Application to predict the new student's score using time series algorithm

Sinar Nadhif Ilyasa [a,1,*], Husni Thamrin [a,2]

[a] Universitas Muhammadiyah Surakarta, Jl. A. Yani, Mendungan, Pabelan, Kartasura 57169, Indonesia
[1] nadhif909@gmail.com [*]; [2] Husni.Thamrin@ums.ac.id
* corresponding author

ABSTRACT

With the rapid development of information technology in this era, data accuracy is essential in our daily lives to solve existing problems. The existence of information is beneficial in helping the decision-making process. Therefore, any existing information can be further processed and analyzed to be used as new knowledge so that it is useful to determine the right decision. The purpose of this research is to determine whether an application using the time series algorithm such as Auto Regression, ARMA (Auto Regression Moving Average), and Triple Exponential Smoothing model. They can forecast prediction scores that may help to solve the student's admission problem. In this case of the project, the researcher found that the Universitas Muhammadiyah Surakarta's admission system is not evaluated correctly in accepting students and controlling incoming students' quality due to the lack of insights. This time series application is one solution to help manage incoming students' quality and quantity, especially in the Universitas Muhammadiyah Surakarta. This application is developed using a web framework called Django, a full-stack Python web framework that encourages rapid growth and clean, pragmatic design. The Auto Regression model is chosen as a prediction model in One Day Service (ODS) Universitas Muhammadiyah Surakarta. It has a better performance than ARMA and Triple Exponential Smoothing and a higher chance to avoid overfitting than the other two models that are more complex for the ODS data.

Keywords:
Time Series
Django framework
Auto Regression
ARMA
Triple Exponential Smoothing

I. Introduction

Management information systems in the academic world with information technology have increased rapidly due to their efficiency and effectiveness [1]. With the rapid development of information technology in this era, data accuracy is essential in our daily lives to solve existing problems. The existence of information is beneficial in helping the decision-making process [2]. Therefore, any existing data can be further processed and analyzed to be used as new knowledge so that it is useful to determine the right decision [3]. According to the standards determined by the following article 1 paragraph 17 Law of Indonesia Number 20 the Year 2003, the university's strategy to obtain qualified students states: "National education standards are the minimum criteria regarding the education system in the entire jurisdiction of the Unitary Republic of Indonesia". This strategy starts by examining the availability of information so that the university can control the quality of students who are accepted [4]. One of the private Islamic tertiary institutions in Surakarta, Universitas Muhammadiyah Surakarta, opens new student registration channels through the One Day Service (ODS). The service has been integrated with information systems to record any input data into the information system automatically. Although the service integrates with the information system, it does not utilize the incoming data to obtain useful knowledge. A system's reliability is measured from the level of information available when it is wanted, and all components operate that make up the system [5]. Available data from admission can be transformed into a prediction student's score data so that it can help to evaluate the admission process. The prediction result will give a new insight to control the number of accepted new students according to the specified quality. One method that utilizes data obtained to predict future data is to use the time series method [6]. The time series forecasting method is a method that collects the same variables to be analyzed and developed into a model of describing the relationships that underline it. There are many models for time series, including ARMA (Auto Regressive Moving Average) [7], ARIMA (Auto-Regressive Integrated Moving Average) [8], SARIMA (Seasonal Auto-Regressive Integrated Moving Average) [9], SARIMAX (Seasonal Auto-

Regressive Integrated Moving Average Exogenous) [10] and ARIMAX (Auto Regressive Moving Average Exogenous) [11].

This study compared three models: The triple Exponential Smoothing model, ARMA model, and Auto Regression model to decide which model is the best to fit the data given in the One Day Service (ODS) dataset. These models are preferred because it is suitable for predicting data that is already stationary. The purpose of this research is to determine whether an application using the time series algorithm such as Auto Regression, ARMA (Auto Regressive Moving Average), and Triple Exponential Smoothing model. They can forecast prediction scores that may help to solve the student's admission problem. In this case of the project, the researcher found that the Universitas Muhammadiyah Surakarta's admission system is not evaluated correctly in accepting students and controlling incoming students' quality due to the lack of insights. The model with the best prediction performance will be applied to forecast the student's score in the One Day Service (ODS) test at the Universitas Muhammadiyah Surakarta to control the admission test's difficulty. This application's development uses a web framework called Django, a full-stack Python web framework that encourages rapid growth and clean, pragmatic design [12]. The remainder of the paper is organized as follows. A proposed model time series is used in Section II to forecast the One Day Service (ODS) student's score. Section III presents the general methodology to build the time series model and build the application. Section IV discusses the result of the dataset analysis and the application's implementation, and Section V states some conclusions.

II. The Proposed Method

This research proposes three models for predicting the student's score in the One Day Service Universitas Muhammadiyah Surakarta: Auto Regression, ARMA (Auto Regressive Moving Average), and Triple Exponential Smoothing, which described as follows,

A. Auto Regression

As defined in (1), auto-regression is a regression model that utilizes the dependent relationship between current observations and past observations.

$$AR(t) = c + \theta_{1y_1} - 1 + \theta_{2y_2} - 2 + \cdots + \theta_{ty_t} - t + \epsilon, \tag{1}$$

with c is a constant, θ_1 and θ_2 are lag coefficients up to order t, ϵ is white noise

B. ARMA (Auto Regressive Moving Average)

ARMA is an Auto Regression combined with the moving average model. The moving average calculates the average value of the time series and then estimates the value in the next period. Moving Average (MA) model for lag order, one follows in (2).

$$MA(t) = \mu + \theta_1 \epsilon_{t-1} + \epsilon, \tag{2}$$

With μ is the expectation of MA (often assumed to be zero), θ_1 is the MA lag coefficient, ϵ is white noise. A combination of Auto Regression and Moving Average obtained ARMA model ($ARMA$) for lag order 1 in (3).

$$ARMA_t = c + AR \cdot y_{t-1} + MA \cdot \epsilon_{t-1} + \epsilon \tag{4}$$

With c is constant, AR is the Auto Regression lag coefficient, MA is the Moving Average lag coefficient, ϵ is white noise.

C. Triple Exponential Smoothing

The Triple Exponential Smoothing assigns exponentially decreasing weights as the observation gets old. Triple Exponential Smoothing introduces an overall smoothing, trend, and seasonal model fitted in the forecasting model. The y is the observation, I is the seasonal index, and t is an index denoting a time period. β, α, γ are constant and must be estimated to minimize the error. L represents the number of divisions per cycle. In our case, L is determined by looking at monthly data that displays a repeating pattern each year. Overall smoothing (S_t) is defined as,

$$S_t = \alpha \frac{y_t}{I_{t-L}} + (1-\alpha)(S_{t-1} + b_{t-1}) \tag{4}$$

The trend smoothing (b_t) in Triple Exponential Smoothing, the equation is denoted as (5),

$$b_t = \gamma(S_t - S_{t-1}) + (1-\gamma)b_{t-1}$$

In the Triple Exponential Smoothing, a seasonal smoothing (I_t) is added to capture seasoanlity that is given by (6),

$$I_t = \beta \frac{y_t}{S_t} + (1-\beta)I_{t-L} \tag{6}$$

After the model is fitted, to forecast a series using triple exponential smoothing is given by (7),

$$F_{t+m} = (S_t + mb_t)I_{t-L+m} \tag{7}$$

For the forecasting model, the variable m indicates how many periods to forecast.

III. Method

The research method uses three models to determine which model better fits in the One Day Service (ODS) data. The tested model in this research is ARMA (Auto Regressive Moving Average), Auto Regression, and Triple Exponential Smoothing. By comparing these three models' performance, the researcher determines which model has the best performance by calculating the error. The research step can be seen in Fig. 1.

The first stage performs a static test in One Day Service (ODS) data that includes students' scores in the Faculty of Engineering that occur in 4 years (2017-2019). A Stationary is compulsory to have in a dataset so that it can use a forecasting model. Stationarity definition is one whose statistical properties do not depend on how the data is observed. A time series with trends or seasonality are not stationary. A common approach to test stationarity is by using the Augmented Dicky Fuller test [13]. Augmented Dicky Fuller is an augmented version of the Dickey-Fuller test for a more extensive and complicated time series model. The Augmented Dicky Fuller test states the null hypothesis is = 1 (this is also called a unit test). A small t-value ($t<0.05$) indicates strong evidence against the null hypothesis. For a t-value less than 0.05, the null hypothesis is rejected for unit root, which means the dataset is stationary.

The second step is determining the prediction model. The model tested in this stage is Auto Regression, ARMA, and Triple Exponential Smoothing.
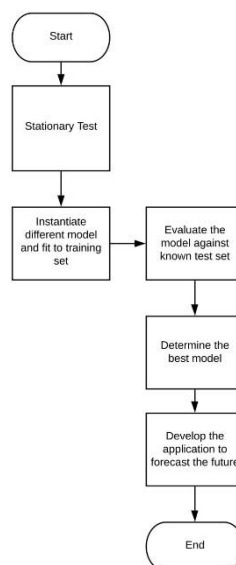

Fig. 1.Research steps.

The third stage is an evaluation between the test data and prediction data to test the accuracy performance by looking at the training and testing data's errors using root mean square error (RMSE). Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are. RMSE is a measure of how spread out these residuals are. The formula for RMSE is stated as (9).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \qquad (9)$$

The last stage predicts the number of engineering students in the future based on the best model obtained after getting a good enough model to be used as a prediction material. The result is applied in the Django web application using a waterfall methodology. The software tools to analyze the One Day Service (ODS) Universitas Muhammadiyah Surakarta prediction and develop the application are listed in Table 1.

Table 1. Development Tools

No	Tools Name	Specification
1	Anaconda Distribution	2019.10 Version
2	Windows OS	Windows 10 Pro
3	Python 3.7	64 bit
4	HighChart	Free version
5	MySQL Database	64 bit
6	Firefox Web Browser	Version 72.0.2 (64-bit)
7	Django Framework	Version 2.2

IV. Results and Discussion

A. Research Data

This research data is a stationary dataset, which is the criteria to implement most forecasting models like the ARMA model, and the proof is shown in the Data Examination section. An example of training data in the ODS (One Day Service) dataset is displayed in Table 2.

Table 2. Training Data

No Participant	Test Attempt	Date	Study Program Choice 1	Study Program Choice 2	Score
1700011112	2	05/01/2017	Chemical Engineering	Civil Engineering	40
1700011113	3	05/01/2017	Chemical Engineering	Architect	52
1700011211	1	05/01/2017	Industrial Engineering	Nutritionist	44
1700011811	1	05/01/2017	Industrial Engineering	Nutritionist	48
1700012011	1	05/01/2017	Industrial Engineering	Economy	42

Data presented in Table 2 is obtained from ODS (One Day Service) dataset archive. The data itself includes the participant number, the participant score, test date, test attempt, and study program choice. For our time series analysis, we will only need three features for this forecasting model, the test date as our timestamp observation (x-axis), the score to forecast data (y-axis), and the study program choice specific study program observation.

For testing purposes, the researcher uses Industrial Engineering data to test which model will be the best fit for prediction. The timestamp used for this forecasting analysis is days because for weekly timestamp, and the monthly timestamp for this dataset case do not offer a large enough training dataset. More data for training is frequently more accommodating, offering more noteworthy open doors for exploratory information examination, model testing, and tuning, and model fidelity. The data set the frequency to days, which means all data set are resampled to days (taking average within that day). The researcher interpolates the missing values with the dataset's mean to run the forecasting model algorithm in Fig. 3 and the raw data before interpolation in Fig. 2. The average score (after adding an average rating to the null values) for students enrolling in undergraduate degrees at Universitas Muhammadiyah Surakarta majoring in Industrial Engineering ranges from January 2017 to August 2019 that can be seen in Fig. 3.

Based on Fig. 3, the graph shows that it does not have precise seasonal data. The amount of data Industrial Engineering majors totaling 960 data will be broken down into training sets and data sets. Details of the dataset used in this prediction are described in Table 3.

Table 3. Dataset feature

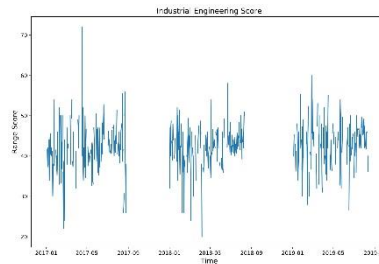| No	Description	Value
1	Count	960
2	Mean	42.63
3	Standard deviation	4.25
4	Minimum Value	20
5	Maximum Value	72

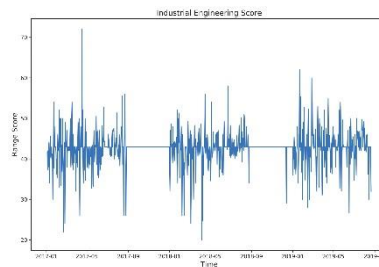Fig. 2. Raw average score registrant industrial engineering student.

Fig. 3. Modified average score registrant industrial engineering.

B. Data Examination

The ARMA model needs to be stationary; therefore, the dataset needs to examine whether the dataset given is stationary or not by looking at the p-value given in the Augmented Dicky Fuller test. The test for the Augmented Dicky Fuller Test is shown in Fig. 4.

```
Augmented Dickey-Fuller Test:
ADF test statistic      -28.218341
p-value                   0.000000
```

Fig. 4. ADF test.

In Fig. 4, the Augmented Dicky Fuller Test, the test result shows that the p-value is 0. Hence, because the p-value is less than 0.05, the data has no root, and it will reject the null hypothesis. The ADF test result shows that the dataset is stationary [13].

C. Prediction Results

For sample purposes, the researcher examines the Industrial Engineering dataset, and the test data set is taken as much as the last 20-200 days of data from Industrial Engineering. The sample dataset average score is 41.59; it is obtained by adding all the numbers in the dataset and dividing them by the amount of the data, calculated using the pandas' library. The stats model library performs a grid search to determine the ARMA order and the lag order in Auto Regression. In Fig. 5, the Auto Regression model (Orange line) almost performs the same as ARIMA predictions (Redline). In contrast, the Triple Exponential Smoothing model (Green line) shows a slight difference from the other model. To obtain the most optimal result, we wanted to achieve a model that matches the closest to the test data set with the prediction result and avoid overfitting. Overfitting is a crucial issue in

supervised machine learning which forestalls us from generalizing the models to fit practical information on preparing information, just as inconspicuous information on testing sets. [14].

Fig. 5 shows that overall there is a slight deviation to the mean among all three algorithms used. The slight variation is likely to occur because of the missing dataset interpolated with the mean. Overall, the prediction result performs well in capturing the trend or moving average, but it does not fit all test data well.

Table 4. Day Forward-Chaining Cross Validation

Data Index	Auto Regression	ARMA	Triple Exponential Smoothing
780-960	4.62	4.65	8.19
800-960	4.50	4.76	6.47
820-960	4.27	4.29	5.33
840-960	4.13	4.12	4.20
860-960	4.32	4.36	4.47
880-960	4.09	4.09	4.23
900-960	3.95	3.95	3.94
920-960	4.06	4.03	4.02
940-960	4.38	4.46	6.40

A cross-validation test evaluates the machine learning model in the Industrial Engineering dataset. This research's cross-validation technique is Day Forward-Chaining; it is based on a method called forward-chaining (also referred to as rolling-origin evaluation) [15]. In this technique, the researcher creates nine train/test splits and calculates the RMSE all over the splits. The technique successively considers each day as the test set and assigns all previous data into the training set. The cross-validation results are shown in Table 4.

Based on Table 2, the results show that the total error obtained in the Auto Regression model is smaller comparing to the ARMA and Triple Exponential Smoothing model when performing cross-validation using the Day Forward-Chaining test. The results also show consistency in performance in terms of error rate when comparing these three models. The Triple Exponential Smoothing has the average highest error among the other models. It is because the Triple Exponential Smoothing is overfitting the model. Evaluating our forecast result, minimizing the RMSE leads to forecasts of the mean. The fewer RMSE values, the better result.
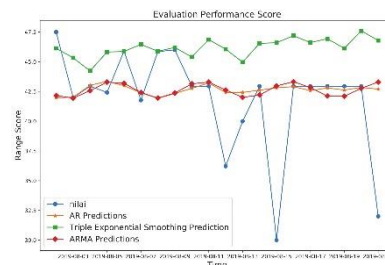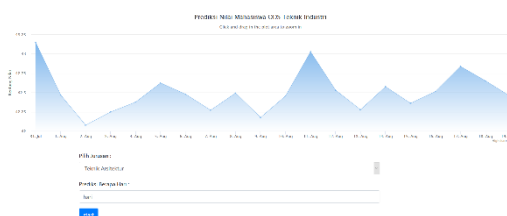

Fig. 5. Evaluation performance model.

Fig. 6. Prediction application.

D. Implementation

The research output is a web application that allows predicting student's admission scores in the One Day Service (ODS) data according to each study program. Prediction starts from the end of the training dataset given. There are two input forms in this program, the first option is to select a study program, and the second input form is to determine how many days to predict.

Fig. 6 shows the program's interface in the web browser running in localhost implemented using the Django Framework. The users can choose the study major and determine how many days to predict. After submitting, it will show the corresponding score prediction result that is visualized by the HighChart graph.

V. Conclusion

This prediction result is useful in capturing the trend or the moving average of the data but does not capture all test data due to some other data points are still not fitted well. Based on the results obtained, the errors between the three types of models are not significant, but overall the Auto Regression has the best performance. The Auto Regression model and ARMA model's error is very close, unlike the Triple Exponential Smoothing that has the most error in this case. In this research, the Auto Regression model is chosen to be used as a prediction model in One Day Service (ODS) Universitas Muhammadiyah Surakarta. It has a better performance than ARMA and Triple Exponential Smoothing and a higher chance to avoid overfitting than the other two models that are more complex. A simple model is preferred when dealing with a series with no trends or seasonality.

References

[1] B. Handaga and H. Amrullah, "Sistem Informasi Akademik Untuk Layanan Mahasiswa Ums Berbasis Mobile," *Emitor*, vol. 16, no. 02, pp. 9–20, 2017.

[2] Nia Komalasari, "Sistem Pendukung Keputusan Kelaikan Terbang (SPK2T)," *J. Ind. Elektro dan Penerbangan 4*, vol. 4, no. 1, pp. 1–11, 2014.

[3] D. H. Kamagi and S. Hansun, "Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa," *J. Ultim.*, vol. 6, no. 1, pp. 15–20, 2014, doi: 10.31937/ti.v6i1.327.

[4] A. Rohman, "Model Algoritma K-nearest neighbor (K-NN) untuk prediksi kelulusan mahasiswa," *Neo Tek. J. Ilm. Teknol.*, vol. 1, no. 1, 2015.

[5] H. Thamrin, "Perancangan Tools Berbasis Python Untuk Memantau Keaktifan Server," *Komuniti*, vol. 2, 2017.

[6] A. Khosravi, L. Machado, and R. O. Nunes, "Time-series prediction of wind speed using machine learning algorithms: A case study Osorio wind farm, Brazil," *Appl. Energy*, vol. 224, no. September, pp. 550–566, 2018, doi: 10.1016/j.apenergy.2018.05.043.

[7] H. Zhang, S. Zhang, P. Wang, Y. Qin, and H. Wang, "Forecasting of particulate matter time series using wavelet analysis and wavelet-ARMA/ARIMA model in Taiyuan, China," *J. Air Waste Manag. Assoc.*, vol. 67, no. 7, pp. 776–788, 2017, doi: 10.1080/10962247.2017.1292968.

[8] N. A. Bakar and S. Rosbi, "Autoregressive Integrated Moving Average (ARIMA) Model for Forecasting Cryptocurrency Exchange Rate in High Volatility Environment: A New Insight of Bitcoin Transaction," *Int. J. Adv. Eng. Res. Sci.*, vol. 4, no. 11, pp. 130–137, 2017, doi: 10.22161/ijaers.4.11.20.

[9] H. Moeeni, H. Bonakdari, and I. Ebtehaj, "Integrated SARIMA with Neuro-Fuzzy Systems and Neural Networks for Monthly Inflow Prediction," *Water Resour. Manag.*, vol. 31, no. 7, pp. 2141–2156, 2017, doi: 10.1007/s11269-017-1632-7.

[10] N. S. Arunraj, D. Ahrens, and M. Fernandes, "Application of SARIMAX Model to Forecast Daily Sales in Food Retail Industry," *Int. J. Oper. Res. Inf. Syst.*, vol. 7, no. 2, pp. 1–21, 2016, doi: 10.4018/ijoris.2016040101.

[11] P. Sutthichaimethee and D. Ariyasajjakorn, "Forecasting energy consumption in short-term and long-term period by using ARIMAX Model in the construction and materials sector in Thailand," *J. Ecol. Eng.*, vol. 18, no. 4, pp. 52–59, 2017, doi: 10.12911/22998993/74396.

[12] M. F. Ani, S. R. Kamat, and M. Fukumi, "Development of Decision Support System via Ergonomics Approach for Driving Fatigue Detection," vol. 1, no. 1, pp. 60–72, 2020.

[13] H. Khan and P. Pohwani, "Testing Phillips Curve in Pakistan," *J. Public Value Adm. Insight*, vol. 3, no. 3, pp. 145–152, 2020.

[14] X. Ying, "An Overview of Overfitting and its Solutions," *J. Phys. Conf. Ser.*, vol. 1168, no. 2, 2019, doi: 10.1088/1742-6596/1168/2/022022.

[15] F. Rahmi Ras, H. Nelly Astuti, and B. Efori, "Perancangan Sistem Pakar Diagnosa Penyakit Asidosis Tubulus Renalis Menggunakan Metode Certainty Factor Dengan Penelusuran Forward Chaining," *Media Inform. Budidarma*, vol. 1, no. 1, pp. 13–16, 2017.