WIA1006/WID3006 MACHINE LEARNING

Prediction of Water Volume Changes

The Role of Weather and Lake Data

A PROJECT BY FIRST YEAR STUDENTS FROM FSKTM, UNIVERSITY OF MALAYA

Machine Learning

HydroDS

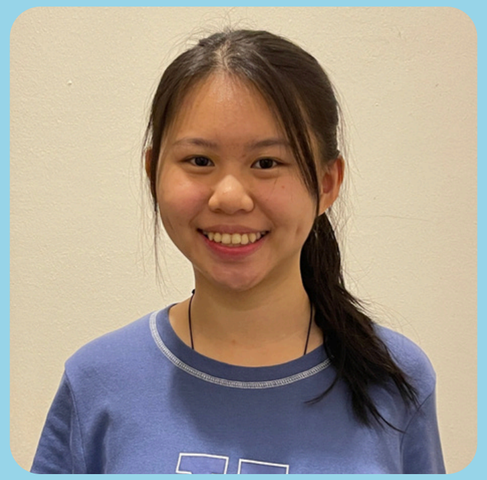
GROUP 15

Team member

NG YU HENG
23054206

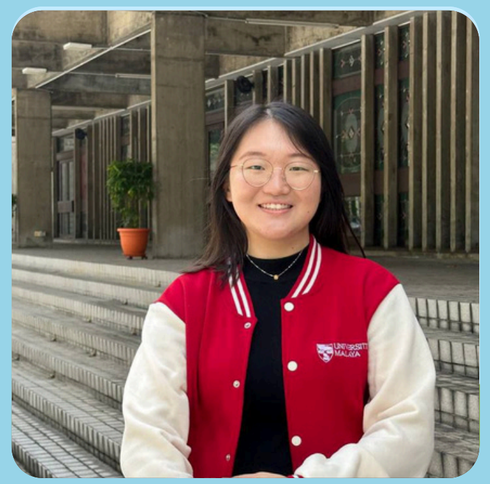
CHANG CAI TING
23005031

CHUA HUI MIN
23005000



KUEH PANG LANG
23005227

NG ZHI WEI
23051966



POH JING MIN
23004979

Part I

Background

Background to the problem

- Weather has been one of the major factors affecting the water volume of a water reservoir / lake / dam
- Fail to take effective action when sudden changes of weather conditions
- Fail to have an effective long term plan when dealing with climate changes

www.todayonline.com

Four dead after flood at Cameron Highlands dam

CAMERON HIGHLANDS — Three foreigners have died after the gates of the Sultan Abu Bakar hydro-electric dam in Cameron Highlands were opened...

24 Oct 2013

The Star

Water in Selangor dams can last for six months with little, no rain

SELANGOR'S seven dams are at optimum level and water supply can last for six months with intermittent rain or no rainfall, says Selangor...

9 Mar 2024

Citizen Journalist Malaysia

Air Itam Dam low on water, PBAPP urges conservation

Penang's Air Itam Dam faces low water levels but assures residents of no immediate shortages. Rain expected soon & water management plans in place.

1 month ago

Solution to the problems

- **Water volume changes prediction for unusual weather condition**
- Water volume changes forecasting for climate changes

Background to the datasets

Lake dataset

Method: Web Scrapping

Website:

Watershed Connection

https://www.watershedconnection.com/

No suitable lake datasets in Malaysia. Using lake data outside of Malaysia as an alternative.

Location:

Phoenix,
Arizona, United States

Number of sample: 36530

Number of features: 13

Weather dataset

Method: Website API

Website:

visualcrossing

https://www.visualcrossing.com/

Getting weather data with high accuracy from reliable weather website helps improve prediction results.

Location:

Falcon Field Airport,
4800 E Falcon Dr, Mesa, AZ 85215, United States

Number of sample: 3662

Number of features: 33

Part II

Data Preprocessing

Data Preprocessing

- Handling missing values
- changing data types
- removing duplicates

DATA CLEANING

- Removing of data
- Removing outlier

HANDLING MISSING DATA

- Combine weather dataset and lake datasets by using similar columns "datetime"

DATA INTEGRATION

- Replacing Categorical Data with Numbers
- Label Encoding
- Transforming Data of Different Scale

DATA TRANSFORMATION

Part III

Explotary Data Analysis

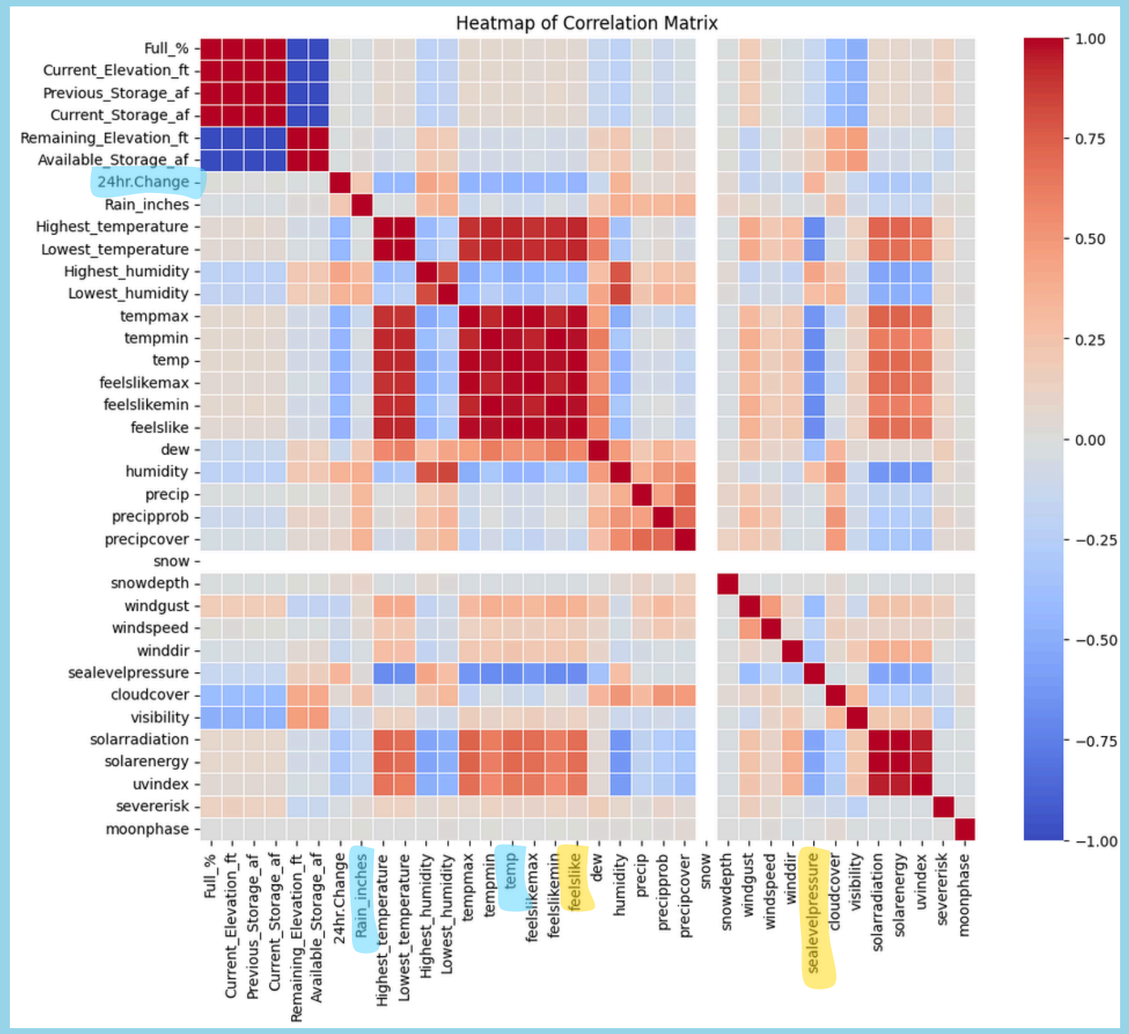
Feature selection

MOST IMPORTANT FEATURE

- TEMPERATURE
- RAIN_INCHES

FEATURE THAT SUPRISINGLY IMPORTANT

- FEELSLIKE
- SEALEVELPRESSURE



Comfirmatory Data Analysis

- **Hypothesis**: Higher temperatures lead to greater evaporation, which decreases the water volume.
- **Question**: Is there a significant relationship between temperature variables (highest, lowest, average) and the 24-hour change in water volume?
- **Hypothesis**: Heavy rainfall in the past 24 hours significantly increases the lake's water volume.
- **Question**: How does the amount of rainfall correlate with the 24-hour change in water volume?
- **Hypothesis**: Higher wind speeds and solar radiation contribute to increased evaporation rates, decreasing the water volume.
- **Question**: How do wind speed and solar radiation affect the 24-hour change in water volume?

Exploratory Data Analysis

- Are there any **missing values** in the dataset, and how should they be handled?
- What are the **characteristics of the outliers** in the 24-hour change data, and how do they affect the overall analysis?
- What are the **correlation patterns** between different features, especially weather-related variables and the 24-hour change in water volume?
- What are the **distributions of key numerical features**, and are there any skewness or outliers?

Part IV

Model Result

Result Summary

		MSE	RMSE	R2 score
Top 1	**Random Forest Regression**	**0.009902**	**0.099508**	**0.690128**
Top 2	Gaussian Process Regression	0.012112	0.110054	0.620963
Top 3	Decision Tree Regression	0.012351	0.111134	0.613483
	Support Vector Regression	0.013088	0.111723	0.590403
	Polynomial Regression	0.013761	0.117309	0.569341
	Neural Network Regression	0.014184	0.119098	0.556102
	Multiple linear Regression	0.016411	0.128105	0.486425

Result Summary

	MSE	RMSE	R2 score

Comparison with ensemble model by auto sklearn

Better	**Auto sklearn (ensemble)**	**0.008574**	**0.092598**	**0.731670**
	Random Forest Regression	0.009902	0.099508	0.690128

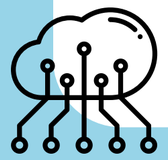
Comparison with single model by auto sklearn

Better	**Random Forest Regression**	**0.009902**	**0.099508**	**0.690128**
	Auto sklearn (single)	0.010473	0.102339	0.672243

Conclusion

- **Random Forest Regression** performed the **best** in predicting the water volume changes in 24 hours.

- An **ensemble model** helps to **produce better prediction results** compared to using a single model.



- Ensuring **data quality**, performing **feature engineering** significantly influenced model performance.

- Machine learning able to help in **optimizing natural resource management** and **preventing disaster**.

- Although prediction helps people to make better decision, individual judgements is required since **prediction is not always right.**

Think Like a Hydrologist!

Thank you

A hydrologist studies how water moves underground and on the Earth's surface.

Machine Learning

HydroDS

GROUP 15