Journal of Theoretical and Applied Information Technology
31st July 2022. Vol.100. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A PREDICTION ON EDUCATIONAL TIME SERIES DATA USING STATISTICAL MACHINE LEARNING MODEL -AN EXPERIMENTAL ANALYSIS

VANITHA.S[1], JAYASHREE.R[2]

[1]Research Scholar, [2]Assistant Professor

Department of Computer Application, SRM Institute of Science and Technology, Kattankulathur, India.

E-mail: [1]vs1703@srmist.edu.in, [2]jayashrr@srmist.edu.in

ABSTRACT

Prediction using time-series data is a vital part of machine learning because it keeps the temporal information of historical data for forecasting. Time series analysis is extensively used in all sectors wherever the data is populated and estimated based on timing such as seconds, minutes, hours, days, months, quarterly, half-yearly, and yearly. However, the model accuracy relies on the number of observations (data), consistency, and the consequence of data. The contribution of this paper is finding the trend of an educational institution enrollment in the upcoming year using the statistical machine learning mode1. Then, a detailed study has conducted to find the capability of the statistical model observed in various scenarios to handle time-series data. The study reveals the factors (Model fitness, Best forecasting duration, Impact of Train/Test ratio in precision) affecting the model accuracy of the statistical algorithms. This work also fulfills the research gap where less work has conducted in year-wise cyclic data without any trend. The methods used for this experiment are Auto-Regressive Integrated Moving Average (ARIMA) and Simple Exponential Smoothing (SES) technique. Finally, the two models are compared and the research objectives are discussed with the experimental result. Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) metrics are used for assessing the model precision. The experimental result proves that the SES model provides better performance than ARIMA and both models are executed with their own merits and demerits.

Keywords: *Time Series, ARIMA, Simple Exponential Smoothing, Student Enrollment Prediction, Factors Affecting the Model*

1. INTRODUCTION

Nowadays, the significance of time-series prediction is progressively increasing due to the power of forecasting. Especially in the business area such as sales quantity and share market price prediction [1-3] to get an idea before investing money, and the prearrangement of required resources, etc. Time-series are also used in energy consumption and production areas such as electricity usage and oil well performance prediction to measure the need in the future and take necessary action according to the result [4, 5]. Time series forecasting extended to nature-related happenings such as forecasting climate and precipitation of a particular place, rainfall, temperature, wind speed, and slope stability in hill stations [6-10]. This prediction helps to avoid major disasters by pre-planning the arrangements. Currently, it gives significant contributions to the medical field [11] to predict chronic diseases early based on past clinical data. In this technology world time-series data also helps to predict network failure, workload and resource usage, network traffic analysis, and road traffic forecasting [12-15] In Education, time series forecasting is used for predicting student enrollment [16-19], student academic performance, and educational institution growth in a country [20]. At all times, getting admission to a reputed institution is a contest for students and parents. The government is also intricate directly by making arrangements for seat allotment and other amenities. These arrangements rely on the number of applications received every year for a particular course. Here machine learning plays a significant role in enrollment prediction by applying statistical calculations to time series data.

1.1 Time Series Data and Properties

Time series is a collection of data recorded at a particular time interval (t). This interval is called lags. The lags may be any time unit such as minutes, hours, and year. The data can have the following properties such as trend, seasonality, cyclic, and noise.

Trend: The Long term upward or downward trend of given data. It may or may not have fluctuation but there is an increasing or decreasing mode in the series. For example, the Indian Population is a perfect example of a growing trend.

Seasonality: The systematic change of given data over a particular period. This period may be a specific day, week, month, or quarter. For example, the sales of cool drinks are usually high every summer season in a year.

Cyclic: The fluctuation of given data is not fixed and the movement of data is irregular then it is called cyclic data. It also takes a long time to change. For example, the selected dataset is cyclic data.

Noise: It is unusual data available in the given dataset and it is independent, not correlated with other values.

1.2 Problem Statement

The contribution of this paper is following:

a) Forecasting the student enrollment trend of an educational institution with the highest accuracy.
b) Identify the factors affecting the model accuracy in the following objectives:
   i. Fitness of statistical model in cyclic data
   ii. Suitable forecasting period (short/long term)
   iii. Impact of train and test ratio in predicted values
   iv. Impact of train and test ratio on accuracy.

In general, less than four year forecasting is considered short-term, and greater than five years are considered long-term. This paper is divided into 5 sections. Section 2 reviews the other works relating to student enrollment and prediction using ARIMA and SES. In section 3, the methodologies followed for this work are explained. In section 4, results and discussions are added. In section 5, the conclusion and future perspectives are included.

2. LITERATURE REVIEW

In this section, a few of the related articles are discussed and this review gives an idea about the existing work. Anggrainingsih et al.[10], developed a model to predict the number of university website visitors using Exponential smoothing (ES). This ES method is chosen due to the reason for adapting all types of data. Optimum constant value selected for this work is α =0.24, β =0.10, and γ =0.01 and the MAPE value is 12.95%.

Slim et al.[11], suggested a model to identify the factors (gender, ethnicity, GPA, first-generation, parent income) that increase the chance of enrolment. It helps to find the right applicant to join that institution using Logistic Regression, SVM, and semi-supervised Probability method on time series data. Sun et al. [12] proposed a model to predict student learning behavior in MOOC data using ARIMA that helps to improve the quality of teaching.

2.1 Related work

Onyeka-Ubaka et al. [13], suggested a model to forecast student enrollment. The selected data divided into two sections one is for the estimation period another one is for the validation period. The result shows the growing trend of future admission. Binu et al.[14], developed a model to predict student enrollment from other states using ANN and compared it with the ARIMA model. The collected historical data managed in the cloud due to its large size.

Yakubu et al. [15], created a model to find the student enrollment trend using the past 15 years of data to forecast the next five years. Also analyzed the factors that affects the enrollment process on management side and identified several reasons (inadequate infrastructure and course-related facilities).

Chen et al. [16], proposed a forecasting model for international undergraduate enrollment. For this study, the author collected semester-wise data (summer, fall, and spring) to increase the data points and to reflect the seasonality. The additional parameters (tuition fee, visa policy, and foreign students) which are affecting the enrollment process is also added for model creation by using one hot ending technique to interpret the availability of data. The best model is selected based on the result of the Schwartz Bayesian Criterion (SBC) and Akaike Information Criterion (AIC) by changing the parameter values of SARIMA (p,d,q)(P, D, Q).

Bousnguar et al. [17], developed four different forecasting models using statistical and deep learning model to predict the new student enrollment of the IBN ZOHR University. The author collects the enrollment details from various departments and there are three categories of student enrollment (new student, transfer student, and former student). The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics are used for model evaluation. In this comparative study, the fuzzy Time Series gives better performance than other models (MAE = 187, RMSE = 211). Here, the notable thing is that the statistical methods provide a lower error rate than the deep learning model due to the insufficient training to the algorithm.

Cruz et al. [18], established a model to predict the future enrollment trend of Cebu Technological University using ARIMA. The selected data set and the predicted values show the growing trend of enrollment. Finally, the best values are selected based on the lowest Akaike Information Criterion (AIC). Since the data is based on trend, second-order differencing is used and Moving Average mode1 (0, 1, 2) is applied.

Hsueh et al. [19], proposed a hybrid model (WOASVR) using Whale Optimization Algorithms (WOA) and Support Vector Regression (SVR) to predict the student enrollment and teacher count based on multiple time series datasets downloaded from the Taiwan government educational database. A case study is done on educational institutions from primary to university level for students and teachers. The proposed model gives better performance than ARIMA and Exponential Smoothing.

Muhammad and Bin Luo [20], proposed an educational institutional prediction model (EIPM) for the yearly time series data collected from 1970 to 2016 with 7 features and compared the effectiveness of the Linear Regression model. Here, the model is developed for all the levels from primary to higher education for both female and male institutions with an accuracy of 60.23%. Finally, the author concluded that the growth of educational institutions is lower than the population growth, especially in female institutions.

2.2 Prediction using ARIMA and SES

Siregar et al. [21] , proposed a model to predict palm oil production using the Exponential Smoothing technique. The authors created different models to compare the performance of smoothing technique variants such as single exponential, double exponential holt, triple exponential additive, and multiplicative. RMSE is used for evaluating the error and finally, the triple exponential additive gives the highest accuracy among others. The selected parameter values are $\alpha = 0.6$, $\beta = 0.02$, and $\gamma = 0.02$ and the RMSE value is 0.10.

Oliveira et al.[22] , established a model to predict the electricity demand across various countries and forecast 24 months in advance using ARIMA and the Exponential Smoothing technique. The author proposed a bagging technique to improve forecasting accuracy. And, the result shows that the proposed method increases the performance in most cases as expected.

In [23] , Taylor James proposed a model to predict short-term electricity usage for the next day and next week using two seasonal pattern time series data collected with the interval of half an hour. The author used multiplicative seasonal ARIMA and the Holt-Winters exponential smoothing to handle this seasonal data. Among these, the double seasonal Holt-Winters give the highest accuracy than the traditional Holt-Winters and double seasonal ARIMA model.

Ravinder [24] , has done a detailed study on the Exponential Smoothing technique for constant value selection using simulated data. This work gives guidelines to select an effective constant value for all kinds of data such as linear, non-linear, and trend-based. The author concluded that less value (0.1- 0.3) gives less error rate if the data is having no trend. Higher values signal the presence of a trend in data and the results for non-linear according to the constant variable are not easily generalizable. The literature using ARIMA and SES is summarized in table 1 and 2.

The previous work exhibits the model accuracy using any metrics such as AIC, RMSE, MAPE, and MAE and the following are the essence of the review. The following are the implications of the review. In Ref. [11], the factors that increase the chance of enrolment are analyzed (gender, ethnicity, first-generation, and parent income) to predict the right applicant. Prediction of other state students focused in [14], the reason why the enrollment trend is not in increasing mode [15] and in [16] the author used endogenous variables to predict international undergraduate enrolment. The variables which are affecting the enrolment process are called endogenous. For example Tuition fees and the Visa policy of that country. But this research focusing the factors affecting the model accuracy of the statistical method by forecasting student enrolment in upcoming years. The figure 1 represents the workflow of the existing and proposed system.

The model accuracy depends on many factors [17]; the fitness of the model selected for that problem, nature of the dataset and size, forecasting period, and the parameters affecting the model (train-test ratio, etc..). These factors are going to find with enough evidence of the required experiment result. Ref. [21-24] added for attaining knowledge about ARIMA and the Exponential Smoothing method using time series data.

Table 1: Related Work

Source	Lags	Nature of data	Methodology	Findings
[13]	Yearly 1962 - 2016	Trend	ARIMA	Enrollment prediction of Logas University in Nigeria. The best model is evaluated using the following metrics: AIC, SBC, AME, RMSE, and MAPE.
[14]	Yearly 2011 - 2018	Trend	Artificial Neural Network, ARIMA	Predicting student enrollment of the university from other states. The result shows both the ANN and ARIMA gives closer performance with above 95% accuracy.
[15]	Yearly 2004 - 2018	Irregular (No trend, seasonal, cyclic)	ARIMA	Forecasting student enrollment of Bolgatanga Polytechnic and the ARIMA (1, 0, 0) is identified as a fitted model to forecast the next 5 years.
[16]	Yearly thrice 1999 - 2014	Cyclic, seasonal with trend	Seasonal ARIMA (SARIMA)	Student enrollment prediction for Midwest University and the best model is selected based on AIC and SBC value.
[17]	Yearly 2003 - 2020	Cyclic	ARIMA, Exponential Smoothing, Fuzzy Time series, Long Short-Term Memory	Comparative study: Student enrollment prediction using IBN ZOHR University data. The fuzzy time series gives the highest score (MAE = 187, RMSE = 211).
[18]	Yearly 2012 - 2019	Trend	ARIMA	Student enrollment prediction for Cebu Technological University and the best model are selected based on the lowest AIC value.
[19]	Yearly 1999 - 2018	Trend (Multiple datasets)	Whale Optimization Algorithms (WOA), Support Vector Regression (SVR), ARIMA and Exponential Smoothing	Predicting student enrollment and teachers' count based on multiple time-series datasets. The proposed method performs well and the metrics used for evaluating all the models are RMSE and MAPE.

Table 2: Prediction using ARIMA and SES

Source	Lags	Nature of data	Methodology	Findings
[3]	Yearly	Seasonal	Seasonal ARIMA (SARIMA)	Prediction on precipitation and temperature in Bhagirathi river basin and the model over-predicts the extreme events in rainfall and temperature.
[4]	Daily	Irregular	ARIMA	Multivariate rainfall prediction using 12 metrological parameters

[5]	Annual	Trend	ARIMA	Prediction on the annual surface temperature of Libya using ARIMA and found two best models out of it which are linear trend (3-1-2) and quadratic trend model (3-2-3).
[21]	5 months once	Trend	Single exponential, Double exponential holt, Triple exponential additive, multiplicative	A comparitive study of palm oil production. Triple exponential additive gives better accuracy(RMSE:0.10).
[22]	Monthly	Cyclic	ARIMA, SES	Long-term electricity demand prediction across various countries. Bagging technique is used to improve performance.
[23]	Half-hourly	Two seasonal pattern	Multiplicative seasonal ARIMA, Holt-Winters exponential smoothing, double seasonal Holt-Winters exponential	Predicting short-term electricity demand and the double seasonal Holt-Winters gives the highest accuracy than others.
[24]	Multiple datasets	Trend, Linear, and Non Linear	Exponential Smoothing	This work gives guidelines to choose smoothing constants effectively; a lesser value (0.1-0.3) gives good accuracy than the higher value.

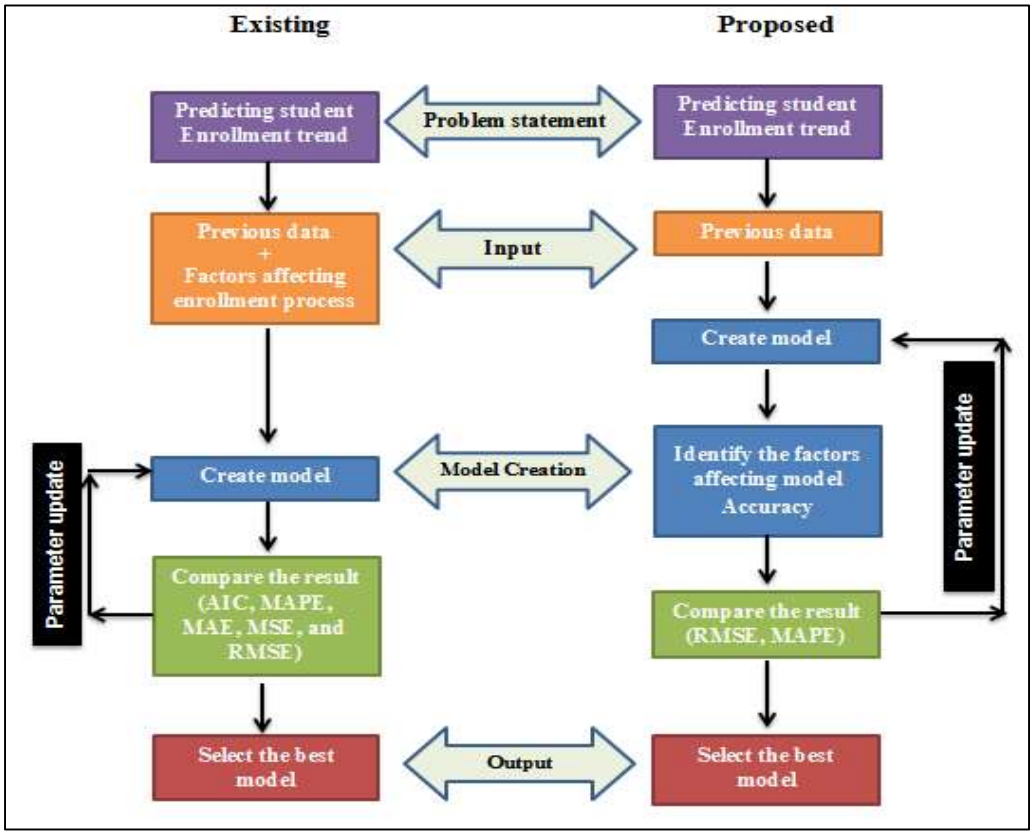

Figure 1: Existing and Proposed model

3. METHODS AND MATERIALS

The framework of enrollment prediction includes the following topics:
3.1 Dataset
3.2 Exploratory Data Analysis (EDA)
3.3 Auto-Regressive Integrated Moving Average (ARIMA)
3.4 Simple Exponential Smoothing (SES)

3.1 Dataset

In an educational institution all the information relating to admission process such as starting date of the application issue, the last date for application submission, the total number of applications sold, and the total number of applications received is notified regularly through the online/offline newspapers, media, and university website. For this research, the total number of applications received every year is collected from online newspapers [25-27]. Totally 14 years of data from 2008 to 2021 and predicted the student enrollment in next 3 years using the ARIMA and SES model.

3.2 Exploratory Data Analysis (EDA)

The figure 2 represents the number of applications received every year. Analyzing the nature of data is necessary to select the right model and it shows that this is a cyclic data without trend and seasonal activity.

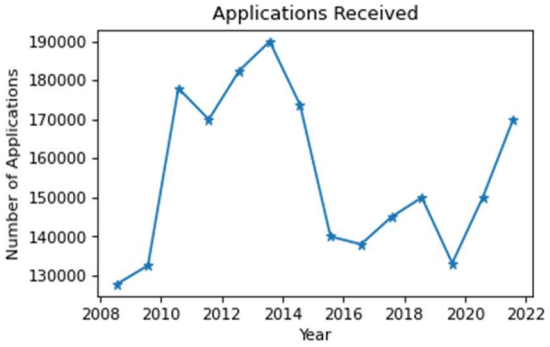

Figure 2: Total Applications Received(2008-2021)

3.2.1 Augmented dickey-fuller test

After analyzing the properties of the given dataset, it is essential to check the stationary of the data for further process. Augmented Dickey-Fuller Test (ADF) is used to identify the stationarity of data, and it shows the given data is stationary.

Null hypothesis (H0) = Data is non-stationary

Alternate hypothesis (Ha) = Data is stationary

In table 3, the p-value shown is 0.0011 and it is less than 0.05. If the p-value is less than 0.05 ($p<0.05$), the given data is stationary or else it is non-stationary. Therefore, the null hypothesis is rejected (H0) and the alternate hypothesis (Ha) is accepted. The ADF test static also shows that it is closer to the critical value in table 3.

Table 3: Augmented Dickey-Fuller Test Result

Parameter	Value
ADF Test Statistic	-4.052627
p-Value	0.001159
Lags Used	5.000000
Number of Observations Used	8.000000
Critical Value (1%)	-4.665186
Critical Value (5%)	-3.367187
Critical Value (10%)	-2.802961

3.2.2 Rolling mean and standard deviation

The figure 3 shows the rolling mean and standard deviation of input data. The rolling mean defines the average value of the particular time window then it is adjusted while moving on to the next value. The length of the time window is selected based on the dataset size. Here, the selected window size is 2, and this is working based on the simple moving average technique. The rolling standard deviation shows the deviation of changes from the global mean and it is the square root of variance.

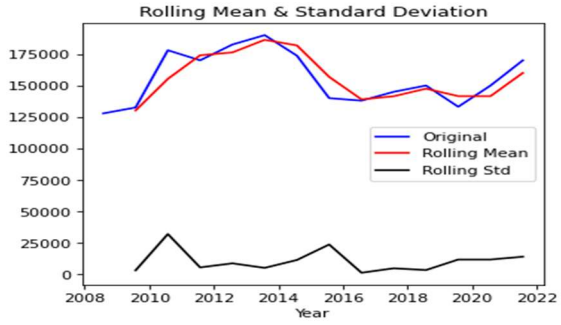

Figure 3: Rolling Mean and Standard Deviation

3.3 Auto-Regressive Integrated Moving Average (ARIMA)

The ARIMA model is suitable for non-seasonal data. It is an integrated (I) version of the Auto-Regressive (AR) and Moving Average (MA) model. AR model uses the preceding lag (t-1) values to predict future value (t+1), and the MA model uses the deviation of previous lags in the series to predict the future value [5]. There are three parameters (p, d, q) that decide the type of ARIMA model. The p represents the AR model, q represents the MA model and the d represents I, the number of

Journal of Theoretical and Applied Information Technology
31st July 2022. Vol.100. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

mathematical differencing is needed to make the data stationary.

A simple AR (p) model is written as the following formula:

$$x_t = c + \sum_{i=1}^{p} \emptyset_i x_{t-i} + e_t \quad (1)$$

Where c is constant, $x_t is$ the value observed at time t, $\emptyset_i$is the parameter of the model, and e_t is the error value. An MA (q) model can be written as the following formula:

$$x_t = e_t + \sum_{i=1}^{q} \theta_i e_{t-i} \quad (2)$$

Where θ_i is the parameter of the model and $e_t$is the error value [4]. Finally, the integrated AR and MA model of ARIMA can be written as the following formula:

$$x_t = c + \sum_{i=1}^{p} \emptyset_i x_{t-i} + e_t + \sum_{i=1}^{q} \theta_i e_{t-i} \quad (3)$$

The p and q parameters are called the AR and MA models, respectively. Partial Auto Correlation Function (PACF) and Auto Correlation function (ACF) are two functions known to suggest the optimized p and q values. It identifies the relationship between the current data with previous values. The difference between the two methods is ACF considers all the time series properties such as trend, seasonality, cyclic, and noise to calculate the correlation but the PACF considers the things partially.

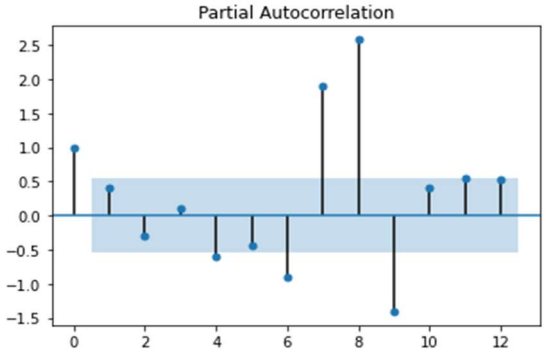

Figure 4: PACF Result

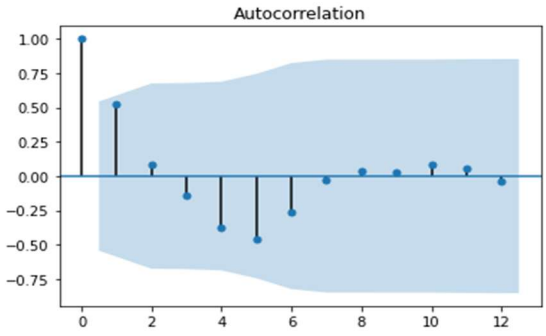

Figure 5: ACF Result

In figures 4 and 5, the blue region shows the confidence level and the spikes outside the region are indicate the lag order to use in AR and MA models. Finally, (p=1, d=0, q=0) parameter values are selected to get better precision than other models. Here the AR model is applied for prediction therefore p=1.

3.4 Exponential Smoothing (ES) Model

Exponential Smoothing (ES) is an alternative forecasting technique to the ARIMA model. ES is following the Simple Moving Average but it differs from weight assignment. The Moving average gives equal weight to the past data but the ES assigns weights specifically higher for the recent data and lowers for old data [22]. ES exponentially decreases the weight of older values hence the older values will not impact the model. Here, α value decides the weightage of the historical data. There are three types of Smoothing methods:

1. **Simple Exponential Smoothing (SES)**: Suitable for the data which have no clear trend and seasonality.
2. **Double Exponential Smoothing (DES)**: Suitable for the data which have trend but no seasonality.
3. **Triple Exponential Smoothing (TES)**: Suitable for the data which have trend and seasonality.

In this given dataset, there is no clear trend and seasonal activity. Thus the Simple Exponential Smoothing (SES) is selected for this experiment. The first step is to find the optimal weight (α=alpha) for better precision. This is also called a smoothing constant (α) and the value is between 0 and 1 [24]. The alpha value (α) can be assigned manually from 0.1 to 0.9 else the model will automatically find the optimal α value. SES is working based on the following formula:

$$p_{t+1} = \alpha x_t + (1-\alpha) p_t (\, 0 < \alpha < 1) \quad (4)$$

Wherep_{t+1} = predicted value at time t+1,
x_t= actual value at time t,
p_t= predicted value at time t,
α = degree of smoothing range 0 to 1.

4. RESULT AND DISCUSSION

Python is used for implementing the ARIMA and SES model in Google Colab. The figure 6 represents the predicted values for the next three years using ARIMA and table 4 shows the predicted and actual values of test data. The ratio of

Journal of Theoretical and Applied Information Technology
31st July 2022. Vol.100. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

train and test split is 65:35. The predicted values are between 1.57 to 1.63 lakh.

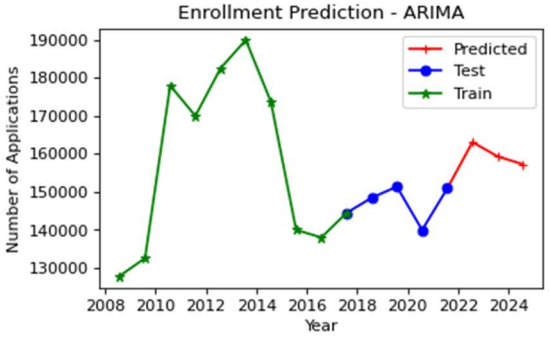

Figure 6: Short Term Prediction using ARIMA

Table 4: Actual and Predicted values of enrollment

Year	Predicted	Expected
2017	144426	145000
2018	148489	150000
2019	151457	133166
2020	139784	150000
2021	150920	170000
2022	163104	
2023	159338	
2024	157280	

Using the SES model, figure 7 shows the predicted values of test data with different α values which are 0.2, 0.5, and 1.0. Among those α=1.0 gives the highest fitting than others. The ratio of train and test split is 80:20. The figure 8 represents the predicted values of next 3 years.

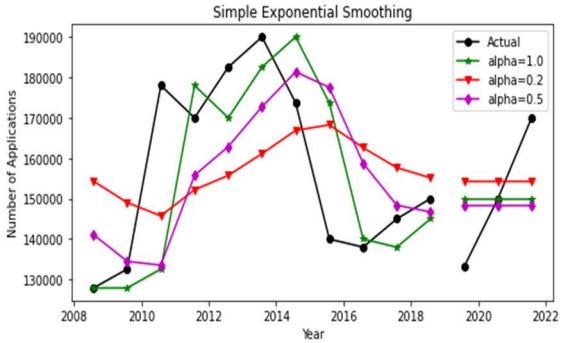

Figure 7: Train & Test (80:20) data of enrollment using SES model

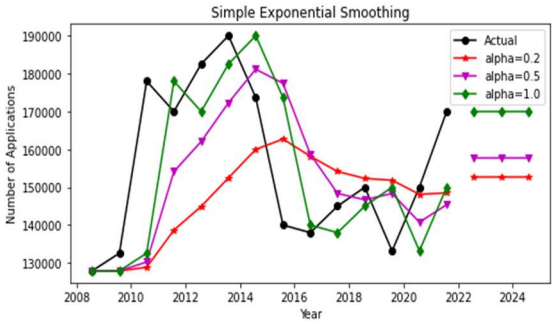

Figure 8: Short Term Prediction using SES

According to the result obtained in the experiment, the following research objectives are discussed.

4.1. Fitness of Statistical Model in Cyclic Data

Time series prediction entirely depends on the previous lags. If the nature of data is linear it can be predicted with the highest accuracy. The figure 6 shows the predicted values of test data using the ARIMA model, and table 4 shows the actual and predicted test values. Here, the 2017 and 2018 predictions are almost closer to the actual values because the values are showing a growing trend from 1.45 lakhs to 1.5 lakhs. But from 2018 onwards, there are sudden changes in data so the predicted values are not closer to the actual value. The figure 7 shows the predicted test data using SES and all the values are the same without any fluctuation. This is the nature of the SES model. In general predicting cyclic data is a bit difficult process. Here, the predicted values of the ARIMA model show the changes according to previous lags but the SES just follows the most recent value of the sequence without any fluctuation.

4.2. Suitable Prediction Period (Short/Long term)

These stats models are suitable for short-term prediction. If we try for long-term (10 years) prediction the values are get flattened at one stage. The figures 6 and 8 show the result for short-term prediction and figure 9 shows the result for long-term prediction using ARIMA.

Journal of Theoretical and Applied Information Technology
31st July 2022. Vol.100. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

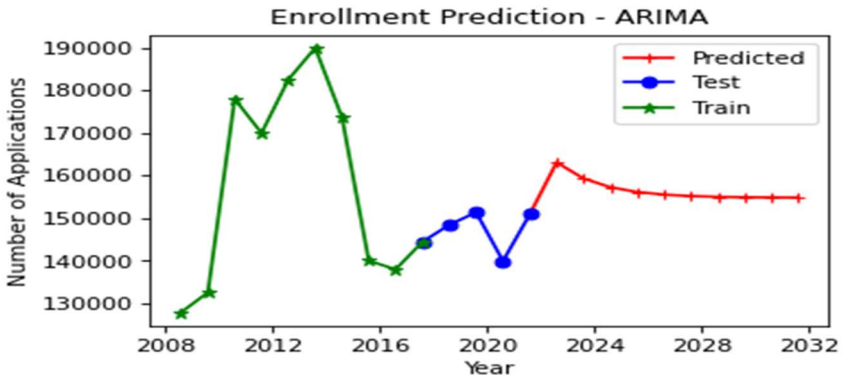

Figure 9: Long Term Prediction using ARIMA

Table 5: ARIMA Model Result

	Impact of Train and Test Ratio in Predicted Values		Accuracy	
Train & Test Ratio	**Test data**	**Forecasting data**	**RMSE**	**MAPE**
65:35	**No**	**No**	**12692**	**6.43%**
75:25	No	No	14188	7.18%
80:20	No	No	16360	8.92%
90:10	No	No	15303	9.10%

Yes - Train and test ratio affecting the test or forecasted values.
No - Train and test ratio is not affecting the test or forecasted values.

Table 6: SES Model Result

	Impact of Train and Test Ratio in Predicted Values		Accuracy			
Train & Test Ratio	**Test data**	**Forecasting data**	**RMSE**	**MAPE**	**α value (manual)**	**α value (optimized)**
65:35	**Yes**	**No**	**11962**	**5.72%**	**0.5**	**1.0**
75:25	Yes	No	13135	6.31%	0.4	1.0
80:20	Yes	No	15092	8.13%	1.0	1.0
90:10	Yes	No	28636	16.44%	1.0	1.0

4.3. Impact of Train and Test Ratio in Predicted Values

Splitting the dataset with the correct ratio for train and testing is an important step because it helps to create a more generalized model. This statement is true only when the data is distributed evenly on both train and test sets. The above tables 5 and 6 shows the result of different train and test ratio of the ARIMA and SES model. The predicted values are not changed in all cases. As per the results shown in tables 5 and 6, changing the ratio of train and test is not affecting the prediction value of the ARIMA model in both test and forecasting.

But in the SES model, the test data is changed because the prediction mainly depends on the most recent past value. Also figure 7 shows the predicted test result is started closely from the last value of training data. SES is not considering all the previous changes in data sequence but it consider α value to assign weight. Likewise, forecasting the next 3 years also follows the last value of the dataset that is 2021(figure 8).

4.4. Impact of Train and Test Ratio on Accuracy

The Accuracy level is changed according to the train and test ratio. The tables 5 and 6 show the highest accuracy in the 65:35 split for both stats models. In SES, selecting the best alpha (**α**) is very

important and there are two ways, assigning α value from 0.1 to 0.9 manually and allowing the system to find the optimized **α** value. Both values are listed in table 6.

The derived result shows that the performance of the model is determined by the various parameters. They are i) Nature of data whether it is linear or stochastic nature data, linear data helps to the selected model to get accurate forecasting as per the first objective. ii) Duration of forecasting, here the short-term prediction gives a better result than the long term as per the second objective. iii) Splitting the dataset for train and test, affect the model accuracy but does not reflect any changes in forecasting values. Finally, the SES model gives less error rate than the ARIMA model.

The study reveals that the existing work done by any of following techniques such as comparative study, hybrid approach, and optimization in parameter selection. Here the technical study of statistical algorithms is missing in ARIMA and SES model [3-5], [13-19], [21-24]. Even in Ref. [24], the author gives guidelines to choose the constant variables in the Exponential smoothing method using simulated data, but that is different from my work. This paper analyzed the technical capability of the statistical model by finding the factors affecting the model accuracy in different scenarios. Fitness for handling non-periodic cyclic data, whether the sudden changes affect the prediction accuracy or not, then analyzed the suitable prediction period and the impact of train and test ratio on the accuracy level. This experimental result proves the merits, demerits of each model, structure of the algorithms, and the influence of data used in the experiment.

5. CONCLUSION AND FUTURE WORK

Predicting student enrollment in an educational institution is a crucial process for management to provide proper facilities in the upcoming year. The following are the external factors such as unemployment, fee structure, government quota, foreign opportunities, and the income level of individual family to impact the enrollment process. According to the collected data, around ten years back, students are curious to join engineering courses and it reached a peak value of 1.9 lakh applications received in 2013. After that, it gradually decreased up to 1.3 lakh then started increasing slowly. This volatility data is predicted successfully using the ARIMA and SES model. Totally 14 years of data is collected for prediction. Both stats models are evaluated by splitting the data with different train and test ratios to answer the research objectives. As per the result obtained 65:35 split gives the highest accuracy in both stats models. Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) are used as error metrics. The lowest error rate of ARIMA is MAPE = 6.43% and RMSE =12692 and for SES, the MAPE = 5.42% and RMSE =11962. The forecasting result shows the downward enrollment trend for the upcoming year. Due to its cyclic nature, the last three years (2019-2021) of data (figure 2) show an increasing trend now the prediction show the decreasing trend. Also, the factors affecting the model accuracy is investigated; i). The selected model predicts the linear data with less error rate than the non-linear cyclic data. ii). Short-term prediction is suitable and effective to follow compared with long-term forecasting. iii) Finally, in the statistical method, even the impact of the train-test ratio is not affecting the forecasting values it alters the accuracy of the model. Thus the experimental result illustrates the importance of choosing the right model for the selected problem, the best forecasting duration, and the influence of the train test ratio on model precision using the statistical time series method.

Limitation: There are a few points to notify here. First, the educational system is not common for all the countries and everyone following different structure. Annual enrollment (yearly) or semester-wise (spring, fall, and summer) enrollment are the two types generally followed by the higher education system. The developed model is suitable for the institution which is following the same structure. Second, collecting previous enrollment data in an educational institution is limited also maintained confidentially by the management. So there are few studies accomplished on this kind of problem statement and implemented with limited time-series data. It has to be solved in further educational research.

Third, Comparison of the current model with the existing one is not feasible. The technical reason is while using a statistical method, it is not mandatory to scale the input values before creating the model. So the error metrics shows the result in different range according to the input values. The evaluation is possible only when all the studies follow the common scaling technique in preprocessing. This drawback is rectified in deep learning time series algorithms.

Future Perspectives: ARIMA is the benchmark of time-series prediction and the SES is an alternative approach. Hence started the work using stats models and planned to work on advanced time

series algorithms such as RNN and its variants (LSTM, and GRU) to increase the accuracy level also avoid the drawbacks of statistical methods. The recent time-series algorithms have enough architecture to handle the previous lags effectively with minimal human intervention.

REFERENCES:

[1] D. Banerjee, "Forecasting of Indian stock market using time-series ARIMA model," *2014 2nd International Conference on Business and Information Management (ICBIM)*, 2014, pp. 131-135, doi: 10.1109/ICBIM.2014.6970973.

[2] F. Mahia, A. R. Dey, M. A. Masud and M. S. Mahmud, "Forecasting Electricity Consumption using ARIMA Model," *2019 International Conference on Sustainable Technologies for Industry 4.0 (STI)*, 2019, pp. 1-6, doi: 10.1109/STI47673.2019.9068076.

[3] T. Dimri, S. Ahmad and M. Sharif, "Time series analysis of climate variables using seasonal ARIMA approach", *Journal of Earth System Science*, vol. 129, no. 1, 2020, pp. 1-16, Available: 10.1007/s12040-020-01408-x.

[4] A.Geetha and G. Nasira, "Time-series modelling and forecasting: modelling of rainfall prediction using ARIMA model", *International Journal of Society Systems Science*, vol. 8, no. 4, 2016, p. 361, doi: 10.1504/ijsss.2016.081411.

[5] E. El-Mallah and S. Elsharkawy, "Time-Series Modeling and Short Term Prediction of Annual Temperature Trend on Coast Libya Using the Box-Jenkins ARIMA Model," *Adv. Res.*, vol. 6, no. 5, 2016, pp. 1–11, doi: 10.9734/air/2016/24175.

[6] F. M. Khan and R. Gupta, "ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India," *J. Saf. Sci. Resil.*, vol. 1, no. 1, Sep. 2020, pp. 12–18, doi: 10.1016/J.JNLSSR.2020.06.007.

[7] Z. Wang *et al.*, "Failure prediction using machine learning and time series in optical network," *Opt. Express*, vol. 25, no. 16, 2017, p.18553-18565, doi: 10.1364/oe.25.018553
doi: 10.1016/j.neucom.2020.11.011.

[8] H. Yang, X. Li, W. Qiang, Y. Zhao, W. Zhang, and C. Tang, "A network traffic forecasting method based on SA optimized ARIMA–BP neural network," *Comput. Networks*, vol. 193, Jul. 2021, p. 108102, doi: 10.1016/J.COMNET.2021.108102.

[9] D. Xu, Y. Wang, L. Jia, Y. Qin, and H. Dong, "Real-time road traffic state prediction based on ARIMA and Kalman filter," *Front. Inf. Technol. Electron. Eng.*, vol. 18, no. 2, 2017, pp. 287–302, doi:10.1631/FITEE.1500381.

[10] R.Anggrainingsih, G. R. Aprianto, and S. W. Sihwi,"Timeseries forecasting using exponential smoothing to predict the number of website visitor of Sebelas Maret University," *ICITACEE 2015 - 2nd Int. Conf. Inf. Technol. Comput. Electr. Eng.*, pp. 14–19, 2016, doi: 10.1109/ICITACEE.2015.7437762.

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

[12] Sun, Dezhi, Tong Li, Fucheng You, Man Hu, and Zhenyu Li. "Prediction of learning behavior characters of MOOC's data based on time series analysis." In Journal of Physics: Conference Series, vol. 1994, no. 1, 2021, p. 012009. IOP Publishing.

[13] 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, 2017, pp. 80–90.

[14] P. K. Binu, A. Chandran and M. Rahul, "A Cloud-Based Data Analysis and Prediction System for University Admission," 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), pp. 1327-1332, 2019, doi: 10.1109/ICICICT46008.2019.8993328.

[15] 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, 2019, pp. 120–135.

[16] Y. Chen, R. Li and L. Hagedorn, "Undergraduate International Student Enrollment Forecasting Model: An Application of Time Series Analysis", *Journal of International Students*, vol. 9, no. 1, 2019, pp. 242-261. doi: 10.32674/jis.v9i1.266.

[17] H. Bousnguar, L. Najdi, and A. Battou, "Forecasting approaches in a higher education setting," *Educ. Inf. Technol.*, vol.27, no.2, July 2021,pp.1993-2011,doi: 10.1007/s10639-021-10684-z.

[18] A.P. Dela Cruz *et al.*, "Higher education institution (Hei) enrollment forecasting using data mining technique," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 2, 2020, pp. 2060–2064, doi: 10.30534/ijatcse/2020/179922020.

[19] S. Yang, H. C. Chen, W. C. Chen, and C. H. Yang, "Student Enrollment and Teacher

Statistics Forecasting Based on Time-Series Analysis," *Comput. Intell. Neurosci.*,2020, doi: 10.1155/2020/1246920

[20] M. S. Iqbal and B. Luo, "Prediction of educational institution using predictive analytic techniques," *Educ. Inf. Technol.*, vol. 24, no. 2, 2019, pp. 1469–1483, doi: 10.1007/s10639-018-9827-y

[21] B. Siregar, I. Butar-Butar, R. Rahmat, U. Andayani and F. Fahmi, "Comparison of Exponential Smoothing Methods in Forecasting Palm Oil Real Production", *Journal of Physics: Conference Series*, vol. 801, 2017, p. 012004, doi: 10.1088/1742-6596/801/1/012004

[22] E. M. de Oliveira and F. L. Cyrino Oliveira, "Forecasting mid-long term electric energy consumption through bagging ARIMA and exponential smoothing methods," *Energy*, vol. 144, Feb. 2018, pp. 776–788, doi: 10.1016/J.ENERGY.2017.12.049

[23] J. W. Taylor, "Short-term electricity demand forecasting using double seasonal exponential smoothing," *J. Oper. Res. Soc.*, vol. 54, no. 8, 2003, pp. 799–805, doi: 10.1057/palgrave.jors.2601589

[24] H. V. Ravinder, "Forecasting With Exponential Smoothing Whats The Right Smoothing Constant?", *RBIS*, vol. 17, no. 3, Aug 2013, pp. 117–126.

[25] R.Sujatha, "Anna University gets 1.80 lakh application forms," *The Hindu*, Jan. 16, 2011. [Online]. Available: https://www.thehindu.com/news/cities/chennai/anna-university-gets-180-lakh-application-forms/article8669827.ece

[26] "Tamil Nadu: 1.7 lakh apply for engineering seats in govt quota," The Times of India, Aug. 25, 2021. [Online]. Available: https://timesofindia.indiatimes.com/city/chennai/tamil-nadu-1-7-lakh-apply-for-engineering-seats-in-govt-quota/articleshow/85613668.cms

[27] "Tamil Nadu: Over 1.5 lakh apply for engineering courses this year," *The new Indian Express*, Jun. 03, 2018. [Online] .Available: https://www.newindianexpress.com/states/tamil-nadu/2018/jun/03/tamil-nadu-over-15-lakh-apply-for-engineering-courses-this-year-1823004.html