comparisons. Firstly, we looked at the difference in feature importance between the best performing machine learning model, that is AdaBoost, and the baseline model, that is logistic regression. Small differences in feature importance were found in the top five most important predictors. For logistic regression, *province of residence* and *married* the fourth and fifth most important predictors, while *frequency of internet use last month* and *household has: a washing machine* were more important in the AdaBoost model. Secondly, the difference in feature importance for boys and girls was investigated. However, no major differences were found.

Overall, the results are in line with what was expected based on previous literature. Prior studies showed good performance for all machine learning models that were studied (Berens et al., 2018; Kemper et al., 2020; Wan Yaacob et al., 2020; Dwi Fajar Maulana et al., 2020; Li et al., 2022; Zulfiker et al., 2021). In some studies machine learning models slightly outperformed logistic regression (Berens et al., 2018; Kemper et al., 2020; Dwi Fajar Maulana et al., 2020; Li et al., 2022; Zulfiker et al., 2021), while in others logistic regression performed slightly better than the machine learning models (Wan Yaacob et al., 2020). The current study contributes to

the literature in which machine learning models slightly outperform logistic regression. The previous literature was divided about which model is the best performing model. Some studies found random forest to perform best (Kemper et al., 2020), while others found AdaBoost to be performing better (Berens et al., 2018; Dwi Fajar Maulana et al., 2020; Zulfiker et al., 2021). The last observation is in line with what the results of this study demonstrate.

Additionally, the literature about the influence of gender on post-secondary education enrolment was inconsistent (Agger et al., 2018; Indrahadi & Wardana, 2020; Richardson et al., 2020; Wenang et al., 2022). This study looked at the difference in feature importance for boys and girls to explain this inconsistency. However, this expected difference was not found. The order of the top five predictors is exactly the same for boys and girls. Country differences might offer an explanation for this. The inconsistent effect of gender on post-secondary education was found in studies considering different countries. It is possible that due to certain cultural norms, boys stay in school longer, while other cultural norms may make girls stay in school longer. Based on the current study, we can conclude that in Indonesia, there is no difference between boys and girls in what factors influence the decision to enrol in post-secondary education.

5.2 Scientific & Societal Contributions

As mentioned before, most findings in this study are in line with the findings of previous studies. Despite this, the current study extends previous research in the domain of post-secondary education prediction by focusing on how to optimize prediction, instead of explaining the relationship between a few predictors and post-secondary education enrolment. Additionally, the use of machine learning models for prediction of post-secondary education is a methodological contribution to the existing literature. Moreover, the result of the second sub research question, which concerned the difference in feature importance between logistic regression and AdaBoost, provides a new perspective on what factors play a roll in post-

secondary education enrolment. Despite there being only a small difference, the features that were found to be of importance, such as owning a bank account or the frequency of reading a newspaper are factors whose effect on post-secondary education enrolment have not been investigated before. Furthermore, although the literature suggested there would be a difference in feature importance between boys and girls, this study found that there is no difference in feature importance between genders in Indonesia. This is an important contribution to the literature, as it suggests that the effect of gender of post-secondary education enrolment may be country-specific.

Not only did the current study contribute new insights scientifically, it also provided relevant results that can be used in practice. The finding that machine learning models can predict secondary education enrolment relatively well can be useful for both secondary schools. Secondary schools can use machine learning to target students that are likely to not enrol in post-secondary education to try and encourage them to continue their education. The study also provides helpful information for policy makers on a higher level, such as the government. Knowing what features are important for prediction can help policy makers develop specific interventions that focus on the most important features. This also has the goal to encourage more students to enrol in post-secondary education. Eventually, this may help Indonesia catch up with the post-secondary enrolment rates of its surrounding countries.

5.3 Limitations

Despite all useful scientific and practical contributions the study provided, there are some limitations in the current study that need to be addressed. Firstly, the lack of a strict method for feature selection is a limitation of this study. Features related to demographics, socioeconomic factors and gender were selected, but there are no specific rules about what makes features related to either of these three concepts. In other words, which features to select is, to a certain

extent, subjective. However, appendix A includes a list of all features that were selected during manual feature selection, which means that is possible to replicate the study. As such, although the lack of strict method for feature selection is a limitation, it should have no effect on the replicability of the study.

Furthermore, the dataset includes many very specific variables, which may measure a bigger construct that is not actually in the dataset. For example, we considered all socioeconomic factors as separate features (e.g. *having a bank account, wealth index, frequency of reading a newspaper).* However, it may be that these features are better represented as one construct, such as socioeconomic status. Although this is something to consider, it should not substantially influence the results of this study as we removed all highly correlated predictors from the model.

5.4 Suggestions for Future Research

Building on to these contributions and limitations, we provide some suggestions for future research. Firstly, the study could be replicated in different countries. it would be interesting to see whether the results of this study are similar in other countries to investigate the effect of culture on post-secondary education enrolment prediction. Moreover, related to the previous point, the effect of gender could be investigated further. We had no success explaining the differences in the effect of gender on post-secondary education enrolment. However, combining the findings from prior studies in different countries with the current study, there seems to be an indication that the effect of gender may be country specific. This could be a starting point for future research. Lastly, future research could look into the use of other machine learning models. Specifically, more complex models such as neural networks. Prediction was already relatively successful with the less complex models used in this study, so it would be interesting to see whether more complex models can do better.

6 Conclusion

The goal of this study was to investigate to what extent demographics, socioeconomic factors and gender are able to predict who will enrol in post-secondary education. This study extends the existing literature by focusing on the prediction of post-secondary education enrolment with machine learning models, rather than explaining the relationship between a small number of features and post-secondary education enrolment with classic statistical models. To achieve the goal of the study, a main research question with three sub questions was formulated. To summarize, the main question with a short answer is stated below.

To what extent can post-secondary education enrolment be predicted by socioeconomic factors, demographics and gender using machine learning techniques?

The study demonstrated that post-secondary education enrolment can be predicted relatively well by various socioeconomic factors, demographics and gender. This is proven by an AUC value between 0.71 and 0.81. AdaBoost performed best with an AUC value of 0.81, but the differences between the models was small. This is in line with previous literature in related areas that also showed a relatively good performance of prediction by all studied machine learning models, with AdaBoost in particular, and small differences between models (Berens et al., 2018; Kemper et al., 2020; Wan Yaacob et al., 2020; Dwi Fajar Maulana et al., 2020; Li et al., 2022; Zulfiker et al., 2021). In previous studies, gender was found to affect post-secondary education enrolment, but findings were inconsistent (Agger et al., 2018; Indrahadi & Wardana, 2020; Richardson et al., 2020; Wenang et al., 2022). This study looked at feature importance in prediction of post-secondary education enrolment for each gender to explain this inconsistency. However, no major differences in feature importance were found.

This has interesting implications for public policy regarding post-secondary education. By using machine learning to predict who will enrol in post-secondary education, secondary

schools and the Indonesian government can develop more targeted interventions to encourage people that are likely to not enrol in post-secondary education to continue their education. As a result, Indonesia has the opportunity to catch up with its surrounding countries in the rising numbers of post-secondary education.

References

Agger, C., Meece, J., & Byun, S. Y. (2018, July 30). The Influences of Family and Place on Rural Adolescents' Educational Aspirations and Post-secondary Enrollment. *Journal of Youth and Adolescence*, *47*(12), 2554–2568. https://doi.org/10.1007/s10964-018-0893-7

Batyra, E., & Pesando, L. M. (2020, July 27). The Selective Impact of Changes in Age-at-Marriage Laws on Early Marriage: Policy Challenges and Implications for Women's Higher-Education Attendance. In *https://repository.upenn.edu/psc_publications/*. University of Pennsylvania Population Center Working Paper (PSC/PARC). Retrieved October 10, 2022, from https://repository.upenn.edu/cgi/viewcontent.cgi?article=1050&context=psc_publications

Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2018). Early Detection of Students at Risk – Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3275433

Bhagat, M., & Bakariya, B. (2022). Implementation of Logistic Regression on Diabetic Dataset using Train-Test-Split, K-Fold and Stratified K-Fold Approach. *National Academy Science Letters*, *45*(5), 401–404. https://doi.org/10.1007/s40009-022-01131-9

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

Chankseliani, M., Gorgodze, S., Janashia, S., & Kurakbayev, K. (2020, May 13). Rural disadvantage in the context of centralised university admissions: a multiple case study of Georgia and Kazakhstan. *Compare: A Journal of Comparative and International Education*, *50*(7), 995–1013. https://doi.org/10.1080/03057925.2020.1761294

Digdowiseiso, K. (2020, February). The Development of Higher Education in Indonesia. *International Journal of Scientific & Technology Research*, *9*(2), 1381–1385. http://repository.unas.ac.id/564/1/The-Development-Of-Higher-Education-In-Indonesia%20%28Kumba%20Feb%202020%29.pdf

Dwi Fajar Maulana, Y., Ruldeviyani, Y., & Indra Sensuse, D. (2020, November 3). Data Mining Classification Approach to Predict The Duration of Contraceptive Use. *2020 Fifth International Conference on Informatics and Computing (ICIC)*. https://doi.org/10.1109/icic50835.2020.9288568

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021, May 11). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, *24*(1), 395–419. https://doi.org/10.1146/annurev-polisci-053119-015921

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN Model-Based Approach in Classification. *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, 986–996. https://doi.org/10.1007/978-3-540-39964-3_62

Harahap, E. S., Maipita, I., & Rahmadana, M. F. (2020, May 8). Determinant Analysis of Education Inequalities in Indonesia. *Budapest International Research and Critics Institute (BIRCI-Journal): Humanities and Social Sciences*, *3*(2), 1067–1082. https://doi.org/10.33258/birci.v3i2.937

Hardy, B. L., & Marcotte, D. E. (2020, October 25). Ties that bind? Family income dynamics and children's post-secondary enrollment and persistence. *Review of Economics of the Household*, *20*(1), 279–303. https://doi.org/10.1007/s11150-020-09516-9

Hossin, M., & Sulaiman, M. N. (2015, March 31). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining &Amp; Knowledge Management Process*, *5*(2), 01–11. https://doi.org/10.5121/ijdkp.2015.5201

Hyseni Duraku, Z., Jemini-Gashi, L., & Toçi, E. (2020, March 2). Perceptions of Early Marriage, Educational Aspirations, and Career Goals among Kosovar Adolescents. *Marriage &Amp; Family Review*, *56*(6), 513–534. https://doi.org/10.1080/01494929.2020.1728006

Indrahadi, D., & Wardana, A. (2020, December). The Impact of Sociodemographic Factors on Academic Achievements among High School Students in Indonesia. *International Journal of Evaluation and Research in Education*, *9*(4), 1114–1120. https://doi.org/10.11591/ijere.v9i4.20572

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020, January 2). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, *10*(1), 28–47. https://doi.org/10.1080/21568235.2020.1718520

Khattab, N. (2018, May 30). Ethnicity and higher education: The role of aspirations, expectations and beliefs in overcoming disadvantage. *Ethnicities*, *18*(4), 457–470. https://doi.org/10.1177/1468796818777545

Khondoker, M., Dobson, R., Skirrow, C., Simmons, A., & Stahl, D. (2016). A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies. *Statistical Methods in Medical Research*, *25*(5), 1804–1823. https://doi.org/10.1177/0962280213502437

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5). https://doi.org/10.18637/jss.v028.i05

Li, Q., Yu, S., Échevin, D., & Fan, M. (2022, June). Is poverty predictable with machine learning? A study of DHS data from Kyrgyzstan. *Socio-Economic Planning Sciences*, *81*, 101195. https://doi.org/10.1016/j.seps.2021.101195

Liao, S. G., Lin, Y., Kang, D. D., Chandra, D., Bon, J., Kaminski, N., Sciurba, F. C., & Tseng, G. C. (2014). Missing value imputation in high-dimensional phenomic data: imputable or not, and how? *BMC Bioinformatics*, *15*(1). https://doi.org/10.1186/s12859-014-0346-6

Ling, C. X., Huang, J., & Zhang, H. (2003). AUC: A Better Measure than Accuracy in Comparing Learning Algorithms. *Advances in Artificial Intelligence*, 329–341. https://doi.org/10.1007/3-540-44886-1_25

Lörz, M., & Mühleck, K. (2018, June 27). Gender differences in higher education from a life course perspective: transitions and social inequality between enrolment and first post-doc position. *Higher Education*, *77*(3), 381–402. https://doi.org/10.1007/s10734-018-0273-y

Ma, J., Pender, M., & Welch, M. (2020, January). Education Pays 2019: The benefits of higher education for individuals and society. In *https://research.collegeboard.org/trends/education-pays* (No. 01469–073). CollegeBoard. Retrieved September 16, 2022, from https://research.collegeboard.org/media/pdf/education-pays-2019-full-report.pdf

Malato, G. (2022, June 14). *Which models require normalized data? - Towards Data Science*. Medium. https://towardsdatascience.com/which-models-require-normalized-data-d85ca3c85388

Mohapatra, N., Shreya, K., & Chinmay, A. (2020). Optimization of the Random Forest Algorithm. *Advances in Data Science and Management*, 201–208. https://doi.org/10.1007/978-981-15-0978-0_19

Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., Prakash, I., & Pham, B. T. (2021). Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil. *Mathematical Problems in Engineering*, *2021*, 1–15. https://doi.org/10.1155/2021/4832864

Niu, L. (2018). A review of the application of logistic regression in educational research: common issues, implications, and suggestions. *Educational Review*, *72*(1), 41–67. https://doi.org/10.1080/00131911.2018.1483892

OECD (2022), *Education at a Glance 2022: OECD Indicators*, OECD Publishing, Paris, https://doi.org/10.1787/3197152b-en.

Ogutu, J. O. (2011, May 27). *A comparison of random forests, boosting and support vector machines for genomic selection - BMC Proceedings*. BioMed Central. https://bmcproc.biomedcentral.com/articles/10.1186/1753-6561-5-S3-S11

Olson-Strom, S., & Rao, N. (2020). Higher Education for Women in Asia. *Diversity and Inclusion in Global Higher Education*, 263–282. https://doi.org/10.1007/978-981-15-1628-3_10

PreProcess function - RDocumentation. (n.d.). Retrieved November 5, 2022, from https://www.rdocumentation.org/packages/caret/versions/6.0-92/topics/preProcess

Richardson, J. T. E., Mittelmeier, J., & Rienties, B. (2020, January 17). The role of gender, social class and ethnicity in participation and academic attainment in UK higher education: an update. *Oxford Review of Education*, *46*(3), 346–362. https://doi.org/10.1080/03054985.2019.1702012

Rumble, L., Peterman, A., Irdiana, N., Triyana, M., & Minnick, E. (2018, March 27). An empirical exploration of female child marriage determinants in Indonesia. *BMC Public Health*, *18*(1). https://doi.org/10.1186/s12889-018-5313-0

Schapire, R. E. (2013). Explaining AdaBoost. *Empirical Inference*, 37–52. https://doi.org/10.1007/978-3-642-41136-6_5

Thailappan, D. (2022, September 19). *AdaBoost : A Brief Introduction to Ensemble learning*. Analytics Vidhya. Retrieved November 20, 2022, from https://www.analyticsvidhya.com/blog/2021/06/adaboost-a-brief-introduction-to-ensemble-learning/

The DHS Program - Data Collection. (n.d.). Retrieved September 30, 2022, from https://dhsprogram.com/data/data-collection.cfm

Toloşi, L., & Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, *27*(14), 1986–1994. https://doi.org/10.1093/bioinformatics/btr300

Vandelannote, I., & Demanet, J. (2021, January 22). Unravelling socioeconomic school composition effects on higher education enrollment: the role of students' individual and shared feelings of futility and self-efficacy. *Social Psychology of Education*, *24*(1), 169–193. https://doi.org/10.1007/s11218-021-09608-z

Vrigazova, B. (2021). The Proportion for Splitting Data into Training and Test Set for the Bootstrap in Classification Problems. *Business Systems Research Journal*, *12*(1), 228–242. https://doi.org/10.2478/bsrj-2021-0015

Wan, C. D. (2017, April 10). Student enrolment in Malaysian higher education: is there gender disparity and what can we learn from the disparity? *Compare: A Journal of Comparative and International Education*, *48*(2), 244–261. https://doi.org/10.1080/03057925.2017.1306435

Wan Yaacob, W. F., Mohd Sobri, N., Nasir, S. A. M., Norshahidi, N. D., & Wan Husin, W. Z. (2020, March). Predicting Student Drop-Out in Higher Institution Using Data Mining Techniques. *Journal of Physics: Conference Series*, *1496*, 012005. https://doi.org/10.1088/1742-6596/1496/1/012005

Wenang, S., Nahdiyati, D., T., A., & Rismawati, I. (2022). The Understanding of Gender Equality in Indonesia by Indonesian Women's Diaspora in Germany. *Prosiding International Conference on Sustainable Innovation (ICoSI)*, *1*(1), 31–36. https://doi.org/10.18196/icosi.v3i1.31

Who We Are. (n.d.). The DHS Program. Retrieved September 23, 2022, from https://dhsprogram.com/Who-We-Are/About-Us.cfm

Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning for kNN Classification. *ACM Transactions on Intelligent Systems and Technology*, *8*(3), 1–19. https://doi.org/10.1145/2990508

Zulfiker, M. S., Kabir, N., Biswas, A. A., Nazneen, T., & Uddin, M. S. (2021, November). An in-depth analysis of machine learning approaches to predict depression. *Current Research in Behavioral Sciences*, *2*, 100044. https://doi.org/10.1016/j.crbeha.2021.100044

Appendix A

Table 3

Variable selected by manual feature selection

Variable	Variable description
CASEID	Case identification
V012	Respondent's current age
V024	Province
V025	Type of place of residence
V100	Participant's sex
V103	Childhood place of residence
V104	Years lived in place of residence
V113	Source of drinking water
V115	Time to get to water source
V116	Type of toilet facility
V119	Household has: electricity
V120	Household has: radio
V121	Household has: television
V122	Household has: refrigerator
V123	Household has: bicycle
V124	Household has: motorcycle/scooter
V125	Household has: car/truck
V127	Main floor material
V128	Main wall material
V129	Main roof material
V130	Religion
V131	Ethnicity

V149	Educational attainment
V151	Sex of household head
V152	Age of household head
V153	Household has: telephone (land-line)
V157	Frequency of reading newspaper or magazine
V158	Frequency of listening to radio
V159	Frequency of watching television
V161	Type of cooking fuel
V169A	Owns a mobile telephone
V170	Has an account in a bank or other financial institution
V171A	Use of internet
V171B	Frequency of using internet last month
V190	Wealth index
V212	Age of participant when first child was born
V221	Time between marriage and birth of first child (in months)
V511	Age of first cohabitation
V613	Ideal number of children
S109H	Type of toilet facility
S121G	Household has: a fan
S121H	Household has: a washing machine
S121G	Household has: an air conditioner