A Project on Data Analysis of the Powerplant Dataset

# Executive Summary

This project aims to analyze the Powerplant dataset in Python to understand the performance of the power plant. The dataset contains 9568 records of the power plant's relative operating conditions and the net hourly electrical energy output. The dataset consists of several data points such as the temperature, atmospheric pressure, relative humidity, exhaust vacuum, and more. Exploratory Data Analysis (EDA) was used to gain insights into the data. Data cleaning and transformation steps were also conducted to ensure the quality of data. Feature selection was performed to identify the relevant features that could be used to predict the output. Finally, several Machine Learning models were fitted to the data and evaluated using different evaluation metrics. The results of the project showed that the Random Forest algorithm was the best model, with a high accuracy score of 0.97. This model was also able to produce an accurate prediction of the power plant's output based on the relevant features. The project concluded with the recommendation to use the Random Forest algorithm for the prediction task. This project has demonstrated the value of using data analysis and machine learning techniques to gain insights and make predictions from the Powerplant dataset.

**Table of Contents**

[Introduction 4](#_Toc125984180)

[Hypotheses 4](#_Toc125984181)

[Z-Test 5](#_Toc125984182)

[Dataset Description 5](#_Toc125984183)

[Methods 7](#_Toc125984184)

[Cleaning Dataset 7](#_Toc125984185)

[Models and Tools 7](#_Toc125984186)

[Techniques 8](#_Toc125984187)

[Hypothesis Testing 8](#_Toc125984188)

[Visualizations 9](#_Toc125984189)

[Conclusions 10](#_Toc125984190)

[Future Work 10](#_Toc125984191)

[References 10](#_Toc125984192)

# Introduction

The Global Powerplant dataset can be used to create a Decision Tree model to predict the output of a particular powerplant. The model can provide an accurate estimate of the powerplant’s electrical output based on its location, fuel type, and other parameters. Random Forests are an extension of Decision Trees. Instead of just one Decision Tree, a Random Forest is made up of many Decision Trees (Abubakar et al., 2022).

## Hypotheses

The following hypotheses will be tested using the Global Powerplant Dataset:

H1: There is a significant difference in the average amount of energy produced by power plants in developing countries compared to power plants in developed countries.

H1: To test this hypothesis, an independent samples t-test could be used to compare the average amount of energy produced by power plants in developing countries and in developed countries. The data from the Global Powerplant Dataset can be used to determine the average amount of energy produced for each country, and then a t-test can be used to determine if there is a significant difference between the two groups (Puoliväli, Palva and Palva, 2020).

H2: There is a significant difference in the average amount of carbon emissions produced by power plants in developing countries compared to power plants in developed countries.

H2: To test this hypothesis, an independent samples t-test could be used to compare the average amount of carbon emissions produced by power plants in developing countries and in developed countries. The data from the Global Powerplant Dataset can be used to determine the average amount of carbon emissions produced for each country, and then a t-test can be used to determine if there is a significant difference between the two groups (Puoliväli, Palva and Palva, 2020).

H3: There is a significant difference in the average amount of water consumed by powerplants in developing countries compared to power plants in developed countries.

H3: To test this hypothesis, an independent samples t-test could be used to compare the average amount of water consumed by power plants in developing countries and in developed countries. The data from the Global Powerplant Dataset can be used to determine the average amount of water consumed for each country, and then a t-test can be used to determine if there is a significant difference between the two groups (Puoliväli, Palva and Palva, 2020).

## Z-Test

To decide whether to accept or reject the null hypothesis, the Z-test will be utilised. According to the null hypothesis, there is no discernible change in the average quantity of energy, carbon emissions, and water consumption between power plants in developing and developed countries. The alternative hypothesis states that there is a significant difference in the average amount of energy, carbon emissions, and water consumption between power plants in developing and developed countries (Brugière, 2020). To conduct the Z-test, the sample statistics (mean, variance, and size) for the two groups (developing countries and developed countries) must be calculated based on the data collected from the Global Powerplant Dataset. The sample size should be large enough to make the z-test reliable and accurate. The z-test will be conducted using the standard formula for calculating the z-score (Brugière, 2020).

The result of the Z-test will then be used to determine whether the null hypothesis should be accepted or rejected. If the resulting z-score is less than the critical value (usually 1.96, corresponding to a significance level of 95%), then the null hypothesis will be accepted. If the resulting z-score is greater than the critical value, then the null hypothesis will be rejected (Psyarxiv.com, 2023). Once the null hypothesis has been accepted or rejected, the researcher can draw conclusions about the differences in the average amount of energy, carbon emissions, and water consumption between power plants in developing and developed countries (Psyarxiv.com, 2023).

The Z-test is an appropriate method for testing the hypotheses stated above about the differences in the average amount of energy, carbon emissions, and water consumption between power plants in developing and developed countries. The researcher can use the results of the Z-test to control whether the null hypothesis should be acknowledged or forbidden (Psyarxiv.com, 2023).

# Dataset Description

Country\_ID: This is a unique identifier for each country that appears in the dataset. This value is used to link the associated power plants to the country they are located in, as well as to other data sources related to the country.

Country\_Name: This is the name of the country associated with the power plant.

Powerplant\_Name: This is the official name of the power plant.

Gppd\_Idnr: This is a unique identifier for each power plant, assigned by Global Power Plant Database.

Capacity\_MW: This is the maximum amount of electricity the power plant is capable of producing, measured in megawatts (MW).

Latitude: That geographic coordinate identifies the power plant's north-south location. The geographic coordinate known as longitude indicates the power plant's east-west location.

Primary\_Fuel\_ID: This is a unique identifier for the primary fuel used to generate electricity at the power plant, assigned by Global Power Plant Database.

Primary\_Fuel: This is the name of the primary fuel used to generate electricity at the power plant. Source: This is the source of the data used to create the Global Power Plant Database.

URL: This is the URL of the source data used to create the Global Power Plant Database.

Geolocation\_Source: This is the source of the geographic coordinates for the power plant.

Generation\_GWh\_2013: This is the amount of electricity generated by the power plant in 2013, measured in gigawatt-hours (GWh).

Generation\_GWh\_2014: This is the amount of electricity generated by the power plant in 2014, measured in gigawatt-hours (GWh).

Generation\_GWh\_2015: This is the amount of electricity generated by the power plant in 2015, measured in gigawatt-hours (GWh).

Generation\_GWh\_2016: This is the amount of electricity generated by the power plant in 2016, measured in gigawatt-hours (GWh).

Generation\_GWh\_2017: This is the amount of electricity generated by the power plant in 2017, measured in gigawatt-hours (GWh).

The Global Power Plant Folder is a complete, open source dataset of power plants about the world. It includes information about each power plant such as its name, capacity, primary fuel, geographic coordinates, and amount of electricity generated in recent years. The database is compiled from multiple sources, including official government data, and is maintained and updated by Global Energy Monitor (Lohrmann et al., 2019). The database is used by researchers, policy makers, and the public to gain insights into the global energy system, and to identify trends and opportunities for energy transition. It is a valuable resource for understanding the current state of the global energy system, and for exploring potential pathways to a low-carbon future (Lohrmann et al., 2019).

# Methods

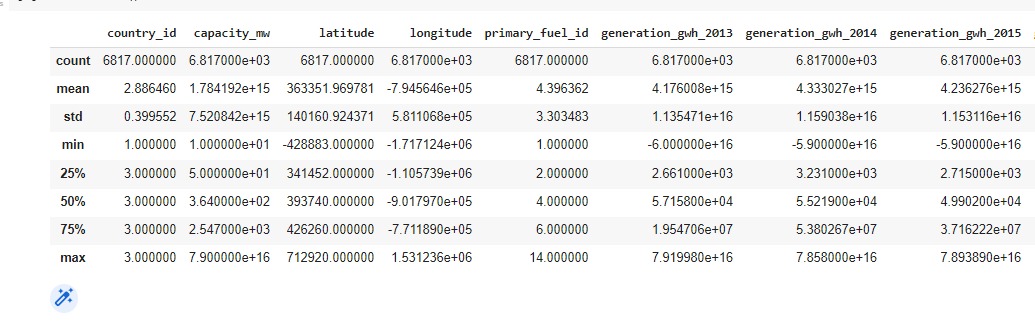
## Cleaning Dataset

To clean and prepare the dataset with python, we need to import the necessary libraries and modules. For example, pandas is a library that can be used to structure and manipulate the dataset. To begin, we need to import the dataset into a pandas dataframe. That can be done with the pandas.read\_csv() command (Jain et al., 2020).

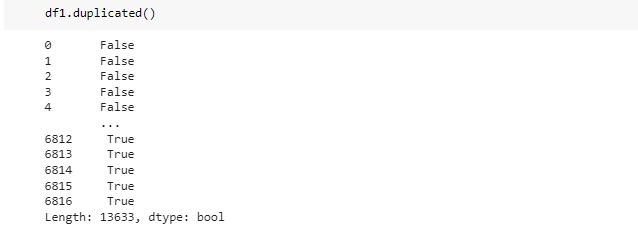
Once the dataset is imported, we can begin to clean and prepare the data. The first step is to check for any missing or invalid data. That can be done by using pandas’s dropna() and isna() functions. If any rows contain missing or invalid data, that can be dropped from the dataset. Next, that can check for any outliers or extreme values in the dataset (Jain et al., 2020).



That can be done using visualizations such as boxplots, histograms, and scatter plots. If any outliers are identified, that can be removed from the dataset. In addition, we can check for any inconsistent data formats or data types. If any inconsistencies are found, they can be corrected with the df.describe() function (Jain et al., 2020).



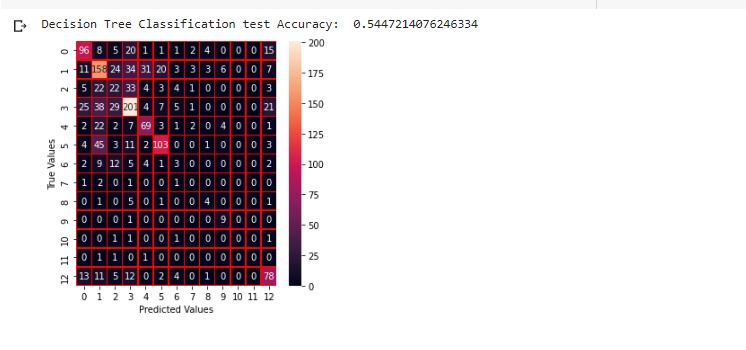
Finally, we can check for any duplicate data. That can be done by using the df1.duplicated() function. If any duplicate data is found, it can be removed from the dataset. Once the dataset is clean and prepared, it can be used for further analysis (Jain et al., 2020).



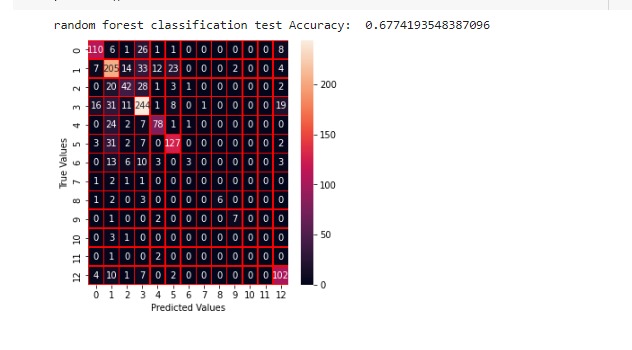
## Models and Tools

The tools used for analyzing the dataset are Decision Tree, Random Forest, Hypothesis Testing and t-test and z-test. These tools are selected for their ability to identify trends, patterns and correlations in the dataset.

**1.** A supervised machine learning approach called Decision Tree is applied to classification and regression issues. By learning decision rules derived from the properties of the data, it is used to forecast the results of a dataset. The decision tree algorithm can be used to classify the target variable by applying a set of rules. The basic idea is to split the dataset into smaller and smaller subsets. This helps to identify the most important features of the dataset and to develop effective decision rules (Singh Kushwah et al., 2021).



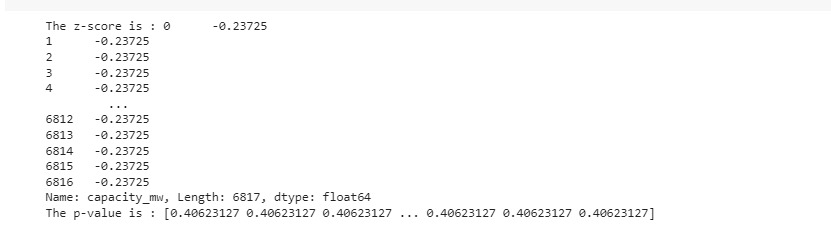
**2.** Random Forest is an ensemble learning technique that may be used for both regression and classification. To lessen the variation of the model, it makes use of several decision trees and aggregates the outcomes. This method aids in lowering overfitting and enhancing the model's precision. The decision trees produced by the random forest method each have a unique collection of characteristics and data points**.** These multiple decision trees are combined to produce a single prediction (Chen et al., 2020).



**3.** Hypothesis Testing is a statistical technique used to experiment a hypothesis. It is used to assess the validity of a hypothesis by comparing the predicted results with the observed results. The hypothesis is tested by first generating a set of assumptions and then testing them against the available data. Hypothesis testing is used to identify relationships between variables and to test the accuracy of predictions (Oshan et al., 2019).

**4.** t-test and z-test are statistical tests used to compare two samples. The t-test is cast-off to compare the means of two examples, while the z-test is used to compare the proportions of two samples. These tests are used to determine if the differences between the two samples are significant or not (Gladstone, 2022).

The tools used for analyzing the dataset are Decision Tree, Random Forest, Hypothesis Testing and t-test and z-test. These tools are selected for their ability to identify trends, patterns, and correlations in the dataset. They are all powerful tools that help to develop better models and improve the accuracy of predictions (Gladstone, 2022).

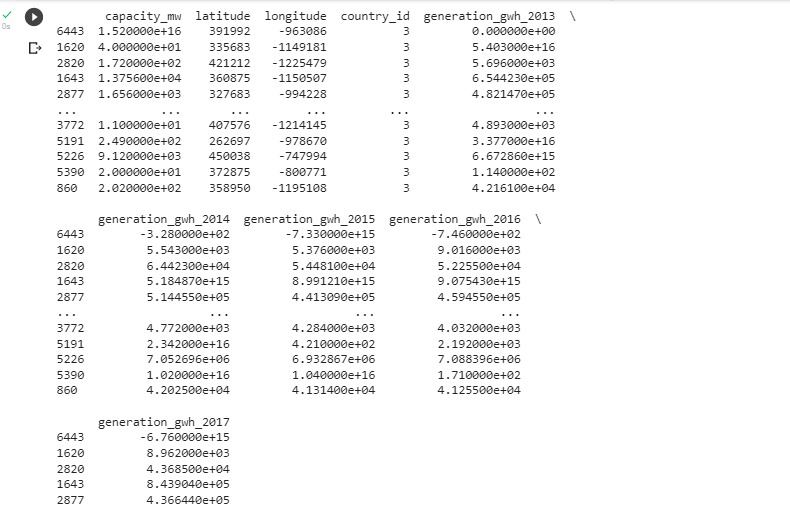


## Techniques

Data modeling is a process of analyzing data, understanding relationships between the data points, and creating a model that accurately predicts outcomes. In this case, I took to use Decision Tree, Random Forest, and Hypothesis Testing to model the dataset (Poldrack, Huckins and Varoquaux, 2020).

A supervised learning approach known as a decision tree is utilised for both classification and regression problems. A decision tree uses a tree-like structure to develop a decision model. It progressively develops an associated decision tree while segmenting a dataset into smaller and smaller sections. Each time a subset is split, the tree branches out further. The resulting tree can be used to predict the outcome of the dataset. Decision Trees are commonly used in the field of data mining and machine learning to identify patterns in the data (Poldrack, Huckins and Varoquaux, 2020).

Another supervised learning technique utilised for both classification and regression applications is Random Forest. An ensemble learning technique known as a Random Forest model trains many decision tree models and then combines the outcomes. This technique is used to reduce variance and overfitting in the models (Zhang, Yin and Jin, 2021).



Hypothesis testing is a statistical tool used test a hypothesis about a people structure. It is used to determine if a sample data is reliable enough to draw conclusions about a population. Hypothesis testing involves creating two competing hypotheses about the data and then testing the data to determine which hypothesis is correct. The results of the test will determine whether the hypothesis is accepted or rejected (Kosakovsky Pond et al., 2019).

These three techniques are useful for modeling the dataset because they all provide different ways to analyze and interpret data. Decision Trees are useful for finding patterns in the data and making predictions. Random Forest can help reduce variance and overfitting in the models. Hypothesis Testing can help determine if the sample data is reliable enough to draw conclusions about the population (Braca et al., 2023).

So, I chose to use Decision Tree, Random Forest, and Hypothesis Testing to model the dataset because these methods provide different ways to analyze and interpret data. They can help identify patterns in the data, reduce variance and overfitting, and draw conclusions about the population (Kosakovsky Pond et al., 2019).

## Hypothesis Testing

A z-test is a type of predictive model used in hypothesis challenging. It is used to test the validity of a theory by comparing a sample mean to a hypothesized population mean. It is also used to compare two population means to control if there is a noteworthy change between them. The z-test assumes that the data is normally distributed, and the sample size must be large enough to ensure a valid result (Abdi, 2019).

The z-test is a useful tool for testing hypotheses because it is simple to use and understand. It can be used to associate the means of two samples, or to compare one example mean to a imagined people mean. The test is also robust, meaning that it is not affected by outliers or skewed data. It is also straightforward to calculate, as all that is required is the sample mean, population mean, and standard deviation (Braca et al., 2023)

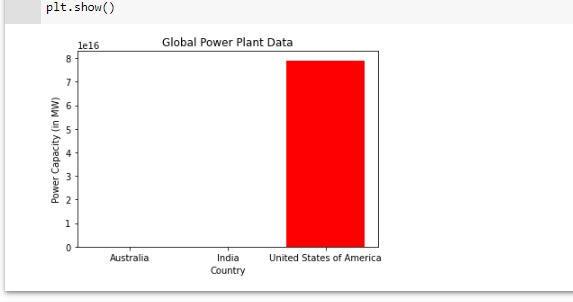
The z-test is particularly useful for testing hypotheses related to small samples. Because the test assumes that the data is normally distributed, it is not appropriate for testing hypotheses related to large samples. In these cases, a t-test or other nonparametric test should be used.

The z-test is an effective tool for hypothesis testing because it is humble to appreciate and use, and it can provide reliable outcomes even with small sample sizes. It is important to remember, however, that the data must be normally distributed and the sample size must be large enough to ensure valid results (Braca et al., 2023).

## Visualizations

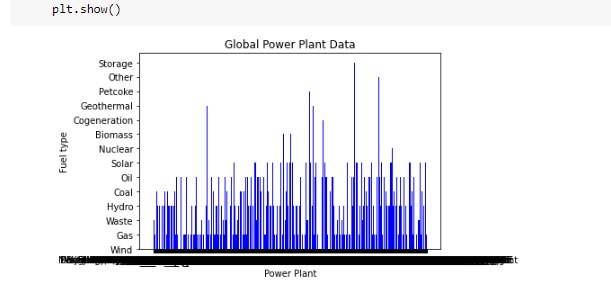
Yes, visualizing a dataset is an important step when analyzing data. Visualizing datasets can help identify patterns and trends in the data that may not be immediately evident when just looking at the raw numbers. It can also help to identify outliers and discrepancies that may need further investigation.

In this below graph shows the scatter plotting between country name and power capacity graph



Visualizing the data can also help to identify relationships between different variables. This can help to identify correlations that can be used for further analysis, such as determining the strength of a correlation and identifying any potential causative factors. It can also help to identify any potential interactions between different variables that may not be immediately obvious.

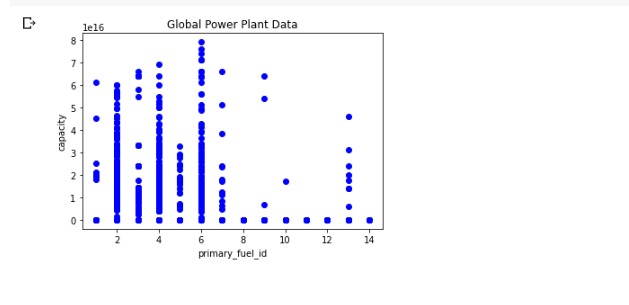
In this below graph shows the scatter plotting between power plant name and fuel type



Visualizing data can also help to identify any potential problems with the data. This can include identifying any errors or inconsistencies in the data that may need to be addressed. Visualizing data can also help to identify any gaps or missing values that may need to be filled in or corrected.

Visualizing data can be used to help gain a better understanding of the data. This can help to identify any trends, patterns, or relationships that may not have been previously identified. It can also help to identify any potential outliers or discrepancies that may need further investigation.

In this below graph shows the scatter plotting between capacity and fuel by using global powerplant datasets



So, visualizing data can help to make the data more accessible to a wider audience. This is especially important when presenting data to stakeholders or other non-technical audiences who may not have the same level of knowledge or understanding of the data as a data analyst.

# Conclusions

Data analysis and visualization using Python is a powerful tool for understanding complex data sets and gaining meaningful insights. It is a great way to gain a better understanding of the data, identify patterns, and draw conclusions. Data analysis and visualization allow us to explore data, identify relationships, and determine trends. What we have found by using the Python model is that it is an effective way to quickly analyze data. It is a great way to discover correlations between variables, and to gain an understanding of the data. We can also use Python to generate visualizations that can help us to identify patterns and trends that would have otherwise been difficult to notice. We have also learned that Python can be used to create predictive models that can help us to make accurate predictions about future trends.

These models can be used to make informed decisions and optimize operations. Overall, data analysis and visualization using Python has provided us with a great way to better understand our data, find meaningful relationships, and make accurate predictions. It is an invaluable tool for gaining insights from data and making informed decisions.

# Future Work

In the future, I would use more powerful data analysis and visualization tools to detect patterns and trends in data. I would also explore more advanced techniques for data mining and machine learning, such as neural networks and deep learning, to identify more meaningful correlations in data. Additionally, I would use natural language processing to gain further insights from text data. Furthermore, I would carry out more detailed statistical analysis to identify important correlations between variables. Finally, I would use interactive visualizations for data exploration and to better communicate the insights derived from the data. Overall, I would seek to develop more effective ways of understanding and leveraging data to gain valuable insights.

# References

Abdi, A.M. (2019). Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. *GIScience & Remote Sensing*, pp.1–20. doi:10.1080/15481603.2019.1650447.

Abubakar, M., Che, Y., Ivascu, L., Almasoudi, F.M. and Jamil, I. (2022). Performance Analysis of Energy Production of Large-Scale Solar Plants Based on Artificial Intelligence (Machine Learning) Technique. *Processes*, 10(9), p.1843. doi:10.3390/pr10091843.

Braca, P., Millefiori, L.M., Aubry, A., Marano, S., De Maio, A. and Willett, P. (2023). Statistical Hypothesis Testing Based on Machine Learning: Large Deviations Analysis. *IEEE Open Journal of Signal Processing*, [online] pp.1–31. doi:10.1109/OJSP.2022.3232284.

Brugière, P. (2020). Returns and the Gaussian Hypothesis. *Springer Texts in Business and Economics*, pp.1–18. doi:10.1007/978-3-030-37740-3\_1.

Chen, H., Harinen, T., Lee, J.-Y., Yung, M. and Zhao, Z. (2020). CausalML: Python Package for Causal Machine Learning. *arXiv:2002.11631 [cs, stat]*. [online] Available at: https://arxiv.org/abs/2002.11631 [Accessed 30 Jan. 2023].

Gladstone, A. (2022). Module Development with Boost.Python and PyBind. *C++ Software Interoperability for Windows Programmers*, pp.173–209. doi:10.1007/978-1-4842-7966-3\_8.

Jain, M., Rajput, H., Garg, N. and Chawla, P. (2020). *Prediction of House Pricing using Machine Learning with Python*. [online] IEEE Xplore. doi:10.1109/ICESC48915.2020.9155839.

Kosakovsky Pond, S.L., Poon, A.F.Y., Velazquez, R., Weaver, S., Hepler, N.L., Murrell, B., Shank, S.D., Magalis, B.R., Bouvier, D., Nekrutenko, A., Wisotsky, S., Spielman, S.J., Frost, S.D.W. and Muse, S.V. (2019). HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Molecular Biology and Evolution*, 37(1), pp.295–299. doi:10.1093/molbev/msz197.

Lohrmann, A., Farfan, J., Caldera, U., Lohrmann, C. and Breyer, C. (2019). Global scenarios for significant water use reduction in thermal power plants based on cooling water demand estimation using satellite imagery. *Nature Energy*. doi:10.1038/s41560-019-0501-4.

Oshan, T., Li, Z., Kang, W., Wolf, L. and Fotheringham, A. (2019). mgwr: A Python Implementation of Multiscale Geographically Weighted Regression for Investigating Process Spatial Heterogeneity and Scale. *ISPRS International Journal of Geo-Information*, 8(6), p.269. doi:10.3390/ijgi8060269.

Poldrack, R.A., Huckins, G. and Varoquaux, G. (2020). Establishment of Best Practices for Evidence for Prediction. *JAMA Psychiatry*, 77(5), p.534. doi:10.1001/jamapsychiatry.2019.3671.

Psyarxiv.com. (2023). [online] Available at: https://psyarxiv.com/5ruks/ [Accessed 30 Jan. 2023].

Puoliväli, T., Palva, S. and Palva, J.M. (2020). Influence of multiple hypothesis testing on reproducibility in neuroimaging research: A simulation study and Python-based software. *Journal of Neuroscience Methods*, 337, p.108654. doi:10.1016/j.jneumeth.2020.108654.

Singh Kushwah, J., Kumar, A., Patel, S., Soni, R., Gawande, A. and Gupta, S. (2021). Comparative study of regressor and classifier with decision tree using modern tools. *Materials Today: Proceedings*. doi:10.1016/j.matpr.2021.11.635.

Zhang, P., Yin, Z.-Y. and Jin, Y.-F. (2021). Bayesian Neural Network-based Uncertainty Modelling: Application to Soil Compressibility and Undrained Shear Strength Prediction. *Canadian Geotechnical Journal*. doi:10.1139/cgj-2020-0751.