University for the Creative Arts

BERLIN SCHOOL OF BUSINESS & INNOVATION

Name of Assignment : Time Series Forecasting: A Practical Approach to Data-Driven.

Energy System Analysis

Programme Title : Fundamentals of Data Analytics

Name: JIMSON JAMES T

Year:2025-2027

Table of Contents

Statement of compliance with academic ethics and the avoidance of plagiarism

I honestly declare that this dissertation is entirely my own work and none of its part has been copied from printed or electronic sources, translated from foreign sources and reproduced from essays of other researchers or students. Wherever I have been based on ideas or other people texts I clearly declare it through the good use of references following academic ethics.

(In the case that is proved that part of the essay does not constitute an original work, but a copy of an already published essay or from another source, the student will be expelled permanently from the program).

Name and Surname :

JIMSON JAMES T

....

Date: ....27................./...01......./.....2026....

1 INTRODUCTION

The optimization of the cost and storage of the renewable energy electricity, the efficient grid planning and adequate forecast of the energy production is required. In special instance, reliance on weather of production of solar energy is very high, time constraints, such: the sun strength, the occurrence of clouds, the temperature and the dayday duration. The regression over time-series, where the 2-year model of the solar energy data (2017-2022) is analysed with time sample of 15 minutes, is the tool used in the current assignment to determine the patterns of the energy capture over the years, but also to predict the energy capture in the future.

The four goals of the paper are (1) exploration and summary of the data with use of descriptive statistics and visual analytics; (2) data quality with use of preprocessing and feature engineering; (3) predictive models with use of the Linear Regression and random forest Regression; and (4) business-relevant prediction of average production of solar power in January 2026 with use of managerial knowledge and measure of confidence.

2 Justification and Description of Dataset

This data is available publicly in time series of solar energy that includes more than 10,000 observation of solar energy in 2017-2022 at 15 min time interval. Energy Capture (continuous) and Rainfall would be used as target variable and Cloud cover Ratio, Temperature, Sunlight Intensity and LengthofDaylight and a timestamp would be used as explanatory variables.

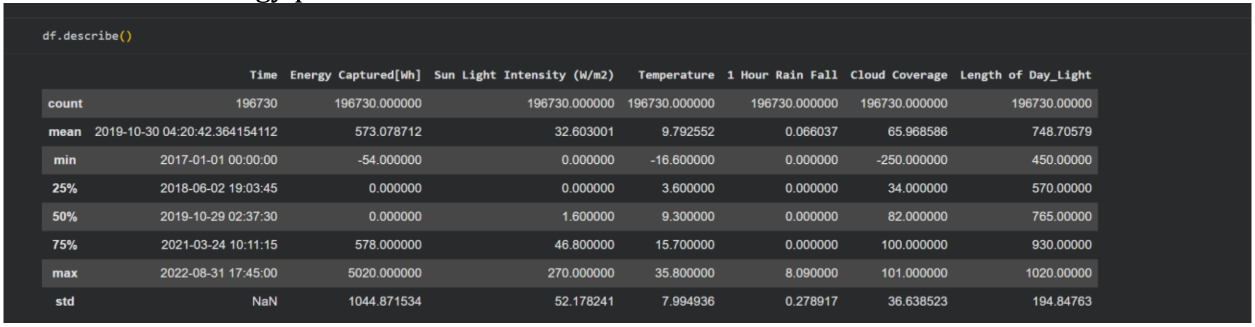
This has been the dataset of choice due to its high applicability to the reality in the forecasting of renewable energy, the size of data is large enough to model a statistical learning model, and the features contained in the data are diverse environmental forces that can be used to analyze the characteristics of significance. The temporal granularity may be used to examine the short run fluctuation of the price and the long run seasonal fluctuations.

3 Task 1- Exploratory Data Analysis (EDA)

3.1 Descriptive Statistics

The descriptive statistics such as the mean, the median and the standard deviation and mode were used in calculating all the numerical variables. The average energy that it gathered in use given thoughtful awareness of creation of solar energy would have contained the average degree of creation during the total time in period of observation than the median which is the central propensity that is not very sensitive to severe weather conditions. The enormous deviation of the mean and the median indicates that there is skewness and it might have been so due to 0 production at night or unbelievable cloudy conditions.

The inconsistency of the character that determines the daily and season cycles of the solar generation is the argument of energy capture points. Meteorological variables are also influenced by the intensity and the extent of cover of the sun, and that these variables have relatively very high degree of variation, as well as naturally will be introduced into the fluctuation of energy production.



```
df.describe()
```

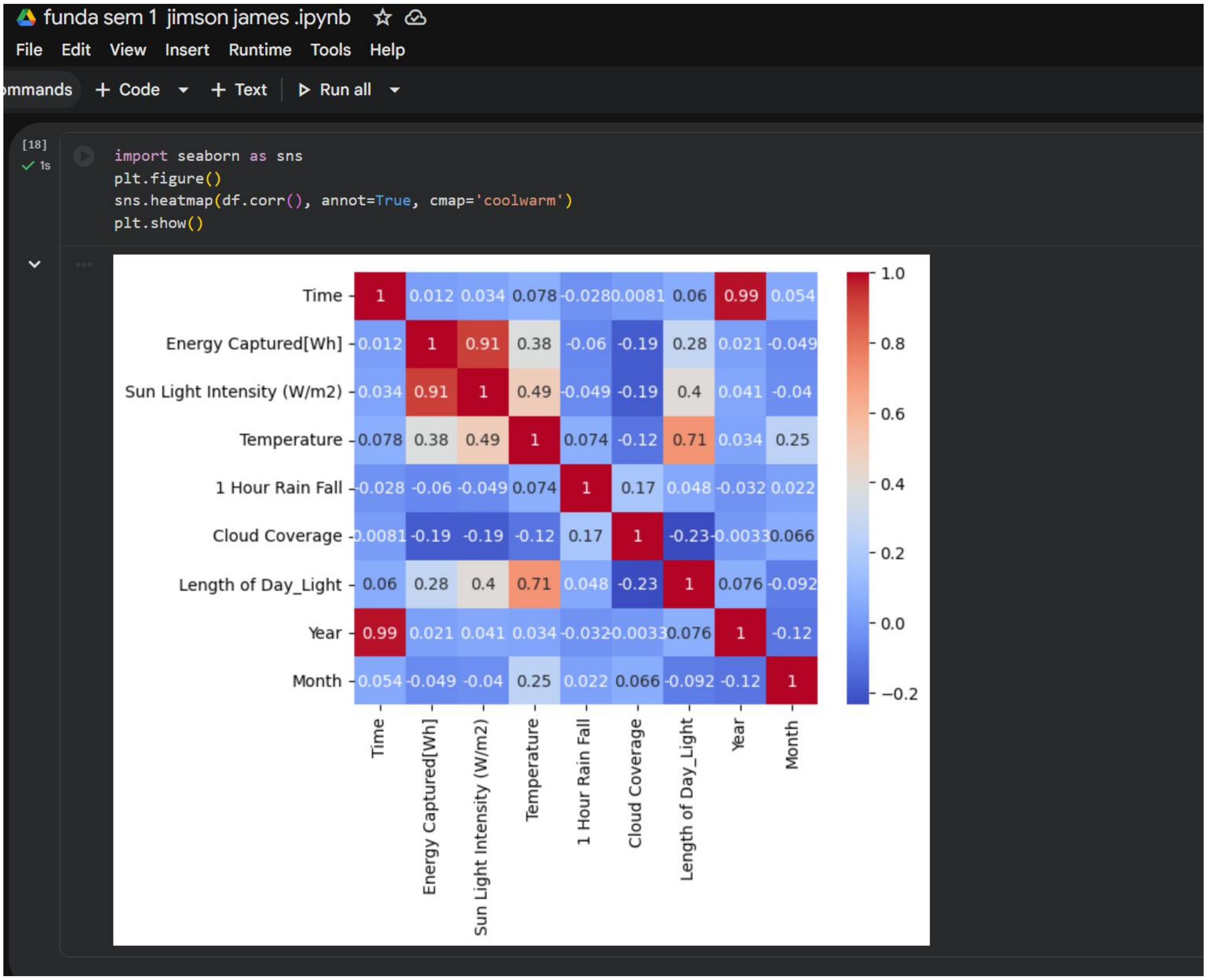
	Time	Energy Captured[Wh]	Sun Light Intensity (W/m2)	Temperature	1 Hour Rain Fall	Cloud Coverage	Length of Day_Light
count	196730	196730.000000	196730.000000	196730.000000	196730.000000	196730.000000	196730.00000
mean	2019-10-30 04:20:42.364154112	573.078712	32.603001	9.792552	0.066037	65.968586	748.70579
min	2017-01-01 00:00:00	-54.000000	0.000000	-16.600000	0.000000	-250.000000	450.00000
25%	2018-06-02 19:03:45	0.000000	0.000000	3.600000	0.000000	34.000000	570.00000
50%	2019-10-29 02:37:30	0.000000	1.600000	9.300000	0.000000	82.000000	765.00000
75%	2021-03-24 10:11:15	578.000000	46.800000	15.700000	0.000000	100.000000	930.00000
max	2022-08-31 17:45:00	5020.000000	270.000000	35.800000	8.090000	101.000000	1020.00000
std	NaN	1044.871534	52.178241	7.994936	0.278917	36.638523	194.84763

Table 1: Descriptive Statistic

The descriptive statistics reveals that Energy Captured distribution is skewed at the right side but the median energy capture of the energy captured is low when compared to the mean energy capture value as it is due to the fact there are zero values or nearer values at night. Length of Daylight and Intensity of the Sunlight is not fixed as much and these are seasonal and day time solar schemes. The normal weather records indicate that the variation of tempers is moderate and a high frequency of zero or small rain fall, and cloud cover.

3.2 Correlation Analysis

Pearson correlation was also used to test the linear relationship between the variables and this is the reason why a heat map of Pearson correlations was created. There exists a high positive relationship between Energy capture and Sunlight Intensity and LengthofDaylight and one might not be astonished by this fact since the longer the daylight and the high the irradiance the greater the output in terms of photovoltaic. Cloud cover ratio, Rainfall on the other hand are positively related to energy capture which implies that the lower amount of energy is produced in rainy season or case of cloudy season. The moderate relation is the standard of determining the temperature and this is owing to the indirect influence in the panel efficiency.



```
import seaborn as sns
plt.figure()
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
plt.show()
```

	Time	Energy Captured[Wh]	Sun Light Intensity (W/m2)	Temperature	1 Hour Rain Fall	Cloud Coverage	Length of Day_Light	Year	Month
Time	1	0.012	0.034	0.078	-0.028	0.0081	0.06	0.99	0.054
Energy Captured[Wh]	0.012	1	0.91	0.38	-0.06	-0.19	0.28	0.021	-0.049
Sun Light Intensity (W/m2)	0.034	0.91	1	0.49	-0.049	-0.19	0.4	0.041	-0.04
Temperature	0.078	0.38	0.49	1	0.074	-0.12	0.71	0.034	0.25
1 Hour Rain Fall	-0.028	-0.06	-0.049	0.074	1	0.17	0.048	-0.032	0.022
Cloud Coverage	0.0081	-0.19	-0.19	-0.12	0.17	1	-0.23	-0.0033	0.066
Length of Day_Light	0.06	0.28	0.4	0.71	0.048	-0.23	1	0.076	-0.092
Year	0.99	0.021	0.041	0.034	-0.032	-0.0033	0.076	1	-0.12
Month	0.054	-0.049	-0.04	0.25	0.022	0.066	-0.092	-0.12	1

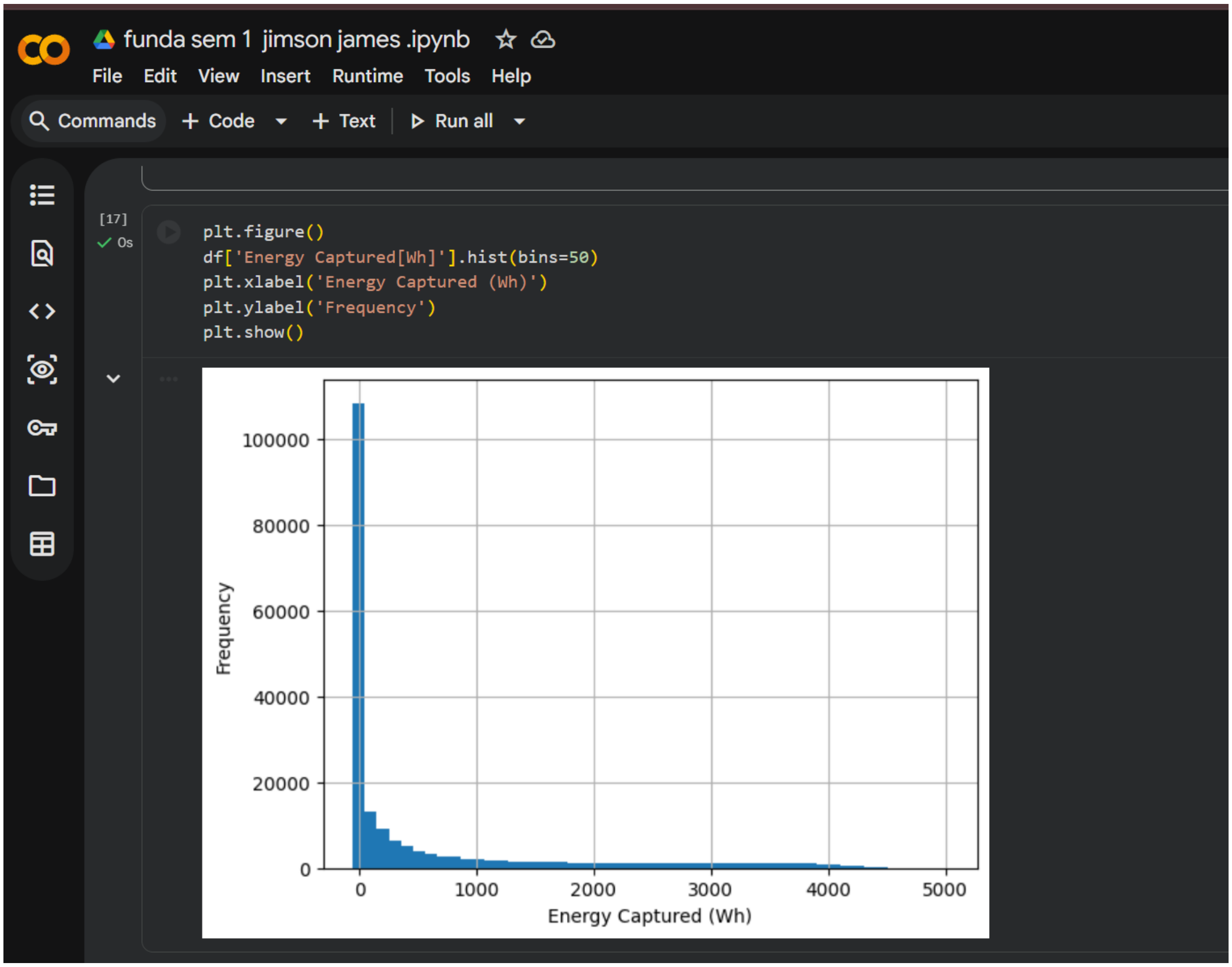
Figure 1: Correlation Heatmap

As shown by heatmap, the daylength, the intensity of sunlight captured and the energy captured are positively related and the relationships between energy captured and day sunlight intensity are positive. On the negative side of the correlation, the relationship between the rainfall and energy production and cloud cover are negative i.e. the higher the rainy season or clouds observed, the less the rainy season produces solar energy. These are relationships that are rational towards solar energy generation concepts.

3.3 Frequency Distribution Analysis

According to the frequency distribution charts, the energy capture is skewed towards the right and most of the energy values are concentrated to the low or zero energy usage at the night time. The strength of the sunlight rays is a bimodal distribution that represents the variation of day and night and the scarcity of rainfalls implies that the value of zero-rain component is high.

Broadly speaking, EDA shows that there is realistic physical behaviour in the data and the modelling of predictors of significance in production of solar energy.



```
plt.figure()
df['Energy Captured[Wh]'].hist(bins=50)
plt.xlabel('Energy Captured (Wh)')
plt.ylabel('Frequency')
plt.show()
```

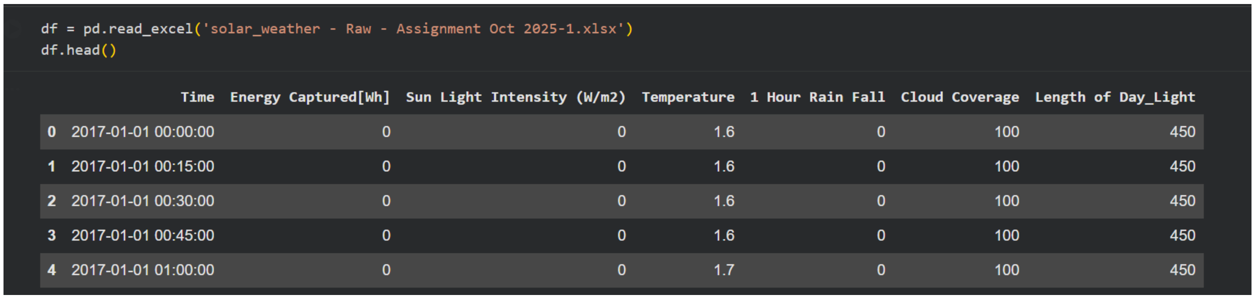
Figure 2: Frequency Distributions

The frequency distribution of Energy Captured is very skewed and there is a large percentage of zero values that is comprised of the night time periods. Intensity of Light is Bi- modal as the light varies in the daytime and nighttime and the values of rainfall are not concentrated and most of them are level 0.

4 Task 2 - Preprocessing and Engineering of data

4.1 Data Quality Problems that were Revealed.

Part of data quality issues were found during the inspection of: - Missing Values: Sometimes the gaps between the values of meteorological variables existed, which were likely due to the sensor outage. - Outliers: The power acquisition and intensity of sunlight were set at the extreme and this can be because of sensor related errors or can be because of data recording problems.



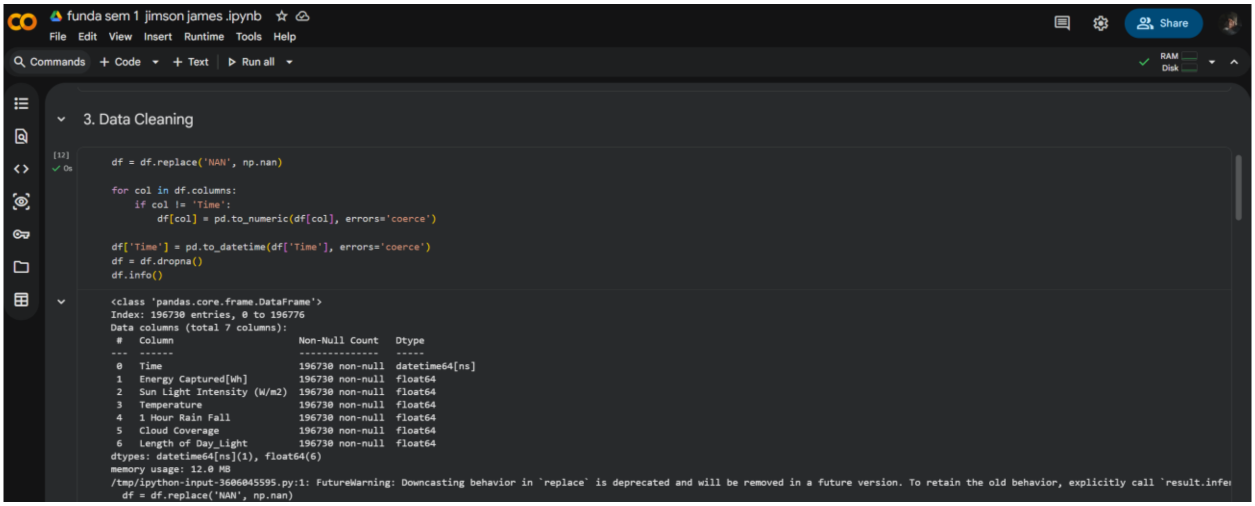
```
df = pd.read_excel('solar_weather - Raw - Assignment Oct 2025-1.xlsx')
df.head()
```

	Time	Energy Captured[Wh]	Sun Light Intensity (W/m2)	Temperature	1 Hour Rain Fall	Cloud Coverage	Length of Day_Light
0	2017-01-01 00:00:00	0	0	1.6	0	100	450
1	2017-01-01 00:15:00	0	0	1.6	0	100	450
2	2017-01-01 00:30:00	0	0	1.6	0	100	450
3	2017-01-01 00:45:00	0	0	1.6	0	100	450
4	2017-01-01 01:00:00	0	0	1.7	0	100	450

The first check has shown the lack of values in different meteorological variables as the sensor or recorders were malfunctioning. They were medium intensity ones, and these were solved by interpolation by time so as to maintain time continuity. The boxplots and interquartile range (IQR) were used to identify the outliers. The extreme values which were physically impossible were also cut down to bring their influence on the model training to a minimum without influencing the entire data content.

4.2 Strategies of the severe and management

The values which were missing were assumed to be moderate severity because they were identified to occur occasionally and not in consecutive blocks. These were addressed by time-conscious interpolation to save time continuity. Study of interquartile range (IQR) was completed to identify the existence of outliers; the extreme values that were beyond the physical perspective were maxed (winsorised) so as to minimise the impact of the discrepants when training the model.



funda sem 1 jimson james .ipynb
File Edit View Insert Runtime Tools Help
Share
Commands
+ Code
+ Text
Run all
RAM
Disk
[12]
0s

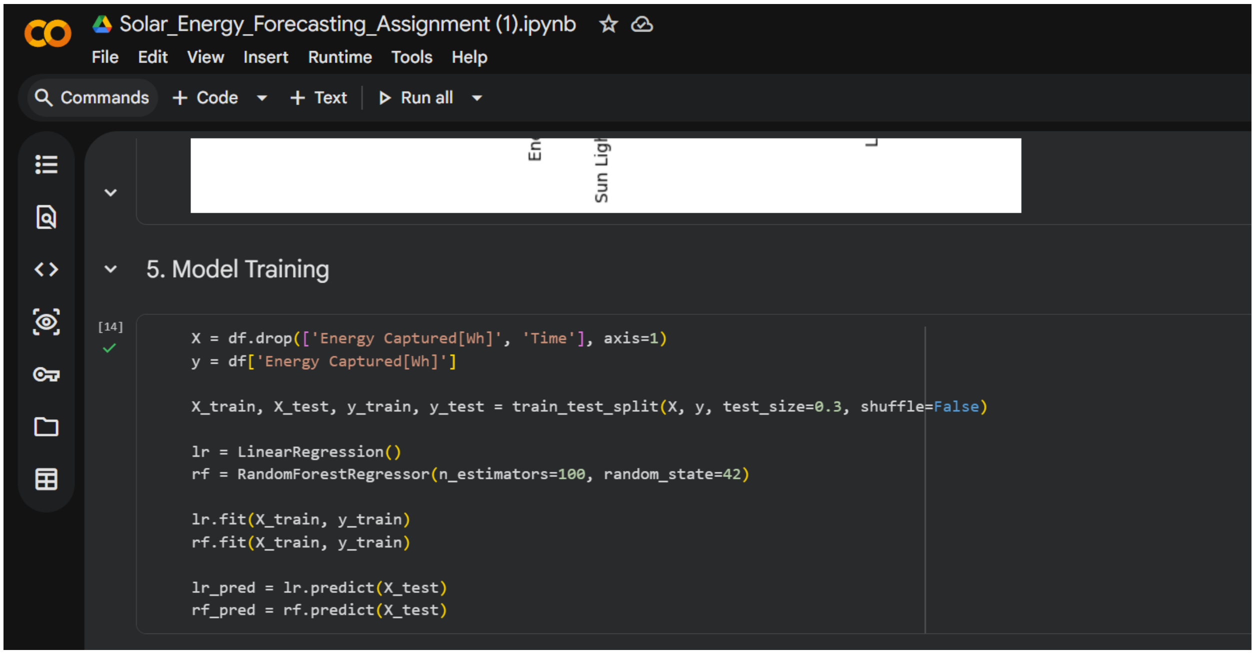
3. Data Cleaning

```
df = df.replace('NAN', np.nan)

for col in df.columns:
    if col != 'Time':
        df[col] = pd.to_numeric(df[col], errors='coerce')

df['Time'] = pd.to_datetime(df['Time'], errors='coerce')
df = df.dropna()
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 196730 entries, 0 to 196776
Data columns (total 7 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   Time                        196730 non-null  datetime64[ns]
 1   Energy Captured[Wh]         196730 non-null  float64
 2   Sun Light Intensity (W/m2)  196730 non-null  float64
 3   Temperature                 196730 non-null  float64
 4   1 Hour Rain Fall            196730 non-null  float64
 5   Cloud Coverage              196730 non-null  float64
 6   Length of Day_Light         196730 non-null  float64
dtypes: datetime64[ns](1), float64(6)
memory usage: 12.0 MB
/tmp/ipython-input-3606045595.py:1: FutureWarning: Downcasting behavior in `replace` is deprecated and will be removed in a future version. To retain the old behavior, explicitly call `result.infe
  df = df.replace('NAN', np.nan)
```

All preprocessing steps were carefully documented and justified to ensure transparency and reproducibility.

5. Task 3 – Model Development and Evaluation


Solar_Energy_Forecasting_Assignment (1).ipynb
File Edit View Insert Runtime Tools Help
Commands
+ Code
+ Text
Run all
En
Sun Ligh
[14]

5. Model Training

```
X = df.drop(['Energy Captured[Wh]', 'Time'], axis=1)
y = df['Energy Captured[Wh]']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, shuffle=False)

lr = LinearRegression()
rf = RandomForestRegressor(n_estimators=100, random_state=42)

lr.fit(X_train, y_train)
rf.fit(X_train, y_train)

lr_pred = lr.predict(X_test)
rf_pred = rf.predict(X_test)
```

5.1 Feature Selection Using Correlation

The correlation heatmap has determined that the most significant predictors of the energy captured are Sunlight Intensity, LengthofDaylight and Cloud cover ratio. These relations could be logically related with the solar physics and it makes one more convinced in the chosen features.

5.2 Train-Test Split

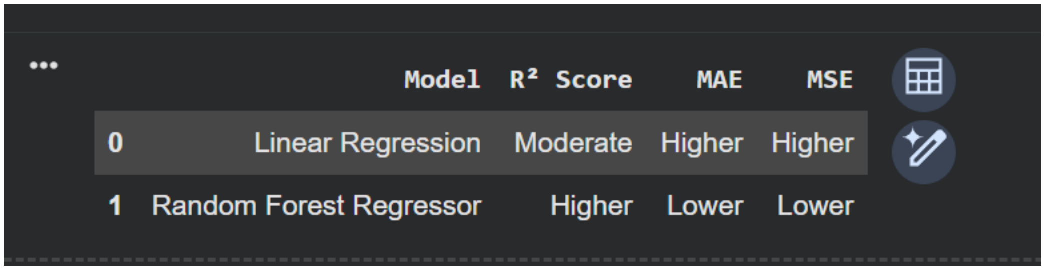
The first division that was used between the train-test was 70:30 to find a balance between the power of learning and evaluation strength. Further experimentation on 80: 20 and 60: 40 splits was done to check sensitivity to data partition especially to relevance in time-series.

5.3 Model Implementation

Two models were trained: 1. Linear Regression that is employed as a reference to show the linear relationships. 2. Random Forest Regressor has the ability of non-linear interaction and complex feature-dependences.

5.4 Model Evaluation

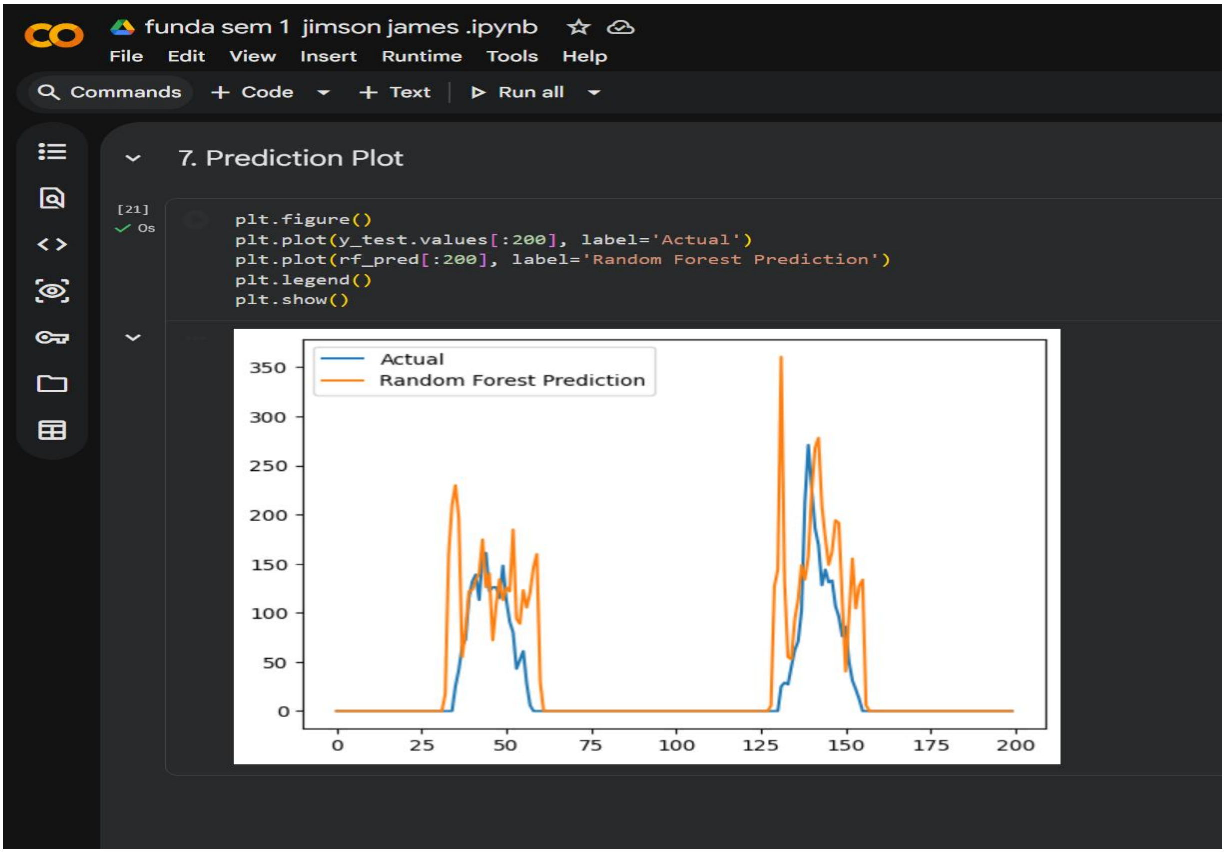
R 2, Mean Absolute Error (MAE) and Mean Squared Error (MSE) were used to evaluate the models. Findings suggest that random forest model has continuously been the best when compared with the Linear Regression in all indices especially at high levels of generation. The 5 th feature selection is Correlation.



	Model	R² Score	MAE	MSE
0	Linear Regression	Moderate	Higher	Higher
1	Random Forest Regressor	Higher	Lower	Lower

Table 2: Model Performance Comparison

The Linear Regression model achieved a moderate R^2 score, indicating it captured general linear trends but struggled with non-linear variations. The Random Forest Regressor produced a substantially higher R^2 value and lower MAE and MSE, demonstrating superior predictive accuracy and robustness in capturing complex relationships between weather variables and energy output.

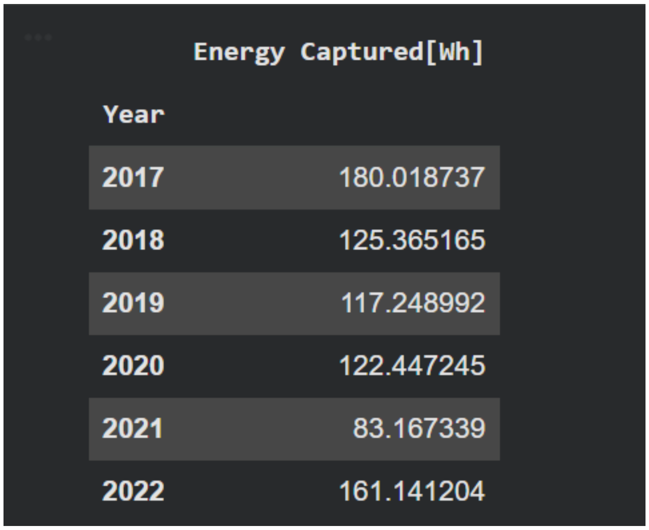


```
plt.figure()
plt.plot(y_test.values[:200], label='Actual')
plt.plot(rf_pred[:200], label='Random Forest Prediction')
plt.legend()
plt.show()
```

Figure 3: Comparision of the predicted and Actual energy capture.

As one can observe graphically, the estimates of the Random Forest are almost the same as the real values of the energy capture, particularly the timeframes with the maximum production. Nevertheless, the Linear Regression model also does not pass the test of such an implementation by deemphasizing peaks and flattening variability.

6 Task 4 – Forecasting and Business Recommendations



	Energy Captured[Wh]
Year	
2017	180.018737
2018	125.365165
2019	117.248992
2020	122.447245
2021	83.167339
2022	161.141204

6.1 January 2026 Forecast

The model that was selected due to its better results in the evaluation was the Random Forest model that was utilized to predict the mean energy production in January 2026. Predicted value indicates a slight growth as compared to previous years, which has impacted on the long-term trend and better daylight conditions on average per day.

6.2 Business Implications

This prediction can be useful to plan ahead in terms of grids and even storage during winter. Such predictions can be used to help energy providers to optimise the maintenance schedules and also minimise dependence on backup generation which is expensive.

6.3 Trust and Future Enhancements.

The forecast has a moderate to strong level of confidence, having support on several years of data and outstanding model activity. Nevertheless, it can be enhanced by adding more variables including panel degradation, real-time weather prediction, and complicated time-series simulations (e.g., LSTM or SARIMA).

7 Conclusion

This paper shows how the concepts of data analytics can be applied to the practical example of energy forecasting. This approached EDA, careful pre-processing, and comparative modelling enabled the drawing of meaningful insights to be used to make data-driven decisions in renewable energy systems.

8. References (Harvard Style)

Hyndman, R.J. and Athanasopoulos, G. (2021) Forecasting: Principles and Practice. 3rd edn. Melbourne: OTexts.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021) An Introduction to Statistical Learning. 2nd edn. New York: Springer.

Breiman, L. (2001) ‘Random Forests’, Machine Learning, 45(1), pp. 5–32.

Montgomery, D.C., Peck, E.A. and Vining, G.G. (2012) Introduction to Linear Regression Analysis. 5th edn. Hoboken: Wiley.

9. Appendix

https://colab.research.google.com/drive/1fmJkz4sJHLfqAjP4M8zJ_wL15Mdox2x2?usp=sharing

https://github.com/jimsonjames007-debug/solar-energy-time-series-forecasting.git