**PROJECT REPORT**

# **Abstract**

The project is a machine learning analysis on globalpowerplant energy production worldwide and similarly an analysis on United States of American nuclear powerplants net generation.

The project contains a Jupyter Notebook, notebook file, two csv datasets, and a worddocument report.

The project contains two coding algorithms using the Python programming language to analyse both of the datasets, the firstdataset (“global\_power\_plant\_database.csv”) uses a Linear Regression code to showcase future prediction, performance of the model, and a visual prediction scatter graph.

Furthermore, the second dataset

(“nuclear\_energy\_overview\_eia.csv”) uses a K Nearest

Neighbour code algorithm to showcasing energy predictions for American nuclear plants.

The core findings of the analysis showcase’s performance metrics of the models used and predictions of both global powerplants and American nuclear plants.

Domain website for (“global\_power\_plant\_database.csv”) ([Global Power Plant Database | Resource Watch](https://resourcewatch.org/data/explore/Powerwatch?section=Discover&selectedCollection=&zoom=3&lat=0&lng=0&pitch=0&bearing=0&basemap=dark&labels=light&layers=%255B%257B%2522dataset%2522%253A%2522a86d906d-9862-4783-9e30-cdb68cd808b8%2522%252C%2522opacity%2522%253A1%252C%2522layer%2522%253A%25222a694289-fec9-4bfe-a6d2-56c3864ec349%2522%257D%255D&aoi=&page=1&sort=most-viewed&sortDirection=-1)).

Domain website for (“nuclear\_energy\_overview\_eia.csv”) ([Nuclear Energy Datasets](https://www.kaggle.com/datasets/alistairking/nuclear-energy-datasets)).

# **Introduction**

## A. Analysis Objectives

Create a Linear Regression for the global powerplants dataset.

Create a K Nearest Neighbour analysis on the nuclear powerplants net generation within USA, dataset.

Outline the performance of Linear Regression and K Nearest Neighbour.

Visualize the predictions.

## B. Motivation

Creating an adequate analysis on net generation production worldwide and American nuclear net generation.

Predicting future energy outputs which can help energy sectors have a recorded analysis on what kind of energy output can be possible if the current trend is continued.

Help in policy making and future decisions regarding this subject.

Predicting trends.

## C. Overview

The next sections are Methodology How the datasets, theme, code, and algorithms were selected and made, Results of Linear Regression, K Nearest Neighbour and the general reason to these results and its implication within the analysis, Discussion of these results, Conclusions regarding everything the project analysis has discovered, Future work that could have been done if the project could have been bigger and other various python code that could have been used, and lastly References of every website from where the project got its datasets, code.

# **Methodology**

## Finding the Datasets and Theme for Selection

The first task that was needed to be done before even starting the analysis was to brainstorm which theme of subject to use. Beginning with scouting for csv datasets. The first website that was found for suitable use, as it contained a large amounts of csv datasets was Kaggle.com. Spending time looking through various themes, it was concluded to make an analysis on energy production or net generation output from various power plants such as electric outputs and nuclear outputs. This theme of subject would be appropriate for machine learning predictions as it would help in outlining energy trends for future uses that can help make conclusions as to what will future energy output can be, making sure that net generation output can be planned for and used suitably.

Likewise, a second dataset was needed for the other task at hand, that required its own dataset of use, this led to more web searching until resourcewatch.org was discovered that contained a great dataset on global energy production within each country.

## Technical

Tools Used

* Anaconda: A downloadable library tool application where Jupyter Notebook is available.

A black text on a white background

Description automatically generated

* Jupyter Notebook: A Python code documentation, visualization, and coding used to create informative machine learning analysis.



* Python: A common programming language.

A blue and yellow snake logo

Description automatically generated

Linear Regression

* Importing the needed libraries.

A screenshot of a computer program

Description automatically generated

* Reading the file with pandas library tool pd.read\_csv(file).

A white background with black and red text

Description automatically generated

* Assigning values to X and y and using them for testing.

A computer screen shot of a computer code

Description automatically generated

* A Predicting the targets and showcasing performance metrics using Mean Squared Error, R squared and Mean Absolute Error.

A screenshot of a computer program

Description automatically generated

* Plotting a visual graph: Using library tool such as matplotlib which helps in creating visually appealing graphs such as histograms, scatter boxes and bar graphs for better understanding and evaluation.

A white background with black text

Description automatically generated

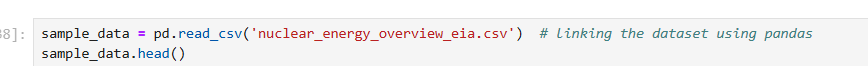
K Nearest Neighbour

* Importing the needed libraries.

A screenshot of a computer program

Description automatically generated

* Reading the file with pandas pd.



* Converting the months into a numerical value with month mapping where you manually add in each month and assign a numerical value to it so it can be used for testing and training.

A calendar with numbers and letters

Description automatically generated

* Assigning values to X and y and using them for testing.

A white background with black text

Description automatically generated

* Training the Knn model on the training data x and y train.

A screenshot of a computer program

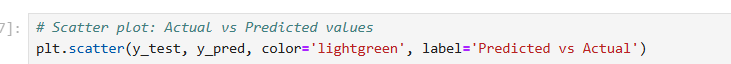
Description automatically generated

* Performance metric using Mean Squared and R squared.

A black and blue text

Description automatically generated

* Plotting a visual graph with matplotlib.

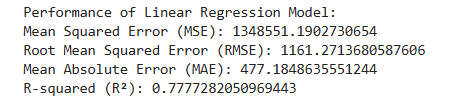


# **Results and Discussions**

## Linear Regression Results

1. Model Performance of Linear Regression

Image:



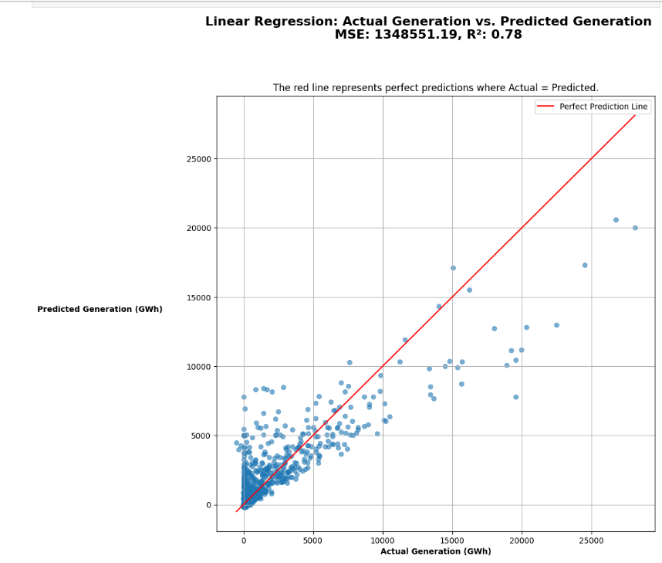
1. Mean Squared Error (MSE): 1348551.1902730654:

Represents the average squared differences between the predicted and actual values.

1. Root Mean Squared Error (RMSE): 1161.2713680587606
2. Mean Absolute Error (MAE): 477.1848635551244
3. R-squared (R²): 0.7777282050969443:

R² measures the proportion of the variance in the dependent variable (actual values) that is explained by the independent variables (features) in the model.

1. Scatter box plot



Showcases a scatter box graph that compares true observed outcomes or actual outcomes to the predicted outcomes represented here as blue dots. The points that lie near towards the red line are the accurate predictions (predicted value that is close to the true value).

The graph showcases the predictions of energy output named as GWh, the lefthand side going up from zero to twenty five thousand is the predicted generation while the bottom across is the actual generation from the global worldwide powerplant generation.

The Red line represents the perfect outcome where y actual is equal to y predicted. Named as the perfect prediction line.

The title includes the Mean Squared Error measurement as the measured value is quite high it might be possible that the chart is inaccurate but since the project is working with high GWh net output results, it is acceptable.

## K Nearest Neighbour Results

1. Model Performance

Image:



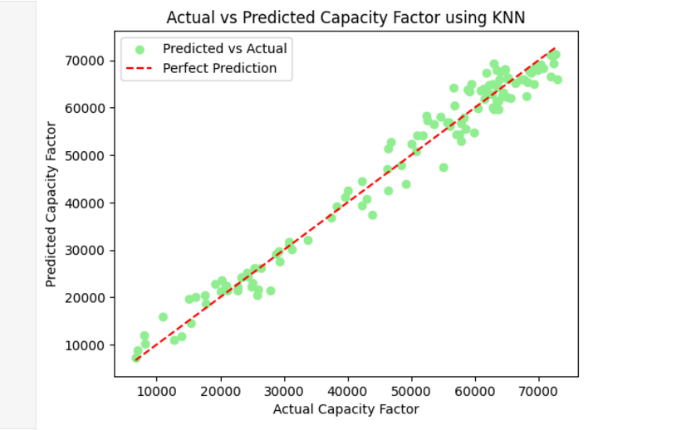
1. Mean Squared Error: 9797408.033170732:

Represents the average squared differences between the predicted and actual values.

1. R-squared: 0.9743002439775875:

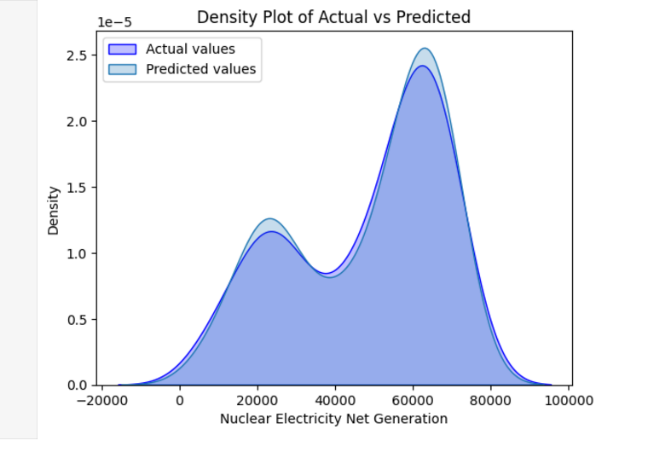
R² measures the proportion of the variance in the dependent variable (actual values) that is explained by the independent variables (features) in the model

1. Scatter box plot



The box plot again as last time represents the actual vs the predicted (y test) vs (y pred). The dots that are closer to the red line are represented as accurate predictions. If dots are further away from the line that indicates prediction errors.

1. Density plot



The density plot or Kernel Density Estimation is used to visualize the actual and predicted values for nuclear energy electricity net generation.

Since the actual and predicted overlap quite commonly this showcases that the model predictions is in a well off alignment with true values, highlighting that the model is doing as intended with no bias.

Both of the values display high peaks which means a concentration of data around two point five density around the six thousand net generation. Likewise, might indicate that it cant capture the full variability of the data.

# **Conclusions and Future Work**

In conclusion, this project aimed to predict global power plant net generation and US nuclear power plant net generation using two machine learning methods, Linear Regression and K Nearest Neighbour. After assigning the training and testing values the key prediction and results were as followed.

1. Model Performance

Linear regression highlighted R squared prediction as (0.7777282050969443

) , Mean Squared Error (1348551.1902730654

), Mean Absolute Error (477.1848635551244

), and Root Mean Error (1161.2713680587606

).

K Nearest Neighbour marked R Squared prediction as (0.9743002439775875

) and Mean Squared Error (9797408.033170732

).

1. Visuale Insights

From the Linear Regression scatter plot most of the predictions were close to the real values, while many of them got clustered to the left, showcasing deviations from the red prediction line.

For the KNN model overall showed much better results than Regression with the prediction line and dots being closer aligned, furthermore the density plot visualized some biases in the prediction with prediction curve diverging slightly in certain areas from the actual.

1. Future Work

For future work if given the time to do it on the project analysis more datasets would be utilized to find similarities or issues that one dataset cannot give within various themes that have a similar link to each other such as nuclear energy with uranium production, assigning more models for evaluation, using simpler machine learning techniques to find new information regarding the dataset being used. Focusing more on nuclear energy as that was the intended goal from the start.

Furthermore, the more detailed evaluation regarding nuclear energy and global energy can be found on my GitHub here ([AlexMGarbalyauskas/Jupyter\_Notebook\_American\_Nuclear\_PowerPlants\_Analysis](https://github.com/AlexMGarbalyauskas/Jupyter_Notebook_American_Nuclear_PowerPlants_Analysis)). Which goes into further analysis regarding nuclear energy.

##### **References**

1. Global Power Plant Database - Resource Watch  
   "Global Power Plant Database", Resource Watch, Available: <https://resourcewatch.org/data/explore/Powerwatch?section=Discover&selectedCollection=&zoom=3&lat=0&lng=0&pitch=0&bearing=0&basemap=dark&labels=light&layers=%255B%257B%2522dataset%2522%253A%2522a86d906d-9862-4783-9e30-cdb68cd808b8%2522%252C%2522opacity%2522%253A1%252C%2522layer%2522%253A%25222a694289-fec9-4bfe-a6d2-56c3864ec349%2522%257D%255D&aoi=&page=1&sort=most-viewed&sortDirection=-1> [Accessed: Nov. 2024].
2. Nuclear Energy Datasets - Kaggle  
   A. King, "Nuclear Energy Datasets", Kaggle, Available: <https://www.kaggle.com/datasets/alistairking/nuclear-energy-datasets> [Accessed: Nov. 2024].
3. Simplilearn, "K Nearest Neighbors (KNN) Algorithm | KNN Classifier | KNN in Machine Learning | Machine Learning Algorithms," *YouTube*, Oct. 15, 2021. [Online]. Available: <https://www.youtube.com/watch?v=OO7Y5wQWnQs>. [Accessed: Nov. 30, 2024].