Predicting Power Output of Solar Panels in Cairo, Egypt

Islam Anwar[1] **and Karim ElBowety**[1]

[1]Faculty of Science & Innovation, University of Prince Edward Island, Charlottetown, Canada

This paper analyses recent historical weather data in Cairo, Egypt and uses it to predict the power output of solar panels. Highlighted key aspects of the data science process are included in this paper such as exploratory data analysis & data pre-processing. Furthermore, various machine learning models were tested and compared using a variety of evaluation metrics to discover which predicted the power output of solar panels most accurately. According to gathered results and figures, Multilayer Perceptron and Lasso Regression produce the most accurate predictions. However, the lack of current data available in our chosen region was not ideal for conducting our research.

Power Output Solar Panels | Regression | Exploratory Data Analysis

Introduction

Solar energy has entered the forefront in terms of renewable energy sources in recent years. With a particular focus on Egypt, it is becoming more essential to implement this energy source considering the nature of the Egyptian weather. Few pieces of research have been conducted towards analyzing the power outputs of solar panels in the region of Cairo, Egypt. This paper aims to fill that gap.

In this paper, we go through a full data science process, beginning with understanding the problem at hand and framing it, followed by data acquisition. Then we performed various data preparation techniques not only to understand the nature of our data and the correlation between features and our target label but to also conduct exploratory data analysis using statistical methodologies.

Data Acquisition

To get a dataset suitable for predicting the power output of solar panels in Cairo, Egypt, we needed a source that collected information regarding solar radiation, sun height, angle, and other relevant data in our chosen location. The data also had to be recent as, due to climate change, data that is a couple of years old may not be relevant anymore. We ended up using PVGIS, which is a tool provided by the EU Science Hub, for multiple reasons. These included the fact that it was free to use and that it had data for Egypt. Furthermore, the data provided by the tool was numerous and looked to be suitable for use in prediction.

We also tried obtaining datasets for air pollution indicators and dust pollution statistics but were unable to find any that would complement our main dataset.

Identified Datasets. The PVGIS tool provides several datasets including one for the performance of grid-connected PV systems, one for the performance of off-grid PV systems, one for tracking PV systems, one for monthly data, one for daily data, and one for hourly data.

Each of those offers more settings to further customize the data to your needs, the first of which of course is the start and end years, which can range from 2005 to 2020. Other than that, it is possible to set the mounting type of the PV system (whether it's tracking or not and on which axes), its slope and azimuth, the type of technology the PV system utilizes, and the installed peak PV power, and the estimated system losses.

Final Acquired Datasets. In the end, we decided on choosing an hourly dataset as that provided a lot more data in comparison with the next tier (daily). For the time, we went with 2019-2020 as we felt that anything before 2019 could produce unsatisfactory or inaccurate results due to climate change changing the Earth's weather significantly every couple of years.

Since there were multiple types of mounts for the PV system, we decided that choosing the most basic one (fixed) along with the optimal slope and azimuth for it would be the best course of action as that would be easiest to generalize for other systems. The optimal slope turned out to be 28°while the optimal azimuth was 0°.

As for the PV technology, we chose the default setting which was Crystalline Silicon, with an installed peak PV power of 1 kWp and an estimated system loss of 1%. The data was exported in JSON format. There was an option for exporting as CSV but all samples produced in that format were not properly separated so we went with the JSON format.

The features collected were:

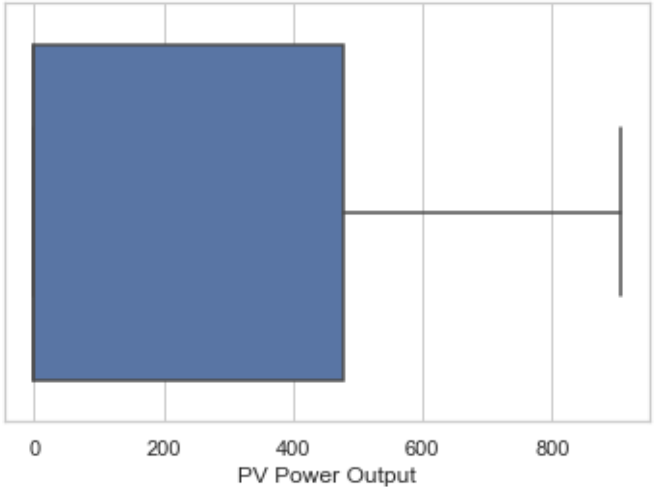
- PV power output in W (Target Variable)
- Direct in-plane irradiance in W/m^2
- Diffuse in-plane irradiance in W/m^2
- Reflected in-plane irradiance in W/m^2
- Sun height in degrees
- Air temperature in Celsius
- Wind speed at 10m in m/s

Data Preparation

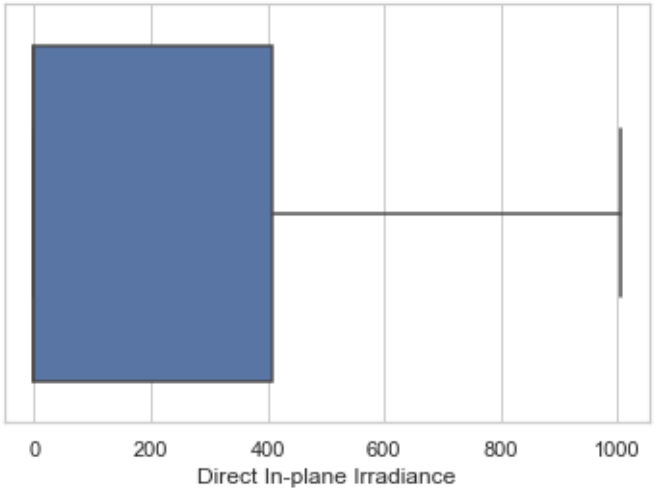
Our first step was converting the JSON file into a Pandas DataFrame for ease of use. This was easily done using the Pandas function json_normalize. After doing that, we noticed that there was a column called Int that consisted entirely of values of 0. Thus, we dropped the column since it was useless. Finally, we changed the names of the columns to more comprehensible ones based on PVGIS's explanation of each feature.

Exploratory Data Analysis.

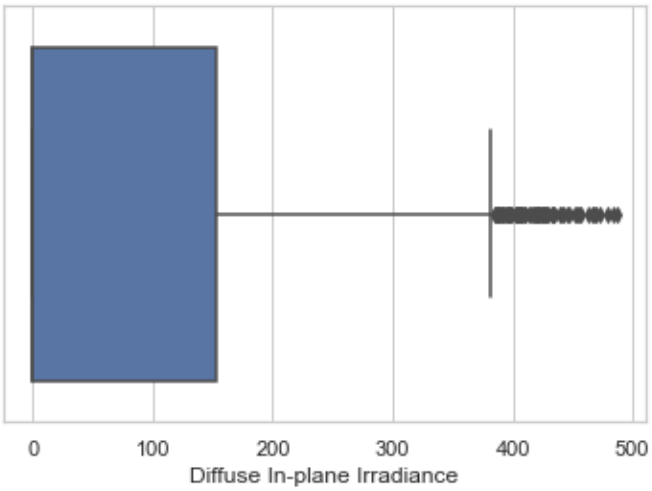
Uni-variate Analysis. For uni-variate analysis, we created a boxplot for each feature including our target variable to better understand their distribution.



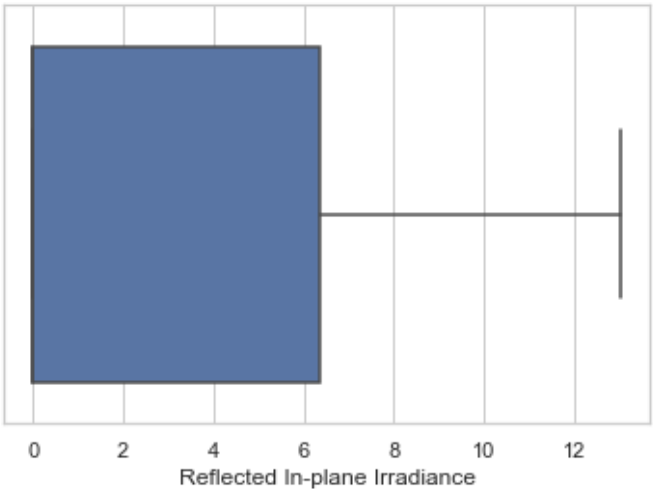
For our target variable, the PV Power Output, we can see that it is highly positively skewed. This will be the trend for most of our features as will be apparent.



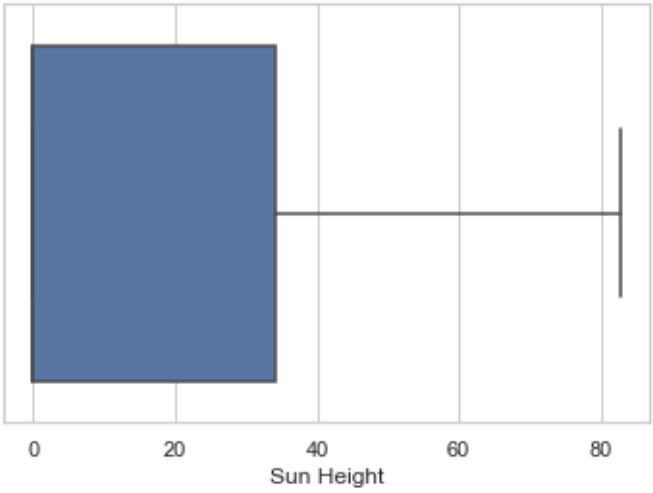
For the feature Direct In-plane Irradiance, it is also highly positively skewed.



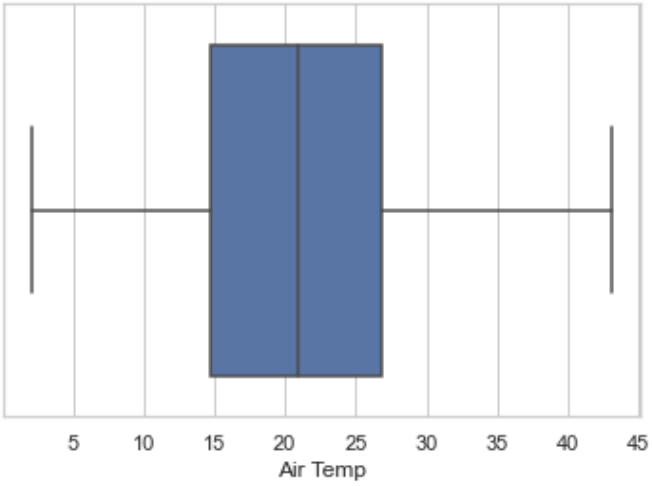
For the feature Diffuse In-plane Irradiance, it is also highly positively skewed but with a significant number of possible outliers.



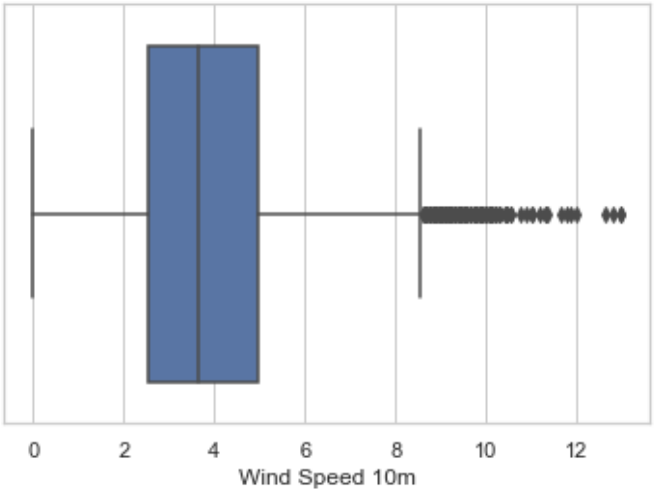
For the feature Reflected In-plane Irradiance, it is also highly positively skewed.



For the feature Sun Height, it is also highly positively skewed.



For the feature Air Temperature, it appears to be almost normally distributed with no skewness.



For the feature Wind Speed at 10m, it is slightly positively skewed but with a significant number of possible outliers.

Bi-variate Analysis. For bi-variate analysis, we performed multiple processes to check how each feature interacts with others.

	PV Power Output	Direct In-plane Irradiance	Diffuse In-plane Irradiance	Reflected In-plane Irradiance	Sun Height	Air Temp	Wind Speed 10m
PV Power Output	1.000000	0.983943	0.870672	0.969558	0.936368	0.340251	0.176114
Direct In-plane Irradiance	0.983943	1.000000	0.775787	0.964400	0.919178	0.367682	0.135549
Diffuse In-plane Irradiance	0.870672	0.775787	1.000000	0.849943	0.869784	0.304729	0.231795
Reflected In-plane Irradiance	0.969558	0.964400	0.849943	1.000000	0.982640	0.442536	0.160086
Sun Height	0.936368	0.919178	0.869784	0.982640	1.000000	0.447678	0.178098
Air Temp	0.340251	0.367682	0.304729	0.442536	0.447678	1.000000	0.234634
Wind Speed 10m	0.176114	0.135549	0.231795	0.160086	0.178098	0.234634	1.000000

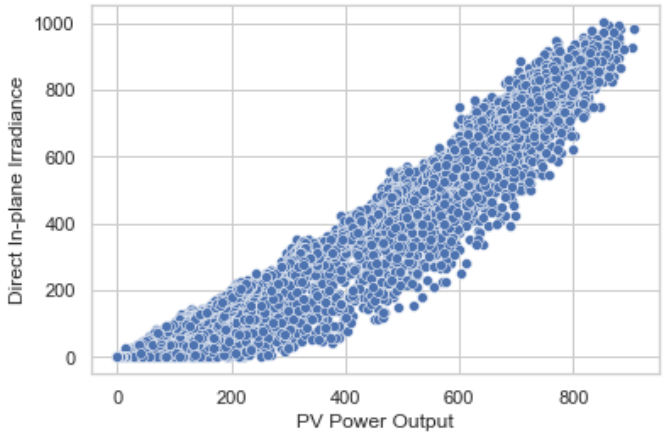
We checked the correlation of each feature with all the others. From this, we can learn multiple things. The first of which is that the Direct In-plane Irradiance, Diffuse In-plane Irradiance, Reflected In-plane Irradiance, and Sun Height are the features most positively correlated with our target variable. On the other hand, Wind Speed is negatively correlated with it which implies that the stronger the wind is, the less effective the PV system is. The remaining feature, Air Temperature, appears to be weakly negatively correlated with the target variable.

We can also note that Direct In-plane Irradiance, Diffuse In-plane Irradiance, Reflected In-plane Irradiance, and Sun Height are all positively correlated with each other. This makes sense as they are all related to the sun and its position in the sky with respect to the PV system.

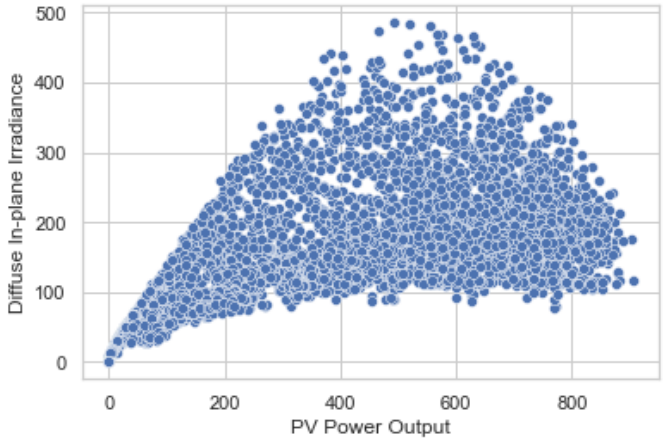
	PV Power Output	Direct In-plane Irradiance	Diffuse In-plane Irradiance	Reflected In-plane Irradiance	Sun Height	Air Temp	Wind Speed 10m
PV Power Output	83319.947117	83203.234982	23606.211011	1143.376132	6247.120456	779.828032	88.697739
Direct In-plane Irradiance	83203.234982	85820.598520	21346.937076	1154.233645	6223.777137	855.249483	69.284732
Diffuse In-plane Irradiance	23606.211011	21346.937076	8822.550565	326.157928	1888.282551	227.266417	37.987855
Reflected In-plane Irradiance	1143.376132	1154.233645	326.157928	16.690982	92.788510	14.355367	1.141137
Sun Height	6247.120456	6223.777137	1888.282551	92.788510	534.216697	82.157869	7.182282
Air Temp	779.828032	855.249483	227.266417	14.355367	82.157869	63.044759	3.250566
Wind Speed 10m	88.697739	69.284732	37.987855	1.141137	7.182282	3.250566	3.044308

We then checked the covariance matrix of the features. The resultant matrix supports our findings from the correlation matrix.

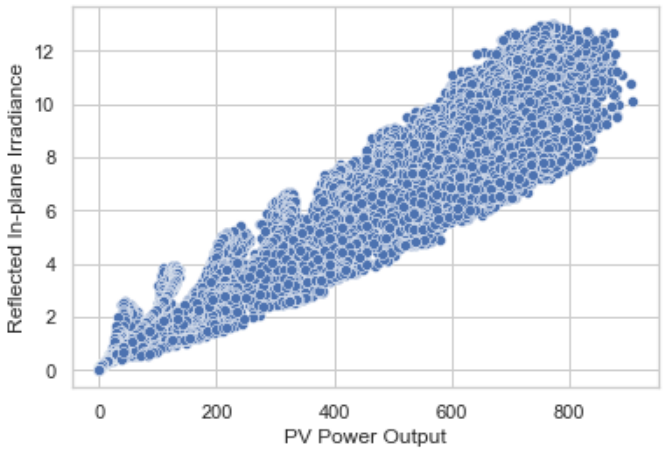
After that, we decided to draw a scatterplot of each feature against the target variable, to better understand the relation between them.



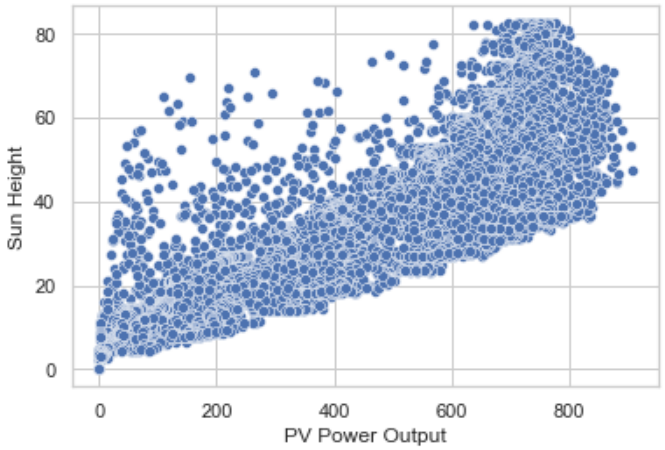
The relationship between the PV Power Output and the Direct In-plane Irradiance appeared to be the most positively correlated. This makes sense as it stands to reason that more sun rays falling directly on the solar panel would generate more power.



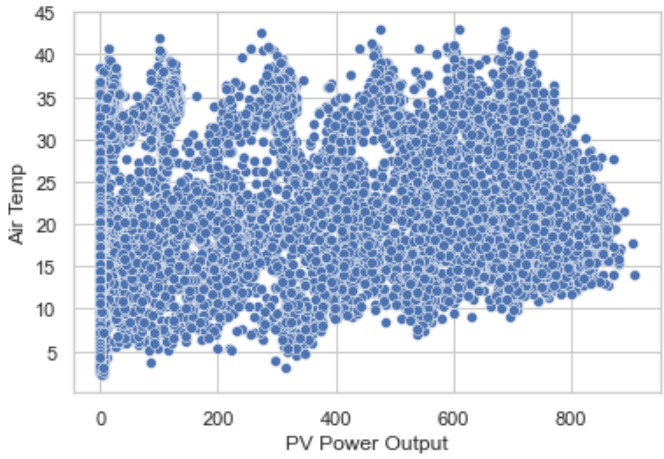
The relationship between the PV Power Output and the Diffuse In-plane Irradiance appeared to be slightly positively correlated at first. However, it then fans out, indicating that past a certain value for the irradiance (which appears to be near 100 W/m^2), the feature doesn't make much of a difference on the power output. This is logical is diffuse irradiance is simply the light received after sun rays diffuse in the atmosphere, which forms a low percentage of the received irradiance.



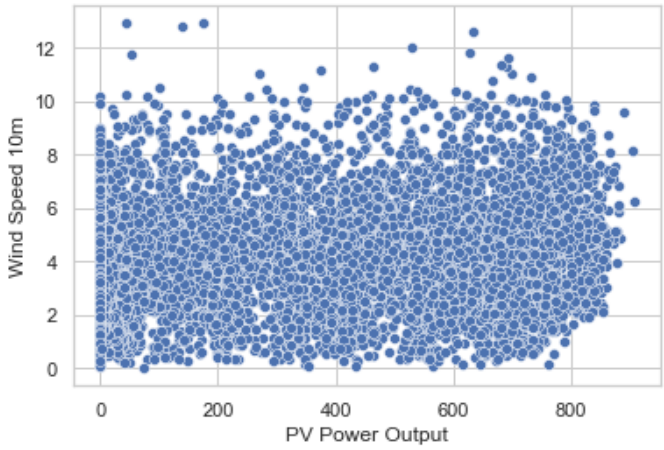
Similar to the Direct In-plane Irradiance, the relationship between the PV Power Output and the Reflected In-plane Irradiance appeared to be strongly positively correlated. This makes sense as reflected irradiance is the light received after sun rays bounce of the ground and reaches the PV system, which forms a not insignificant percentage of the received irradiance.



The relationship between the PV Power Output and the Sun Height appears to be weakly positively correlated. This can be explained by the fact that even though the higher the sun's height, the more energy is received by the PV System, this feature doesn't include the irradiance received, which is ultimately the more important factor in calculating how much power is created.



Their appears to be no clear relationship between the PV Power Output and the Air Temperature.



Similarly, their appears to be no clear relationship between the PV Power Output and the Wind Speed.

Data Cleaning. We first began cleaning the dataset by checking if there are any null values present so we can deal with them. It turned out that there were no null values so nothing had to be done here.

After that, we started checking for uni-variate outliers using the z-score method. This showed that there aren't any outliers as no values were above 3 or under -3. 3/-3 here represents that a value is further from the mean from 99.7% of the values.

We then checked for multi-variate outliers, by calculating the Mahalanobis Distance. We then applied the Chi-Square test to the calculated distances to see if any are statistically significant.

	PV Power Output	Direct In-plane Irradiance	Diffuse In-plane Irradiance	Reflected In-plane Irradiance	Sun Height	Air Temp	Wind Speed 10m	Mahalanobis	P value
8	671.03	637.45	173.17	6.42	30.96	13.39	3.66	12.362390	2.067955e-03
9	764.29	775.75	162.03	7.55	35.72	14.64	3.86	15.042657	5.414129e-04
10	723.01	649.96	232.27	7.34	36.88	15.60	4.55	10.707274	4.730914e-03
11	612.03	471.57	267.46	6.33	34.20	16.06	4.90	10.822679	4.465654e-03
12	635.67	622.63	145.00	5.90	28.20	16.20	4.97	12.712586	1.735789e-03
31	508.97	424.44	169.46	4.55	23.38	11.10	5.93	10.738885	4.656728e-03
32	690.83	647.69	167.10	6.44	30.99	12.49	6.41	18.796167	8.288275e-05
33	778.98	759.62	170.32	7.52	35.78	14.07	6.97	21.602319	2.037586e-05
34	789.54	754.76	185.46	7.67	36.97	15.21	8.69	24.815074	4.087663e-06
35	204.69	16.87	235.46	3.10	34.32	15.64	9.59	27.721029	9.559935e-07

It turned out that there are values that are statically outliers. However, due to the fluctuating nature of weather, we decided to treat these as meaningful outliers and leave them in the dataset as they weren't large enough to consider as actual outliers.

We finally decided to apply feature selection using a variance threshold to remove any features that have no variance at all and it turned out that there are no such features in our dataset.

Data Transforming. We decided to scale the data as some of our chosen models work using gradient-based methods which requires that all the features in the dataset be similarly scaled or else the model may not converge. Thus, we decide to scale the data using the Robust Scaler as it is based on percentiles and is not affected by outliers, of which our dataset contains an amount.

As a final step, we split the dataset into 2 for training and testing, with the testing dataset 20% of the original dataset.

Data Analysis

Post-data preparation, for our data analysis, we selected a collection of models we thought were suitable for our problem to evaluate the results and perform comparative measures. This was conducted firstly by defining our models and their respective hyperparameters, then fitting as well as predicting.

Model Selection. Regarding our selected models, the table below provides information regarding the models we chose including Linear & Lasso Regression, MLP (Multilayer Perceptron), Random Forest, etc. This variety of models was chosen to display the variation of results by applying models that use different approaches such as instance-based regressor with KNR, backpropagation with MLP, multiple decision trees with Random Forest, and many more.

Models
Linear Regression
Support Vector Regressor
Multilayer Perceptron
K-Neighbors Regressor
Random Forest Regressor
Gradient Boost Regressor
Lasso Regression

Hyperparameter Optimization. With the aspect of hyperparameter optimization, the following table illustrates the chosen values of our hyperparameters for each model.

Model	Hyperparameter	Value
Linear Regression	Fit Intercept	False
Support Vector Regressor	C	200
	Epsilon	5
MLP	Epsilon	$1e^{-8}$
	Max Iterations	1000
K Neighbors Regressor	N Neighbors	6
Random Forest Regressor	N Estimators	20
	Max Depth	10
Gradient Boosting Regressor	Random State	0
Lasso Regression	Fit Intercept	False
	Alpha	0.01

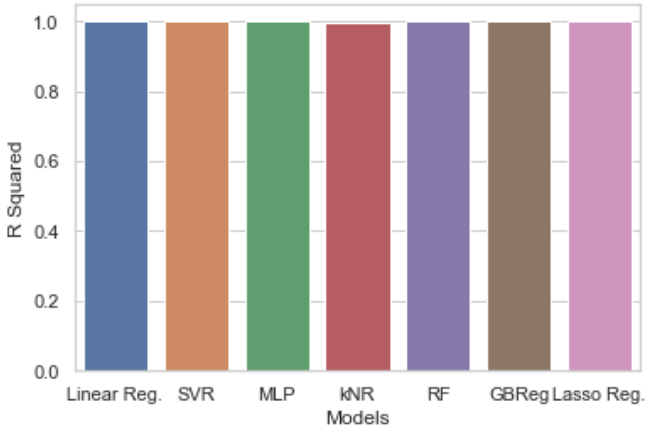
These hyperparameters were chosen based on several trials, and indicated the most consistent and accurate results.

Results

Evaluation Metrics. For our evaluation, we used 3 metrics: R Squared, Root Mean Squared Error and Mean Absolute Error.
The R Squared metric, also known as the Coefficient of Determination, is the amount of variance in the target variable that can be predicted from the features.
The RMSE is the standard deviation of the residuals, which can show if a model is usable for real-life prediction or not.
The MAE is the average of the absolute residuals. Similarly, it can show if a model is usable for real-life prediction or not.

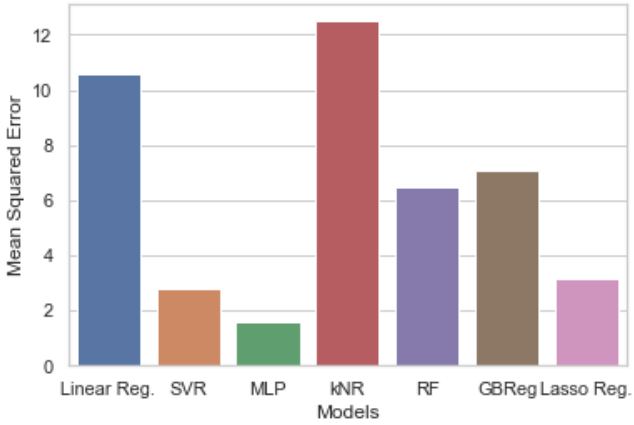
Results. According to the evaluation metrics we opted for, the following figures display the comparison of results between our models on the dataset.

Coefficient of Determination. Results for the coefficient of determination produce very little insight between the differences in our models and which works best as they all performed near perfectly to the dataset as shown in the figure below.

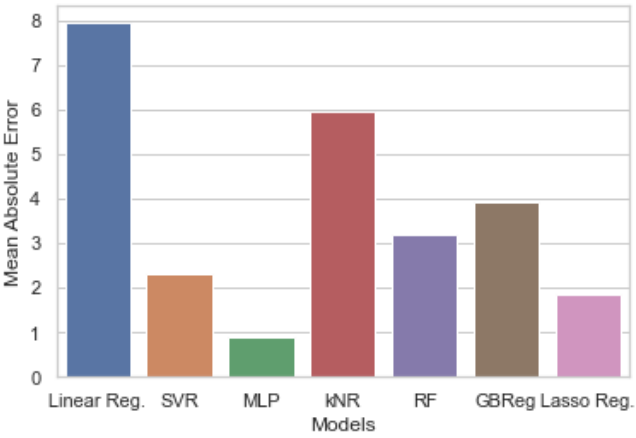


Standard Deviation of Residuals. Concerning the standard deviation of the residuals, we were able to extract very meaningful information from the figure below. It shows that the multilayer perceptron model is the most effective in terms of

minimizing the mean squared error - with lasso regression and support vector regressor closely behind. However, models such as linear regression, random forest, and others didn't work particularly well at minimizing the mean squared error.



Mean Absolute Error. As for the mean absolute error, we also deduct valuable information from the figure showing that the multilayer perceptron model performed very well for this evaluation metric as well, with lasso regression and SVR slightly lagging. This illustrates that MLP is a very strong model corresponding to this dataset and following the various pre-processing techniques performed.



Conclusions

In conclusion, it is possible to predict the output of a PV system with a high degree of accuracy if you have enough recent solar information about the region meant for prediction. Our results show that 2 years' worth of data should be enough to get usable results.

Many models are suitable for this challenge. The best 2 we tested were Multilayer Perceptron (Neural Network) and Lasso Regression (Linear Regression with L1 Regularization). These 2 models produced comparable results and were the lowest amongst the 7 that we have tested, while also not being computationally expensive.

Still, working on a region without available data such as air pollution indicators statistics is limiting and to reach better results, this data needs to be gathered somehow.