A machine learning modeling prediction of enrollment among admitted college applicants at University of Santo Tomas

Cite as: AIP Conference Proceedings **2472**, 020004 (2022); https://doi.org/10.1063/5.0100174
Published Online: 19 August 2022

Arturo J. Patungan and Mari Loren M. Francia

View Online

Export Citation

AIP Conference Proceedings **2472**, 020004 (2022); https://doi.org/10.1063/5.0100174 **2472**, 020004

A Machine Learning Modeling Prediction of Enrollment among Admitted College Applicants at University of Santo Tomas

Arturo J. Patungan, Jr. [1, a)] and Mari Loren M. Francia [1, b)]

[1]*Department of Mathematics and Physics, College of Science, University of Santo Tomas, Philippines*

[a)] Corresponding author: ajptaungan@ust.edu.ph
[b)] mariloren.francia.sci@ust.edu.ph

Abstract. Predicting the enrollment has become a critical part of institutional planning processes in higher education. The forecast of enrolled number of students annually, represents a very important role, because the foundation of the budget and expenditures is based on the number of students enrollees. Data mining enables organizations to use their current reporting capabilities to uncover and understand hidden patterns in vast databases. These patterns are then built into data mining models and used to predict individual behavior with high accuracy. The University of Santo Tomas Entrance Test (USTET) and College Application Form is part of the admission requirements of the University. Every year, a new batch of examinees takes their chance in passing USTET. However, there are times that USTET passers do not enroll. The dataset used in this study was gathered from the UST Office for Admission and UST Office of the Secretary-General. In order to predict the enrollment at University of Santo Tomas, the Admission details of students in the past five academic years (A.Y.2015-2016 up to A.Y.2018-2019) were used. The attributes in the Application Form with no data mining value to classify the admitted applicant's enrollment behavior were ignored and the attributes with limited data gathered were not considered. The data set contains twenty-four (24) main attributes. Twenty-three (23) of those are the input variables, and one output variable for label, which has two categories: ENROLLED and NOT ENROLLED. Four data mining algorithms namely Artificial Neural Network, Bayesian Network, Decision Tree, and Logistic Regression were implemented and tested for efficiency using the software RapidMiner Studio Version 9.0.003. The performances of the algorithms were evaluated using the following parameters: accuracy, sensitivity, specificity, precision, negative predictive value. The most efficient model produced was determined by comparing the performance parameters. Thus, the results showed that Naïve Bayes have the highest accuracy in the validation or holdout dataset, which is 99.98%.

INTRODUCTION

One of the biggest challenges that higher educational institutions face nowadays is predicting the paths of students [1]. Institutions would like to know which among the students and how many students will enroll in particular course programs [2]. With the rapid opening of educational institutions, admissions in any universities and educational institutions are likely to face imminent admission crisis nowadays [3]. To have good quality of students and to achieve desired number of students, the institutes have some standard admission procedures like entrance test, interviews, or group discussions but some of the institutes allow direct admission also. In addition to this challenge, is the enrollment management, which continues to be a challenge that motivates higher education institutions to search for better solutions [4]. One way to effectively address these challenges is through the data mining [5]. Data mining is applied to determine set of patterns of students seeking admission in college collected over years and use those set of patterns for future prediction [6],[7]. Thus, this study will be focused on comparing the effectiveness of four data mining algorithms in the prediction of enrollment behaviors [8] of admitted applicants at University of Santo Tomas.

METHODOLOGY

Broadly, there are three types of Machine Learning – Supervised Learning, Unsupervised Learning, and Reinforcement Learning. In this study, Supervised Learning type of machine learning is used. The Supervised Learning is an algorithm consists of a target variable or outcome variable (dependent variable), which is to be predicted from a given set of predictors (independent variables). This study is a quantitative research; it emphasizes objective measurements and the statistical, mathematical, or numerical analysis of data collected using data mining techniques to formulate facts and

The 5th Innovation and Analytics Conference & Exhibition (IACE 2021)
AIP Conf. Proc. 2472, 020004-1–020004-4; https://doi.org/10.1063/5.0100174
Published by AIP Publishing. 978-0-7354-4387-7/$30.00

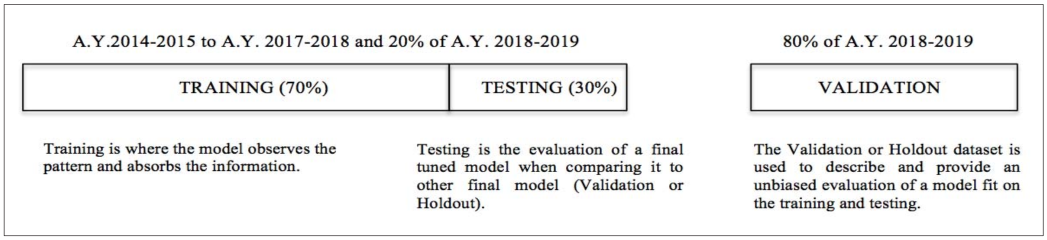
uncover patterns in research.

The mathematical models used in this study are models that forecast whether a University of Santo Tomas Examination Test (USTET) passer will enroll or not. The model will take the applicant's attributes into consideration in the prediction of enrollment. The attributes that will be used in this research were chosen from the USTET College Application Form. Some of the attributes in the Application Form that was not pertinent (relevant) to the data mining experiment goal were ignored. The attributes name, application number, address, date of birth, and contact number were not used as having no data mining value to classify the admitted applicant's enrollment behavior.

The data set contained twenty-four (24) attributes and twenty-three (23) of those are the input variables, in which three (3) are numerical and nineteen (19) are nominal, and one output variable for class which has two categories: ENROLLED and NOT ENROLLED. This attribute included the gender of the applicants, region of the school where he graduated, if he graduated in UST or not, Is the school a private or public school, his religion, its nationality, the program where he applied for as a priority program, alternative, and the one he enrolled in, his priority program status, his alternative program status, if he is a varsity player, the applicants family monthly income, if he is a member of any cultural, civic or church organization, a member of fraternities/sororities, if the applicant have a family member that are graduates or student of UST and a family member that are employed in UST, if he planned to pursue further studies, in need of financial help, applied in other schools, his USTET result in English, Mathematics, and Science, and finally if he enrolled in UST

The USTET admitted applicants and enrollees in the past five academic years were used. The dataset of A.Y.2014-2015 up to A.Y.2017-2018 and 20% of A.Y. 2018-2019 will be used for training and testing. The remaining 80% of A.Y.2018-2019 will be used for validation of the effectiveness of the model. The data of A.Y. 2018-2019 is divided into 20% and 80%, which was randomly selected. About 20% of A.Y. 2018-2019 was put in training and testing set due to the limited amount of data in some programs. There were new programs in A.Y. 2018-2019, which needed to be in the training and testing in order to make predictions for the validation set, which is the 80% of A.Y.2017-2018. The distribution of those who enrolled and not enrolled in the dataset cannot be divulged because it is one of the restrictions of a confidentiality agreement that the researchers agreed to the university when they were allowed to use the data for the study but were considered in the process.

The model used a split ratio of 70% and 30% of the training and testing dataset, the model was then iteratively trained and validated on these different sets. Overfitting led to poor prediction, to avoid over fitting the split validation was used, this is basically using the training set to generate multiple splits of train and testing sets. Figure 1 below shows split validation of dataset.



Training is where the model observes the pattern and absorbs the information.

Testing is the evaluation of a final tuned model when comparing it to other final model (Validation or Holdout).

The Validation or Holdout dataset is used to describe and provide an unbiased evaluation of a model fit on the training and testing.

FIGURE 1. Split Validation of DataSet

Four data mining algorithms- Artificial Neural Network, Naïve Bayes, Decision Tree, Logistic Regression are used in the study in order to satisfy the specification and conditions of the most effective algorithm in the prediction of enrollment and were implemented and tested for efficiency using the software RapidMiner Studio Version 9.0.003. These four classification models were the most common models used and is readily available in Rapidminer software. A model for each technique will be constructed and tested for efficiency using the performance parameters based on the Confusion Matrix – accuracy, sensitivity, specificity, and precision

DISCUSSION OF RESULTS

This study consisted personal records in the database. Thus, this study followed a confidentiality clause. This confidentiality clause restricted the classified data to be exposed or stated on the study. Thus, only the statistical summaries were presented and interpreted.

The four models used the following parameters in Rapidminer and yielded to the following models. The Naïve Bayes with no laplace correction. The Decision Tree had an accuracy as criterion, 10 as its maximal depth, with prepruning and pruning application, a minimal gain of 0.01, minimal leaf size of 2 and minimal split size of 3. The Logistic Regression used auto solver, standardization of values, removes multicollinearity among attributes and using mean imputation for missing data. The Artificial Neural Net uses 2 hidden layers, 3 training cycles, 0.05 learning rates, and 0.5 momentum. It also applies decay and normalization.

Table 1 and Table 2 show the accuracy, sensitivity, specificity, and precision of the Training and Testing Dataset Performance, and the Validation or Holdout Dataset Performance of each model.

TABLE 1. Summary of the Model Performance in the Training and Testing Data Set

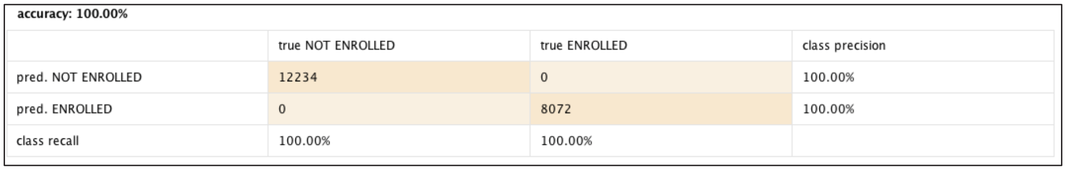
Classification Models	Accuracy	Sensitivity	Specificity	Precision
Artificial Neural Net	96.14%	91.30%	99.33%	98.90%
Decision Tree	68.37%	41.60%	86.03%	66.27%
Logistic Regression	97.46%	97.51%	97.43%	96.16%
Naïve Bayes	100.00%	100.00%	100.00%	100.00%

TABLE 2. Summary of the Model Performance in the Validation or Holdout Data Set

Classification Models	Accuracy	Sensitivity	Specificity	Precision
Artificial Neural Net	95.33%	89.66%	98.88%	98.04%
Decision Tree	76.65%	69.71%	80.98%	69.63%
Logistic Regression	96.90%	96.89%	96.91%	95.14%
Naïve Bayes	99.97%	99.94%	100.00%	100.00%

The Naïve Bayes showed the highest performance in the training and testing dataset, and validation or holdouts dataset among the four data mining models.

Figure 2 which shows summary of prediction of enrollment of admitted applicants in which, 8,072 students who enrolled, were correctly predicted as enrolled (True Positive), and 12,234 students who did not enroll, were correctly predicted as not enrolled (True Negative). There are no students who were incorrectly predicted. Using the Confusion Matrix, the accuracy, class precision and class recall were computed. Table 8 shows that the testing dataset performances in Bayesian. All the performance of Bayes showed 100.00%.

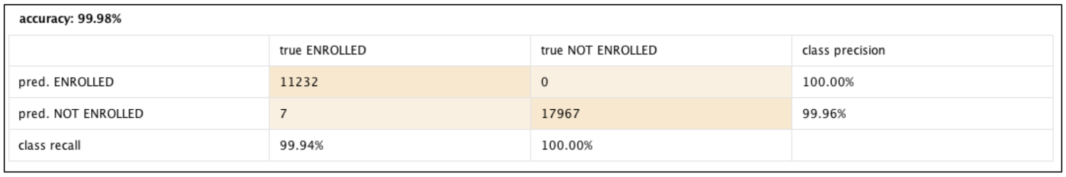


accuracy: 100.00%

	true NOT ENROLLED	true ENROLLED	class precision
pred. NOT ENROLLED	12234	0	100.00%
pred. ENROLLED	0	8072	100.00%
class recall	100.00%	100.00%	

FIGURE 2. Naïve Bayes Training and Testing Dataset Performance and Confusion Matrix

The Figure 3 shows summary of prediction of enrollment of admitted applicants in which, 11,232 students who enrolled, were correctly predicted as enrolled (True Positive), and 17,967 students who did not enroll, were correctly predicted as not enrolled (True Negative). However, there were 7 students who did not enroll were incorrectly predicted as enrolled (False Negative), and there are no students who enrolled were incorrectly predicted as not enrolled (False Positive). Using the Confusion Matrix, the accuracy, class precision and class recall were computed. Table 9 reveals that the testing dataset performances in Bayesian. The accuracy rate was 99.98%, which is the ratio of correctly predicted observations over the total observations. The Sensitivity (True Positive Rate) in validation was 99.94%, which is the proportion of the actual enrollees that were correctly predicted to be enrolled. Both the Specificity (True Negative Rate) and Precision (Positive Predictive Value) in testing is 100.00%. Lastly, the Negative Predictive Value in validation is 99.96%, which is the ratio of correctly predicted not enrollees (true negatives) to the total predicted negative observations.



accuracy: 99.98%

	true ENROLLED	true NOT ENROLLED	class precision
pred. ENROLLED	11232	0	100.00%
pred. NOT ENROLLED	7	17967	99.96%
class recall	99.94%	100.00%	

FIGURE 3. Naïve Bayes Validation or Holdout Dataset Performance and Confusion Matrix

CONCLUSION

This paper is focused on the prediction of enrollment behavior at University of Santo Tomas. Understanding of data mining techniques is important to find which technique could give the best result and could contribute informative knowledge. Thus, each technique has its own strengths and weaknesses. Comparing the results of the four data mining techniques, Naïve Bayes gave the highest accuracy with 99.98%, Sensitivity of 99.94% and a Specificity and Precision of 100% in the Validation or Holdout Dataset. Therefore, Naïve Bayes is the most efficient model produced for the prediction of enrollment. Thus, the produced algorithm is suggested to the University. However, Artificial Neural Network, and Logistic Regression, and Decision Tree can still be an efficient model since their parameters can be changed and the accuracy rate are very close to one another. Hence, these techniques are also highly recommended and widely used in data mining.

REFERENCES

1. G. B. Tarekegn and V. Sreenivasarao, International Journal of Research Studies in Computer Science and Engineering, **3**(2), 10 (2016)
2. N. Abu Haris, M. Abdullah, A. Othman and F. Rahman, “Admission Management through Data Mining using WEKA”, in *Knowledge Management International Conference (KMICe)*. (University of Kuala Lumpur, 2014).
3. S. Singh and V. Kumar, International Journal of Computer Science and Network (IJCSN). **1** (4), 121(2012)
4. B. Nakhkob and M. Khademi, Journal of Advances in Computer Research, **7**, 125 (2015).
5. J. Luan, New Directions for Institutional Research, **2002** (113), 36 (2002). https://doi.org/10.1002/ir.35
6. L. Chang, New Directions for Institutional Research. 131, 53 (2006).
7. R. K. Arora and D. Badal, International Journal of Advanced Research in Computer Science and Software Engineering. **3**(10), 674 (2013).
8. D. Doreswamy and K. Hemanth, Artificial Intelligent Systems and Machine Learning, **3**(3), 162 (2011).