ISSN 2278-3091
Volume 9 No.2, March - April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse179922020.pdf
https://doi.org/10.30534/ijatcse/2020/179922020

Higher Education Institution (HEI) Enrollment Forecasting using Data Mining Technique

Adeline P. Dela Cruz[1], Ma. Leslie B. Basallo[2], Benjamin A. Bere, III[3], Jerome B. Aguilar[4], Cheneta Kenny P. Calvo[5], Jan Carlo T. Arroyo[6], Allemar Jhone P. Delima[7]
[1-5,7]College of Engineering, Technology, and Management, Cebu Technological University-Barili Campus, Cebu, Philippines
[6]College of Computing Education, University of Mindanao, Davao City, Davao del Sur, Philippines
adeline.purisima.delacruz@gmail.com[1], basallomaleslie@gmail.com[2], benjie2x2@gmail.com[3], jaconfig@yahoo.com[4], keken517@gmail.com[5], jancarlo_arroyo@umindanao.edu.ph[6], allemarjpdjca@yahoo.com[7]

ABSTRACT

Prediction plays a vital role used for strategic and tactical decision-making undertaking sthat pave the way for efficient and effective management. Prediction is beneficial in HEI mining in its continued quest to study on the historical, current, and the continuance data relationships for particular situations from an educational context. This paper employed the famous ARIMA(p,d,q) model in forecasting HEI general student enrollment count for S.Y. 2019-2020 to S.Y. 2024-2025 using the university's overall enrollment data for S.Y. 2011-2012 to 2018-2019. Different p,d,q values were tested, and the model with the lowest Akaike Information Criterion (AIC) value was used for prediction. The simulation result showed that ARIMA(0,2,1) model appeared to be the statistically appropriate model to forecast enrollment in the university. The forecast showed an increasing trend in enrollment for the succeeding school years. Future researchers may utilize other data mining algorithms and consider the specific prediction of enrollment counts per colleges for better enrollment trend analysis and knowledge extraction.

Key words: ARIMA algorithm, data mining, forecasting, HEI, prediction

1. INTRODUCTION

Data mining (DM) or knowledge discovery (KD) is the process of extracting implicit information or knowledge from databases that are drawn from the field of statistics, which uses mathematical and machine learning techniques and algorithms[1]. The application of DM or KD area is essentially dependent on the problem the researcher sought to answer. The Higher Educational Institution (HEI), being the area where DM or KD is commonly applied, is greatly concerned with the student's enrollment data to look for patterns and possible influences on student's decision to attend their institution. Prediction, as one of the commonly used data mining techniques in the literature, is considered as a practical method for HEI management in generating knowledge to be used for decision making [2].

Prediction plays an essential role in HEI management as it draws inferences relevant for decision and policy-making undertakings. The administration and the researchers have the leeway to apply prediction on various problems with a variety of complexity. Prediction, when used for student enrollment, primarily aids understanding for enrollment trend analysis, which constitutes knowledge on the significant impact of enrollment for revenue outcomes. The generated information could influence future strategy and resource decisions [3]. Understanding patterns, association, changes, significant structures, and anomaly detection are some of the benefits perceived after knowledge extraction.

Student enrollment prediction is beneficial for HEI as the knowledge generated could leverage on the use of optimal decision-making strategies needed for future planning. However, the selection of the correct prediction methods for a particular problem is still a quest since data mining algorithms to be used are dependent on the availability of the HEI data. In this paper, the famous ARIMA(p,d,q) algorithm which is a type of time series analysis model is implemented since the dataset used is a univariate historical data of enrollees from the Cebu Technological University-Barili Campus, Philippines, from the school years 2011-2012 to 2018-2019, obtained from the University's Office of the Registrar. Different ARIMA(p,d,q) models were tested, and the optimal model to be used for forecasting was selected from it.

The purpose of this study is to provide a 6-year forecast on the number of future enrollment on the abovementioned university satellite campus, specifically for the school years 2019-2020 to 2024-2025. The end result of this study can provide the following benefits to wit: strengthen the existing admission and retention policies of the campus, make central decisions on the university's long term enrollment management strategy, develop an annual marketing and recruitment plan, and to determine internal and external factors affecting the drop and increase of enrollment, among others. The projected enrollment can also serve as a basis for proposing additional classroom buildings to meet increased demand. Generally, this study is hoped to contribute to the two major literatures; (1) on the use of the ARIMA algorithm

and (2) on the literature of HEI mining.

2. RELATED LITERATURE

A time series model approach was used in [4] to predict student enrollment in basic public schools in Ghana. Fifty-four data points were generated from the 1961 to 2014 enrollment dataset provided by Ghana's Ministry of Education. Findings showed that ARIMA(0,2,2) model identified using AIC forecasted increase in enrollment within the next five years, with a gradual decrease after each year thereof.

Additionally, [5] described an aggregated enrollment prediction approach through support-vector machine and rule-based predictive models. The initial predictive results are generated by the SVM, which are then fed to a tool called Cubist [6], which produces the rule-based predictive model. Results of the study show that SVM and Cubist made fairly accurate predictions with MAPE ranged from 0% - 11% for the SVM model and 2.13% to 15.59% with an average error of 5.5% for the Cubist model.

Data mining techniques were also used in [7] to predict student enrollment using Apriori and Naïve Bayes Algorithm. The study proposed a system based on data mining techniques to recommend an academic track for students to take. The study used two prediction models: branch prediction and stream analysis. The Naïve Bayes algorithm was used to the most suitable track based on a student's answer on a questionnaire. On the other hand, the Apriori Algorithm was used to provide suggestions to alternative tracks that students could take in case they would not opt for the recommended track by the Naïve Bayes algorithm.

In [8], a Seasonal Autoregressive Integrated Moving Average (SARIMA) model was used to project enrollment of international students at Midwest University. The model is represented as SARIMA(p,d,q) x $(P,D,Q)_s$where p is the autoregressive terms, d is the differences, q is the moving average terms, P is the seasonal autocorrelation, D is the seasonal trend, Q is the seasonal moving average and s is seasonal period. Through this model, seasonality was captured through the seasonal enrollment pattern of international undergraduate students by semester, using a limited fifteen to twenty years of data. The study was able to establish a forecasting model using several crucial indicators, particularly visa policies, the rapid increase in Chinese enrollment, and tuition rate.

The paper [9] captured enrollment prediction of student applicants at a cohort level for the University of New Mexico. The dataset used for the prediction was gathered from 2003 to 2016 through the seasonal enrollment periods of spring, summer, and fall semesters. AIC was used to identify the best ARIMA model for the series. The SARIMA(0,0,0)x$(1,0,3)_3$ model was selected as it has the lowest AIC. It was used to predict the enrollment at UNM for the year 2017, resulting in a reliable accuracy at 80% confidence interval.

On the other hand, a five-year student enrollment forecast was done for the Bolgtanga Polytechnic in West Africa. The dataset utilized for the prediction spanned from the years 2004 to 2018. Forecasting validation tools like AIC, BIC, and HQ were employed to identify the best prediction model. Upon assessment, it has been identified that the best model to be used is ARIMA (1,0,0), as it has the minimum values. The results of the study indicated an insufficient increase in students' enrollment over time [10].

The study [11] utilized the ARIMA model to predict student admissions at the University of Lagos in Nigeria. Data from undergraduate admissions from 1962 to 2016 were collected and analyzed. The selection of the best ARIMA model has been made using information criteria such as AIC, SBC, AME, RMSE, and MAPE through splitting data into estimation period and validation period. Findings showed a continual increase in student population annually.

3. METHODOLOGY

3.1 Dataset

The datasets used in this paper are the historical data of the overall enrollment count from Cebu Technological University-Barili Campus, Philippines, constituting the general enrollment count of all the colleges from the university satellite campus for S.Y's 2011-2012, 2012-2013, 2013-2014, 2014-2015, 2015-2016, 2016-2017, 2017-2018, and 2018-2019. The datasets were obtained from the university's Office of the Registrar.

3.2 ARIMA Algorithm

The ARIMA(p,d,q) model is used in time series forecasting. The p variable denotes the autoregressive order, while dis for differenced t times whereas d represents the moving average order. The ARIMA(p,d,q) model is shown in equation (1) below.

$$\emptyset(B)(w_t - \mu) = \theta(B)a_t \quad (1)$$

where t,is represented as time index and backshift operator for symbol B, autoregressive parameter assigned as$\emptyset(B)$, $\theta(B)$ for MA, w_t for d value in the ARIMA(p,d,q) model and$a_t$for white noise[12], [13].

3.3 Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)

The basis of ARIMA p and q assignment is determined through its ACF and PACF plot. The equation is stressed as:

$$P_k = \frac{\sum_{t=k+1}^{r}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{r}(Y_t - \bar{Y})^2} \quad (2)$$

where P_k denotes the ACF coefficient in lag k, and the observed period is expressed as t, while observations in period t is denoted by Y_t. The$\bar{Y}$denotes as the mean, and the observation in t-k is expressed as Y_{t-k} [12], [13]. The autoregressive (p) order is represented in the Partial Autocorrelation Function (PACF) plot while the Autocorrelation Function (ACF) plot denotes the moving average (q) of the model.

3.4 Akaike Information Criterion (AIC)

In various papers where ARIMA modeling is observed, the most appropriate model to be used to forecast is the candidate model with the lowest AIC value. The AIC is expressed as equation (3) below:

$$AIC = 1n\frac{\sum_{i=1}^{t}\hat{\varepsilon}_i^2}{T-n} + \frac{2n}{T} \quad (3)$$

where $\hat{\varepsilon}_i^2$ denotes squared residual estimates, T for observation size within samples, and n for the estimated parameters [14], [13].

4. RESULTS AND DISCUSSION

4.1 Graphical and Statistical Methods

Fig. 1 shows the time series plot for CTU-Barili's enrollment rate from S.Y. 2011-2012 to 2018-2019. An upward trend is evident in the graph, which denotes an increasing number of enrollees over the years.

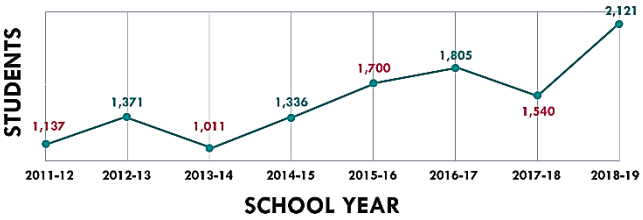

Figure 1: Time series plot for CTU-Barili's enrollment rate from S.Y. 2011-2012 to 2018-2019

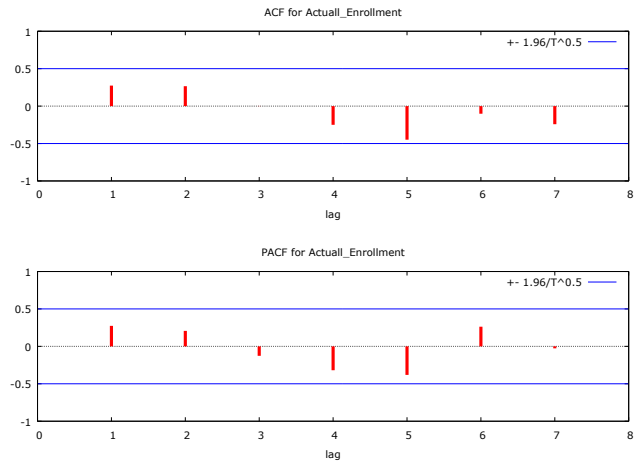

Figure 2: Correlogram plot of the dataset

Figure 2 shows the ACF and PACF correlogram plot for (MA)(q) and (AR)(p) value for lags 1 to 8. The ACF and PACF showeda decaying pattern throughout the lags. The ACF is decaying, while the PACF is also decaying but not in an abrupt manner. This denotes a zero value for autoregressive (p) and a process for the moving average (q). Meanwhile, a trend is evident in Figure 1, making the data non-stationary, therefore adding value in d in the ARIMA(p,d,q) model.

The optimal p,d,q model is chosen based on the model with the lowest Akaike Information Criterion (AIC),as shown in Tables 1 and 2.

Table 1: ARIMA (0,1,*) selection using AIC

ARIMA Model	AIC
****0,1,1****	**102.1462**
0,1,2	102.0872
0,1,3	104.0393
0,1,4	104.4132
0,1,5	106.1677

Table 2: ARIMA (0,2,*) selection using AIC

ARIMA Model	AIC
****0,2,1****	**94.68104**
0,2,2	95.22491
0,2,3	97.49976
0,2,4	100.3610

Assigning 1 to d denotes the adoptive process level while assigning 2 to d denotes adoptive trend in addition to the level of the process. The ARIMA(0,1,1) has the lowest AIC value considering d=1 while ARIMA(0,2,1) has the lowest AIC value considering d=2 as evident in Tables 1 and 2.

Table 3: Summary of ARIMA(p,d,q) Model

ARIMA Model	AIC
0,1,1	102.1462
**** 0,2,1 ****	**94.68104**

Table 3 shows that ARIMA(0,2,1) model appeared to be the statistically appropriate model to forecast the enrollment rate of the CTU-Barili Campus for S.Y. 2019-2020 to 2024-2025. The model established the lowest AIC value and is optimal for prediction.

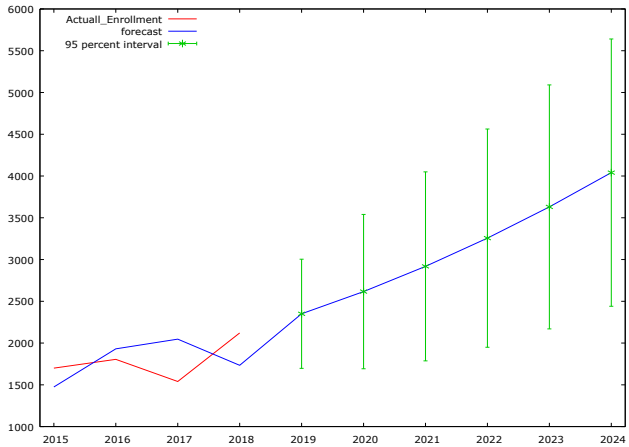

Figure 3: The forecasted number of enrollees for the succeeding school years

Figure 3 shows the graphical representation of the forecasted enrollment data of the CTU-Barili Campus, with its 95% confidence interval using the ARIMA(0,2,1) model. An increasing trend is projected from S.Y. 2019-2020 onwards with the specific forecasted values for S.Y. 2019-2020 to 2024-2025, as shown in Table 4.

Table 4: Forecasted enrollment number with a 95% confidence interval

School Year	Forecast	Lo 95	Hi 95
2019-2020	2,350	1,697	3,003
2020-2021	2,616	1,692	3,540
2021-2022	2,918	1,786	4,049

2022-2023	3,256	1,949	4,562
2023-2024	3,630	2,169	5,090
2024-2025	4,040	2,440	5,640

5. CONCLUSION AND RECOMMENDATION

In this paper, the optimal ARIMA(p,d,q) model was identified to forecast the future enrollment trend at Cebu Technological University-Barili Campus, Philippines, for the school years 2019-2020 to 2024-2025. The ACF, time difference and the PACF process were identified in order to come up with an appropriate model for prediction. The selection of the model with the lowest AIC value was observed in order to come up with the best model. The simulation results showed that ARIMA(0,2,1) model was identified as the best ARIMA(p,d,q) model to forecast enrollment in the HEI. The forecast showed an increasing trend in enrollment for the succeeding school years. Future researchers may consider predicting specific enrollment counts per colleges for a better understanding of enrollment trend analysis and knowledge extraction.

It is recommended that the use of data mining techniques and algorithms [15]–[27] be observed for further knowledge extraction.

REFERENCES

[1] A. J. P. Delima, A. M. Sison, and R. P. Medina, "Variable Reduction-based Prediction through Modified Genetic Algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 5, pp. 356–363, 2019. https://doi.org/10.14569/IJACSA.2019.0100544

[2] N. A. Haris, M. Abdullah, N. Hasim, and F. A. Rahman, "A study on students enrollment prediction using data mining," in *10th International Conference on Ubiquitous Information Management and Communication*, 2016.

[3] J. Ward, "Forecasting enrollment to achieve institutional goals," *Coll. Univ. Journals*, pp. 41–46, 2007.

[4] I. NWI-MOZU, S. K. ASIEDU-ADDO, A. R. OPPONG, and M. ALI, "Primary One Enrolment in Public Basic Schools in Ghana Using Time Series Forecasting," *J. Innov. Educ. Africa*, vol. 1, no. 2, pp. 29–39, 2017.

[5] S. S. Aksenova, D. Zhang, and M. Lu, "Enrollment Prediction through Data Mining," *IEEE Int. Conf. Inf. Reuse Integr.*, 2006.

[6] "RuleQuest Research." .

[7] H. Sabnani, M. More, P. Kudale, and S. Janrao, "Prediction of Student Enrolment Using Data Mining Techniques," *Int. Res. J. Eng. Technol.*, vol. 5, no. 4, pp. 1830–1833, 2018.

[8] Y. Chen, R. Li, and L. S. Hagedorn, "Undergraduate International Student Enrollment Forecasting Model : An Application of Time Series Analysis," *J. Int. Students*, vol. 9, no. 1, pp. 242–261, 2019.

[9] A. Slim, D. Hush, T. Ojah, and T. Babbitt, "Predicting Student Enrollment Based on Student and College Characteristics," *11th Int. Conf. Educ. Data Min.*, pp. 383–389, 2018.

[10] I. D. Yakubu and J. A. Awaab, "Assessing Students ' Enrolment in Bolgatanga Polytechnic Using Time Series Analysis," *East African Sch. J. Eng. Comput. Sci.*, vol. 2, no. 4, pp. 120–135, 2019.

[11] J. N. Onyeka-Ubaka, S. O. N. Agwuegbo, and O. Abass, "Application of the ARIMA Models for Predicting Students ' Admissions in the University of Lagos," *J. Sci. Res.*, vol. 17, no. 1, pp. 80–90, 2017.

[12] J. D. Urrutia, F. L. T. Mingo, and C. N. M. Balmaceda, "Forecasting Income Tax Revenue of the Philippines Using Autoregressive Integrated Moving Average (Arima) Modeling : a Time Series Analysis," *Am. Res. Thoughts*, vol. 1, no. 9, pp. 1938–1992, 2015.

[13] A. J. P. Delima, "Application of Time Series Analysis in Projecting Philippines ' Electric Consumption," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 5, pp. 694–699, 2019.

[14] K. Molebatsi and M. Raboloko, "Time Series Modelling of Inflation in Botswana Using Monthly Consumer Price Indices," *Int. J. Econ. Financ.*, vol. 8, no. 3, pp. 15–22, 2016.

[15] M. Y. Orong, A. M. Sison, and A. A. Hernandez, "Mitigating Vulnerabilities through Forecasting and Crime Trend Analysis," *2018 5th Int. Conf. Bus. Ind. Res.*, pp. 57–62, 2018.

[16] M. Y. Orong, A. M. Sison, and R. P. Medina, "A Hybrid Prediction Model Integrating a Modified Genetic Algorithm to K-means Segmentation and C4.5," in *TENCON 2018 - 2018 IEEE Region 10 Conference*, 2018, pp. 1853–1858.

[17] M. Y. Orong, A. M. Sison, and R. P. Medina, "A new crossover mechanism for genetic algorithm with rank-based selection method," in *5th International Conference on Business and Industrial Research: Smart Technology for Next Generation of Information, Engineering, Business and Social Science, ICBIR 2018*, 2018, pp. 83–88.

[18] P. G. L. Denila, A. J. P. Delima, and R. N. Vilchez, "Analysis of IT Graduates Employment Alignment Using C4.5 and Naïve Bayes Algorithm," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1, pp. 745–752, 2020.

[19] A. J. P. Delima, "Predicting Scholarship Grants Using Data Mining Techniques," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 4, pp. 513–519, 2019.

[20] A. J. P. Delima, "Applying Data Mining Techniques in Predicting Index and non-Index Crimes," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 4, pp. 533–538, 2019.

[21] G. R. Jadraque, A. J. P. Delima, and R. N. Vilchez, "Algorithmic Analytics for Outcomes-based Tertiary Education Performance Assessment," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1, pp. 766–773, 2020.

[22] J. L. D. Mercaral, A. J. P. Delima, and R. N. Vilchez, "Prediction of Employees' Lateness Determinants Using Machine Learning Algorithms," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1, pp. 779–783, 2020.

[23] U. O. Cagas, A. J. P. Delima, and T. L. Toledo, "PreFIC : Predictability of Faculty Instructional Performance through Hybrid Prediction Model," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 7, pp. 22–25, 2019.

[24] E. L. Polinar, A. J. P. Delima, and R. N. Vilchez,

"Students Performance in Board Examination Analysis using Naïve Bayes and C4.5 Algorithms," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1, pp. 753–758, 2020.

[25] J. C. Alejandrino, A. J. P. Delima, and R. N. Vilchez, "IT Students Selection and Admission Analysis using Naïve Bayes and C4.5 Algorithm," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1, pp. 759–765, 2020.

[26] J. S. Gil, A. J. P. Delima, and R. N. Vilchez, "Predicting Students' Dropout Indicators in Public School Using Data Mining Approaches," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1, pp. 774–778, 2020.

[27] A. J. P. Delima and M. T. Q. Lumintac, "Application of Time Series Analysis for Philippines ' Inflation Prediction," *Int. J. Recent Technol. Eng.*, vol. 8, no. 1, pp. 1761–1765, 2019.