Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna

Activity 8			
Topic:	**Topic 8: Time Series Analysis**	**Week No.**	9
Course Code:	**CSST104**	**Term:**	2nd Semester
Course Title:	**Machine Learning**	**Academic Year:**	2023-2024
Student Name		**Section**	
Due date	**March 27, 2024 12:00 PM**	**Points**	

Assessment Task: Pollution Data Time Series Analysis Using ARIMA

Objective:
The goal of this assessment is to simulate a time series dataset representing a country's annual pollution levels over a decade. You will apply time series analysis techniques, including the ARIMA (Autoregressive Integrated Moving Average) model, to understand the data's behavior, test for stationarity, and forecast future pollution levels.

Dataset Simulation:
Simulate a time series dataset for a country's annual pollution levels over 10 years. Assume a trend component, a seasonal component, and some random noise in the data.

The dataset "most-polluted-countries.csv" contains several columns related to pollution and country characteristics for the year 2023. Here are some key columns in the dataset:

- **pollution_2023:** Pollution figures for the year 2023.
- **pollution_growth_Rate:** The growth rate of pollution from the previous year.
- **country_name:** Name of the country.
- **country_region:** The region where the country is located.
- **united_nation_Member:** Whether the country is a member of the United Nations.
- **country_land_Area_in_Km:** The land area of the country in square kilometers.
- **pollution_density_in_km:** Pollution density per square kilometer.
- **pollution_density_per_Mile:** Pollution density per square mile.
- **share_borders:** Countries with which the country shares its borders.
- **pollution_Rank:** The rank of the country based on its pollution.
- **mostPollutedCountries_particlePollution:** Particle pollution levels in the country.

For a time series analysis using the ARIMA model, we typically need a time series data column (e.g., dates or years) and a numerical value column to predict (e.g., pollution figures over time for a specific country).

Tasks:

1. **Dataset Preparation:**
    - Simulate the time series data with a clear trend and seasonality to reflect hypothetical annual pollution levels.
    - Plot the time series to visualize the trend and seasonality.

2. **Stationarity Testing:**
    - Perform a stationarity test (e.g., Augmented Dickey-Fuller test) to check if the time series is stationary.
    - Discuss the implications of the test results for time series analysis.

3. **ARIMA Model Identification:**
    - Use plots (e.g., autocorrelation and partial autocorrelation plots) to identify the ARIMA model parameters (p, d, q).
    - Explain your choice of parameters.

4. **ARIMA Model Fitting:**
    - Fit an ARIMA model to the simulated data using the identified parameters.
    - Evaluate the model's fit and discuss any adjustments needed based on diagnostics plots or performance metrics.

5. **Forecasting:**
    - Use the fitted ARIMA model to forecast pollution levels for the next 2 years.
    - Plot the forecast along with a confidence interval to visualize the expected future values and the uncertainty.

6. **Report and Insights:**
    - Provide a detailed report of the analysis process, model fitting, and forecasting results.
    - Discuss the potential real-world implications of your findings and how they could inform policy or decision-making related to environmental management.

Deliverables:

- A Jupyter Notebook containing the Python code and analysis for each task.
- A final report summarizing the methodology, findings, and insights from the time series analysis.

Evaluation Criteria:

- Accuracy of the simulated time series data.
- Correct application of stationarity tests and interpretation of results.
- Appropriate identification and justification of ARIMA model parameters.

- Quality of the ARIMA model fit and the rationale for any model adjustments.
- Clarity and accuracy of the forecast, including the interpretation of confidence intervals.
- Depth of analysis and insights provided in the final report.

This task aims to assess proficiency in time series analysis techniques, model identification and fitting, and the ability to interpret and communicate findings effectively. It offers a practical understanding of how ARIMA models can be applied to environmental data, even when actual time series data are not available for analysis.

Submission Instruction:

- Share the Google Collab Activity to markbernardino@lspu.edu.ph
- Filename Format: **2A-BERNARDINO-EXER8**

Inability to follow this instruction will be deducted 5 points each for filename format and late submission per day. Also, cheating and plagiarism will be penalized.