WQD7001
Principles of Data Science

Group Assignment
Prediction of Water Usage in Malaysia

Group 5

Name	Student ID	Roles
Rahayu Rianti	23104958	Group Leader
Arun Kumar	23082194	Maker
Siti Nur Liyana Binti Roslan	23067122	Detective
Karenina Kamila	23117951	Oracle
Looi Xue Ying	24053855	Detective

1. Project Background

The rapid urbanization and population growth in Malaysia have led to increasing pressures on water resources, resulting in heightened water demand across various states. These changes present challenges such as water scarcity, infrastructure limitations, and inefficiencies in resource allocation. To address these issues, there is a critical need for predictive modelling of water usage based on population data.

According to WHO (2024), the current population of Malaysia in 2023 is around 33.40 million and is projected to grow to 44.3 million by 2050 which is an increase by 10.9%. With the assumption of steady population growth along with the growth of the industrial sector and economic development, the demand for water is expected to increase as well for the next few decades, therefore, coming out with a long-term plan in ensuring sustainable water usage is required.

The citizens and government of Malaysia are our target users in obtaining the data since we are evaluating the water usage rate of this country. Therefore, this project is suitable for government and private sectors of policymakers such as the government, water industry and even NGOs to get insight into water consumption in Malaysia before making a data-driven decision.

The potential benefits of this project is that we are able to provide insight by making predictions of water usage in Malaysia so that the stakeholders are able to come out with a plan or policy to ensure the water is sustainably used to avoid water poverty in the long run. We had specifically chosen the environment as our domain because this project is aligned with the United Nations' Sustainable Development Goals (SDG) 6 which is about 'Clean Water and Sanitation'. Our country of focus is Malaysia, where multiple datasets such as population, water consumption, water production and water access were used. Thus, this plan is beneficial to everyone, as water is the source of life and the main component of sustainable development on this planet.

2. Project Objectives

I. To analyze the water usage against the population in Malaysia.
II. To build predictive models of water consumption in Malaysia.
III. To evaluate the accuracy of the prediction models used.

3. Exploratory Data Analysis (EDA)

First and foremost, all the datasets used were obtained from Malaysia Open Data Source (OpenDOSM). The types of data collected are structured data which includes data for population, population by states, water consumption, water production and water access. Since the source of data is obtained directly from the government official database, we can assure that the datasets are reliable and are of high quality for us to process and use for this project. The datasets are downloaded and kept as a csv (comma-separated value) file.

Next step, we explored and quality checked on the datasets to decide which attributes and how much data we were planning to use. As for all datasets, those have different period ranges. Then, we decided to use the same data range between 2003 and 2022 by reducing the amount of data. As for population data, we used only "overall" data for age and ethnicity attributes and "both" data for gender attributes. Meanwhile, the remaining dataset with spatial context will be maintained since all the attributes are useful for our project. Furthermore, we also check for missing values in the dataset, the gaps or interpolation and any data duplicity. Next, we also add derived attributes into the dataset which are growth rate and water consumption per capita as part of the pre-processing. From this, we transform the dataset by merging multiple datasets into a new dataset using year as foreign key. All the processes were being made on Google Colab using the Python programming language.

Then, we did Exploratory Data Analysis (EDA) to investigate the data in terms of anomalies, and to check assumptions using statistics and graphical representations. Here are some insight that we get by doing EDA:

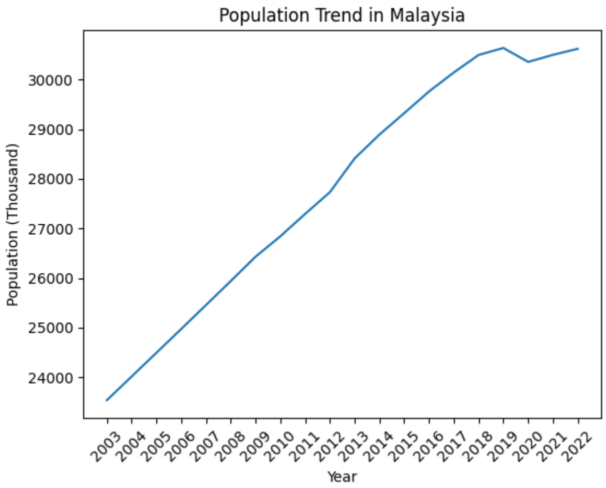


Figure 1. Population in Malaysia Trends

Over the years, Malaysia's population has increased from 2003 to 2019. However, there is a slight decline in 2020 during Covid-19 also potentially decreasing in birth-rate. In the following years, the population increased slightly.

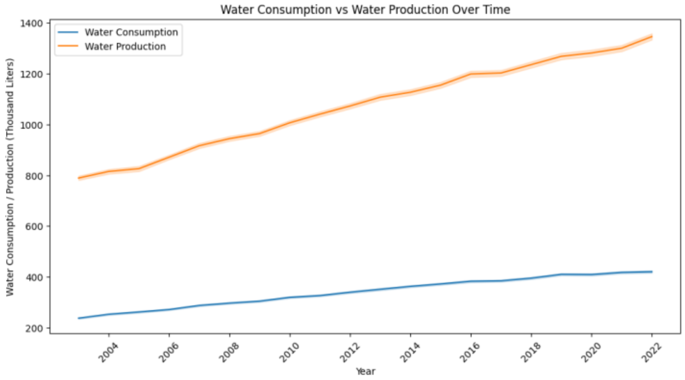


Figure 2. Water Consumption against Water Production line chart

There are gaps between water consumption and production, where production grows steadily while consumption remains comparatively lower, suggesting inefficiencies or underutilization in water distribution relative to population needs.

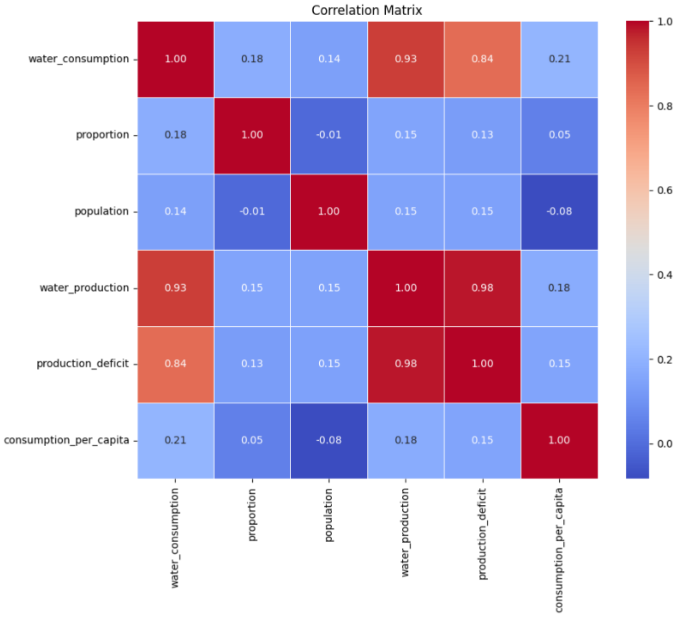


Figure 3. Correlation Matrix between all variable

The most significant relationships are between water consumption, water production, and production deficit. As water consumption goes up, production must increase to prevent a larger deficit.

4. Machine Learning Modelling

In this section, we discuss the machine learning models used to predict water consumption trends in Malaysia. The selection of models was guided by the need to capture diverse data patterns, including linear relationships, non-linear interactions, time-series dependencies, and long-term trends. By comparing the performance of multiple models, we aim to identify the most suitable one for this use case.

A. Linear Regression

Linear Regression is a statistical approach used to model the relationship between independent variables (e.g. population, water production) and a dependent variable (water consumption). In this project, Linear Regression was chosen due to its simplicity and effectiveness in identifying linear patterns in the data.

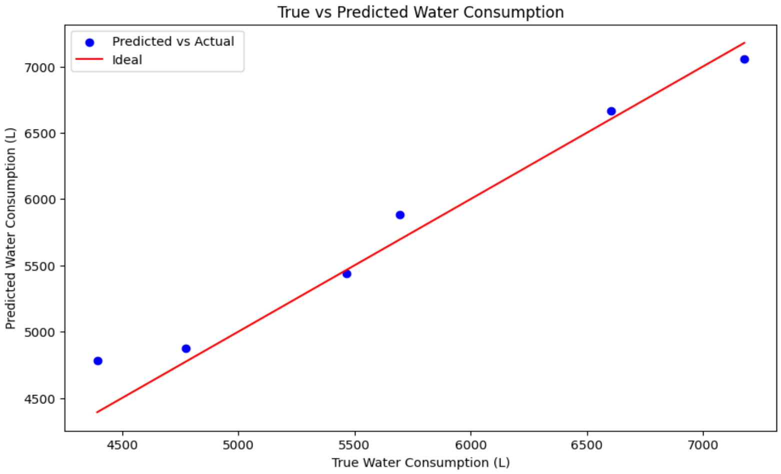


Figure 4. True vs Predicted Water Consumption for Linear Regression

B. Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and combines their outputs to improve predictive accuracy and control overfitting. For this project, Random Forest was applied to evaluate its capability to capture non-linear relationships between predictors and water consumption.

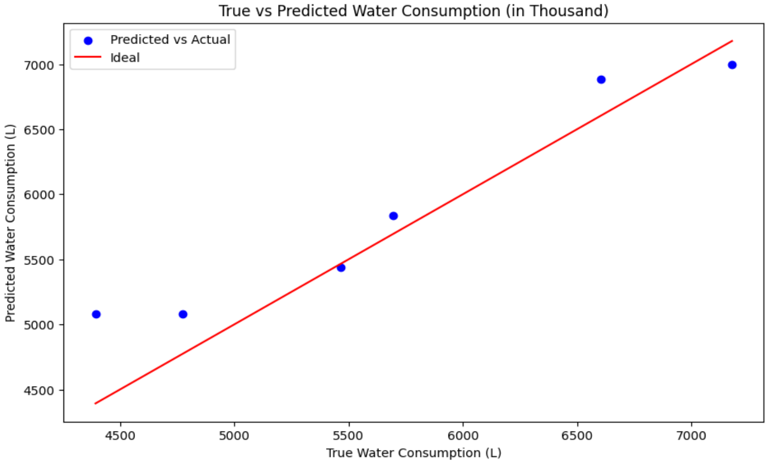


Figure 5. True vs Predicted Water Consumption for Random Forest

C. Long short-term memory (LSTM)

LSTM is a type of recurrent neural network (RNN) designed to handle sequential data and capture long-term dependencies. In this project, LSTM was applied to model the temporal patterns in water consumption.

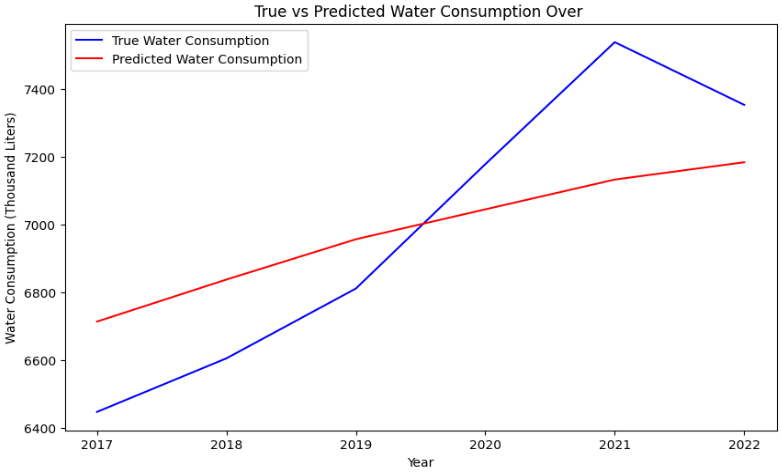


Figure 6. True vs Predicted Water Consumption for LSTM

D. Autoregressive Integrated Moving Average (ARIMA)

ARIMA is a popular time-series forecasting model that captures trends and seasonality in univariate data. For this project, ARIMA was used to evaluate its suitability for water consumption forecasting.

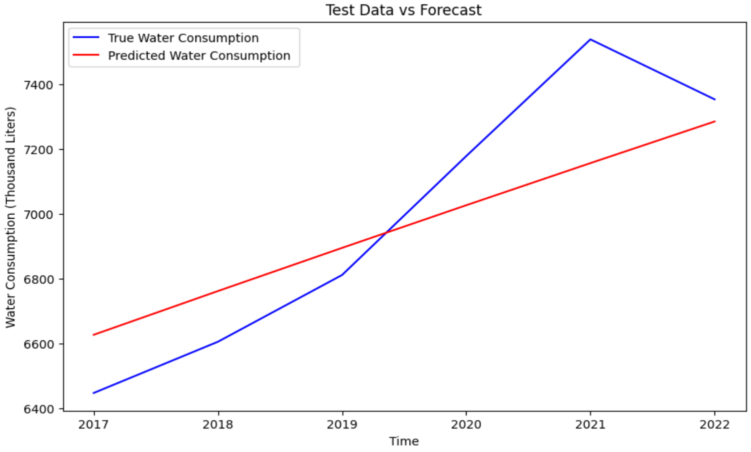


Figure 7. True vs Predicted Water Consumption for ARIMA

5. Machine Learning Model Evaluation

The evaluation of the models is based on four metrics:

- Mean Absolute Error (MAE): Average magnitude of errors, where lower values are better.
- Mean Squared Error (MSE): Penalizes larger errors more, where lower values are better.
- Root Mean Squared Error (RMSE): The square root of MSE, providing error magnitudes in the same unit as the dependent variable. Lower values are better.
- R² (Coefficient of Determination): Explains the variance (accuracy) in the data; values closer to 1 indicate better fit.

	Mean Absolute Error(MBE)	**Mean Squared Error(MSE)**	**Root Mean Squared Error(RMSE)**	**R-squared**
Linear Regression	149.341	36743.342	191.686	0.961
Random Forest	271.795	117605.461	342.937	0.875
LSTM	224.084	58806.312	242.500	0.627
ARIMA	170.251	39524.914	198.809	0.748

a. Linear Regression:

This model showcases exceptional predictive accuracy for water usage in Malaysia. The low error metrics, which indicate that the model consistently produces predictions close to the actual data points (Plevris et al., 2022). The model also explains 96.1% of the variance in water consumption, highlighting its strong ability to capture the relationship between population, water production and water consumption. This performance suggests that the underlying patterns in water consumption are largely linear, making Linear Regression a highly suitable choice for this dataset. For instance, an increase in population likely results in a significant increase in water consumption, which is effectively captured by the model.

Limitations: This model assumes a linear relationship between population, water production and water consumption, which may fail to capture more complex, non-linear patterns or events like sudden policy changes or global pandemic outbreaks.

Recommendations: Requires regular assessment of its performance, especially if the nature of the predictors or water consumption patterns evolves over time.

b. Random Forest:

This model produced relatively high error metrics, indicating poorer predictive performance compared to other models (Plevris et al., 2022). The value of R^2 (0.875) explains only 87.5% of the variance, which is quite high and in contrast to the value obtained from the model's error metrics. Its performance here suggests that it may not effectively handle the structure of this dataset.

Limitations: There is a possible overfitting, where the model focuses too closely on the training data, resulting in reduced generalizability to unseen data (Bejani and Ghatee, 2021). Moreover, the relatively high error metrics suggest that the model might be overly sensitive to noise or irrelevant features in the dataset (Montesinos López et al., 2022). This could be a result of lacking optimal hyperparameter tuning.

Recommendations: This model can be improved further by incorporating additional predictors, and optimizing hyperparameters such as tree depth, number of estimators, or minimum leaf size. In addition, applying hybrid approaches that combine Random Forest with models adept at time-series prediction, such as ARIMA or LSTM, could yield improved results.

c. LSTM (Long Short-Term Memory):

The LSTM model produces intermediate MAE (224.084) and RMSE (242.500), with moderate MSE (58806.312) values with low R^2 (0.627), which suggests the model also struggles to explain the variance in the data.

Limitations: This might be the model's weakness, being a data hungry model that requires large datasets to achieve high accuracy (Faheem, 2022). Besides, this model is also highly sensitive where the model requires meticulous tuning of hyperparameters, while improper configurations can lead to suboptimal results.

Recommendations: Its weaker performance in this case may be attributed to insufficient data, inadequate preprocessing, or training inefficiencies that were reflected with the relatively low R^2, which indicates it failed to fully capture patterns in the dataset. With optimal implementation such as providing larger datasets, refined preprocessing, and careful tuning, the LSTM could outperform other models, especially for datasets with complex temporal patterns and dependencies.

d. ARIMA (Auto Regressive Integrated Moving Average):

This model demonstrated strong performance with the lowest MAE (170.251) and RMSE (198.809), and relatively low MSE (39,524.914). It indicates that the model is suitable for univariate time-series data, which highlights ARIMA's effectiveness in capturing trends and seasonality in univariate time-series data (Kontopoulou et al., 2023). However, its moderate R^2 value (0.748) indicates that it explains 74.8% of the variance, which is lower than Linear Regression and Random Forest.

Limitations: ARIMA is unable to handle external predictors using population and water consumption data, unlike other models (Elsaraiti and Merabet, 2021). While it excels in short-term forecasting and univariate time-series analysis but struggles with complex, non-linear relationships or external predictors. This makes it less versatile compared to models like Linear Regression, which can leverage additional predictors to enhance accuracy.

Recommendations: Using hybrid approaches that combine other machine learning models such as SARIMA or ARIMAX could yield improved results as the models further enhanced ARIMA due to the additional parameters added.

To summarize our findings, the Linear Regression model is the best choice for the regression modelling case as the model had produced high accuracy value while LSTM performed the worst. For scenarios where error metrics such as MAE and RMSE are prioritized, ARIMA is an excellent alternative compared to Random Forest. Although the Random Forest model had achieved higher accuracy than ARIMA, the model produced relatively high error metrics, which makes the accuracy produced less reliable.

6. Deployment of Data Product

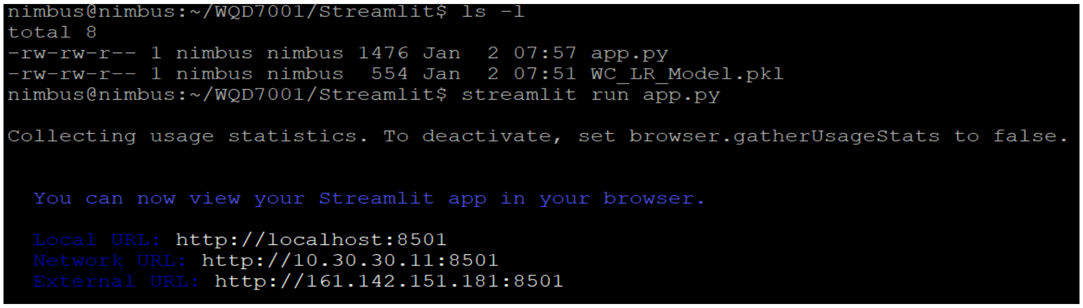
The trained machine learning models were evaluated based on Malaysia's population data (X1) and water production data (X2), to predict the target value of water consumption (Y). In this project, we have chosen two features to predict the water consumption. For that we have evaluated Linear Regression, Random Forest, LSTM and ARIMA. But to predict the water consumption based on two features, hence we had chosen the best performing model which is Linear Regression.

A. Store the ML Model

The trained Linear Regression model will be saved as a model file in a binary format which is called as pickle file(.pkl) format using the “pickle” library in python. The saved model will be stored in a central model repository, where we can version the models and keep track of the performance, so that we can deploy the model to the appropriate model for the prediction purpose at different environments such as User Acceptance Test (UAT) and Production.

B. Deploy the ML Model

The specific version of the model from a central repository will be fetched over the secured network into the model prediction environment. Using the python framework, the model will be loaded into the prediction application and made to be ready for inference, in which it does prediction based on the input values X1 and X2.

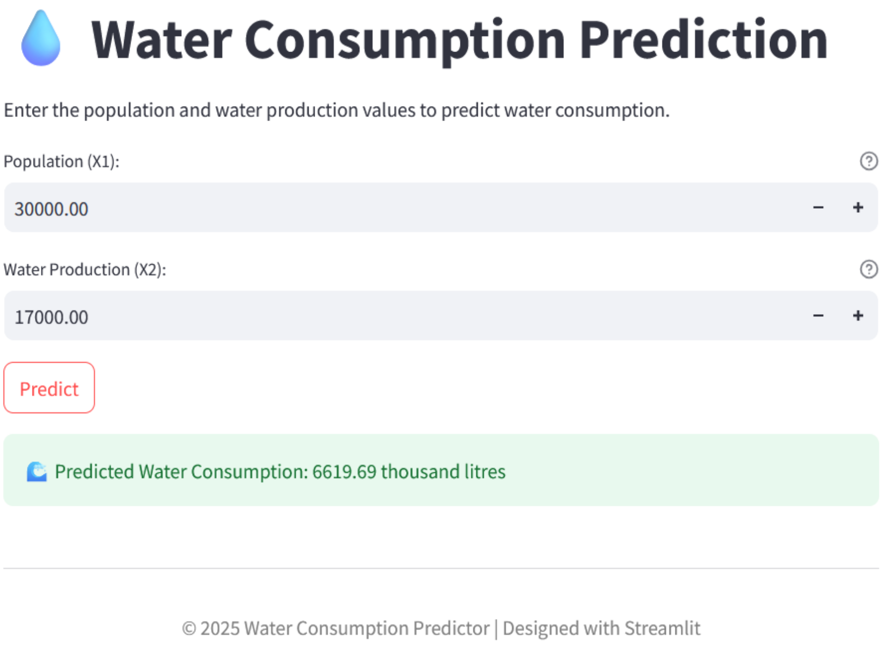


```
nimbus@nimbus:~/WQD7001/Streamlit$ ls -l
total 8
-rw-rw-r-- 1 nimbus nimbus 1476 Jan  2 07:57 app.py
-rw-rw-r-- 1 nimbus nimbus  554 Jan  2 07:51 WC_LR_Model.pkl
nimbus@nimbus:~/WQD7001/Streamlit$ streamlit run app.py

Collecting usage statistics. To deactivate, set browser.gatherUsageStats to false.


  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://10.30.30.11:8501
  External URL: http://161.142.151.181:8501
```

Figure 8. Deploy the ML Model

C. Data Product using Machine Learning Model

The data product we developed is a web based UI application which takes input values such as population data (X1) and water production data (X2) from users and outputs the predicted (Y) value as water consumption. Here is the screenshot of our data product,



Water Consumption Prediction

Enter the population and water production values to predict water consumption.

Population (X1): 30000.00

Water Production (X2): 17000.00

Predicted Water Consumption: 6619.69 thousand litres

© 2025 Water Consumption Predictor | Designed with Streamlit

Figure 9. Data Product Web Based UI Application

- **Input**
  - Population (X1) = 30,000,000 people
  - Water Production (X2) = 1700000 litres
- **Output**
  - Water Consumption (Y) = 6619.69 thousand litres

7. Plan for Reproducible Research

The current problem about science in general is that many projects cannot be reproduced. The scientific result can be accepted too if it has undergone a peer review, and transparent process.

It is the same that the lack of details about the data analysis in particular can make it impossible to recreate any of the results presented in a paper (Peng, 2011). To address this communication problem, a concept has emerged known as reproducible research. In the computational era, there are additional benefits to doing it. In addition to being able to fully understand the process by which the results were obtained, readers also have access to the data and the computer code, both of which are valuable to the extent that they can be reused or repurposed for future studies or research. Another goal of reproducible research is to provide a kind of audit trail, should it be needed (Peng, 2021).

As for data science projects, adopting reproducible research can be done by sharing the measured data, analytic data, computational results, presentation code and articles. Regarding this, we decide to use GitHub as an implementation adopting reproducible research. Considering few reasons such as:

a. GitHub is open to the public (transparent).
b. GitHub has features to save the data, code, and article/result (communicating the result).
c. GitHub has many users, it is possible to get independent reviews for this project.
d. GitHub provides tracking changes.

The project code repository in GitHub as follows: https://github.com/kareninakamila/Data-Science_Water-Consumption-Prediction-in-Malaysia

8. Conclusion

Following the OSEMN phase and other parts of the project, we can conclude that datasets are easy to get and clean, with no missing value and no duplication. Moreover, from other phase we can get insights:

1. Exploring:
    a. Overall Malaysia's population has increased year by year except in Covid-19 (2019-2020) period.
    b. Overall Malaysia’s water production has increased year by year.
    c. Overall Malaysia’s water consumption has increased. There was a slight decrease in domestic sector usage in 2017 and 2022. In the non-domestic sector the slight decrease was 2020 and 2021 during Covid-19.
    d. Total Malaysia’s water production is higher than water consumption, however it suggests inefficiencies in water distribution.
    e. Correlation matrix shows there is a significant relationship between water production and consumption, followed up by production deficit. As for the population, it has a weak relationship with those variables.
2. Modeling:
    a. To predict water consumption, we used 4 (four) models such as Linear Regression, Random Forest, LSTM, and ARIMA.
    b. From model evaluation results, we are able to provide recommendations for optimizing water resource allocation through prioritization based on variance insights. The R^2 scores indicate how well the models capture variability in the data. Linear Regression is the model with highest R^2 (0.961) value that is better suited for identifying key predictors influencing water usage, enabling resource prioritization for high-demand regions. Other than that, we can utilize ARIMA for forecasting short-term water demand in regions with consistent patterns. This can help in proactive planning of water distribution, especially during peak demand periods.
3. Other phases:
    a. To build the data product, we used a web based UI application as input to predict water consumption in Malaysia. As for the process, the trained Linear Regression model is being saved into a binary file and then deployed to the local production environment which has web prediction for water consumption.
    b. The reproducible research for data science can be achieved using GitHub.

In conclusion and overall, we are able to achieve all of our project objectives with code and explanations available in Google Colab. As well as, we are able to build the prediction of a water consumption web. However, there are few things can be improved or added which are:

a. Adding a new objective to explore and predict water consumption based on state or other variables in more granular ways.
b. Adding and using available automation features to improve efficiency, reduce error and streamline processes.
c. Launching the prediction water consumption website publicly.
d. Adding another way to reproduce research by using the open science platform.
e. Adding prescriptive analytics to give more recommendations and solutions for stakeholders.

9. References

Bejani, M. M., & Ghatee, M. (2021). A systematic review on overfitting control in shallow and deep neural networks. *Artificial Intelligence Review*, *54*(8), 6391-6438.

Elsaraiti, M., & Merabet, A. (2021). A comparative analysis of the arima and lstm predictive models and their effectiveness for predicting wind speed. *Energies*, *14*(20), 6782.

Faheem, M., Aslam, M., & Kakolu, S. (2022). Artificial intelligence in investment portfolio optimization: A comparative study of machine learning algorithms. *International Journal of Science and Research Archive*, *6*(1), 335-342.

Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., & Matsopoulos, G. K. (2023). A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks. *Future Internet*, *15*(8), 255.

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Overfitting, model tuning, and evaluation of prediction performance. In *Multivariate statistical machine learning methods for genomic prediction* (pp. 109-139). Cham: Springer International Publishing.

Peng, R. D. (2011). Reproducible research in computational science. Science 334(6060),1226–27

Peng, R. D., & Hicks, S. C. (2021). Reproducible research: A retrospective. Annual Review of Public Health, 42(1), 79–93.

Plevris, V., Solorzano, G., Bakas, N. P., & Ben Seghier, M. E. A. (2022). Investigation of performance metrics in regression analysis and machine learning-based prediction models. In *8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS Congress 2022)*. European Community on Computational Methods in Applied Sciences.

World Health Organization. (2024). *Country Overview: Malaysia.* World Health Organization.