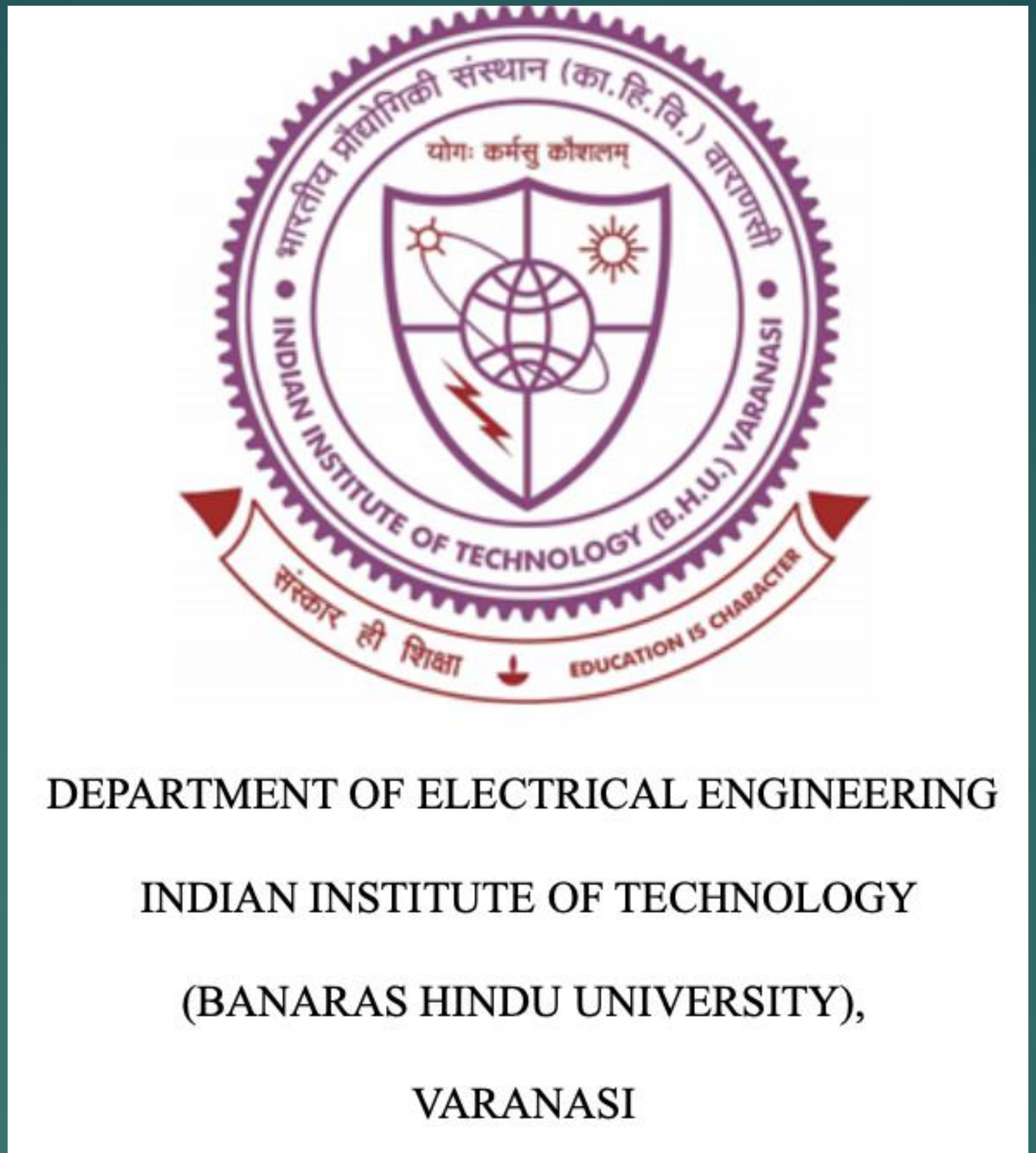


DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY

(BANARAS HINDU UNIVERSITY),

VARANASI

B.Tech Project

Professor Incharge :

Dr. Avirup Maulik sir

Submitted By :

Vaibhav Joshi(20085118)

Vedansh Pandey(20085120)

Navapallav Borthakur(20085123)

INTRODUCTION

- Renewable energy is the type of energy generated from inexhaustible natural resources since it is always regenerated after use. Usage of sources of renewable energy for electricity production helps to reduce greenhouse gases generated by non-renewable energy sources and it also helps to protect
environment by conserving conventional energy sources and renewable
sources are economical as they are available in plenty.

Need for Forecasting

Solar forecasting

Solar Energy is inconsistent in nature as it depends on factors like the position of the sun, time of the day, atmospheric conditions, season, characteristics of solar plant etc. Critics of solar energy claim that it is unreliable in nature. It becomes necessary to forecast solar power generation for the efficient usage of Solar Energy, and for accurate management of loads in combination with a grid. With the help of solar forecasting the unreliability of solar power can be reduced to some extent and it can be used extensively.

Why Machine Learning?

Machine learning models are more accurate than mathematical models because they can learn from data and improve their accuracy over time . Mathematical models are based on assumptions and require a lot of data to be accurate . Machine learning models can learn from data and improve their accuracy over time . They can also handle complex relationships between variables that are difficult to model mathematically . Machine learning models can also be used to identify patterns in data that are not easily visible to humans .

Exploratory Data Analysis

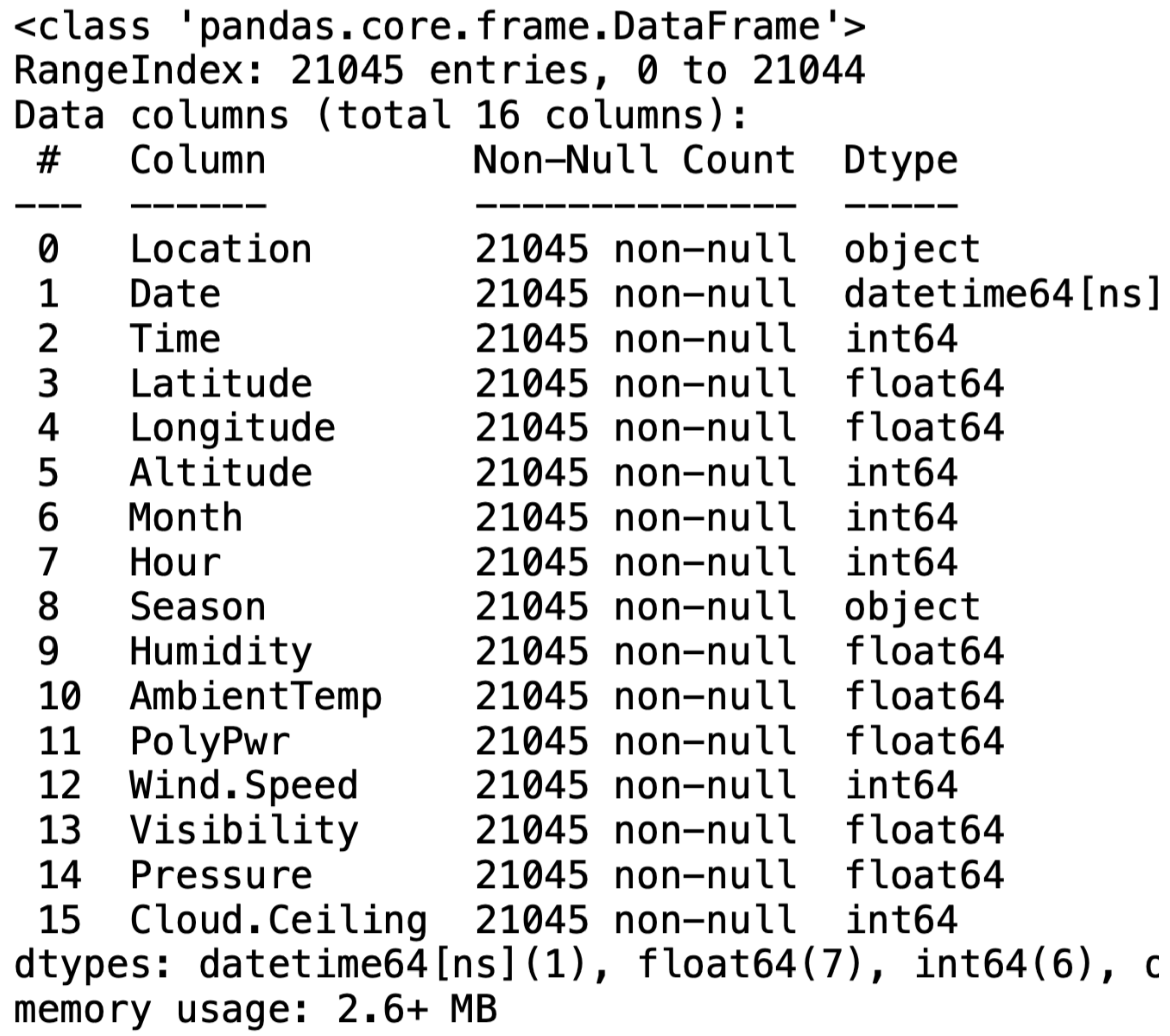
- The dataset consists of 21,045 rows and 16 columns.

- It is a Time Series data - since there’s a time attribute to dataset , hence it need to be sorted according to Time.

- The target variable is PolyPwr (power output).

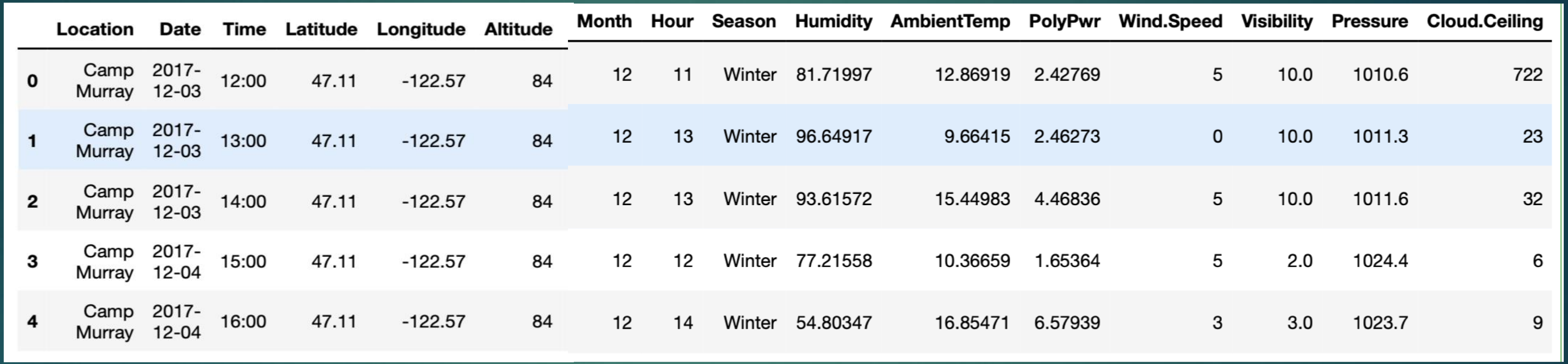
- Dataset Link -

- https://www.kaggle.com/datasets/northern-hemisphere-horizontal-photovoltaic



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21045 entries, 0 to 21044
Data columns (total 16 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Location       21045 non-null  object
 1   Date           21045 non-null  datetime64[ns]
 2   Time           21045 non-null  int64
 3   Latitude       21045 non-null  float64
 4   Longitude      21045 non-null  float64
 5   Altitude       21045 non-null  int64
 6   Month          21045 non-null  int64
 7   Hour           21045 non-null  int64
 8   Season         21045 non-null  object
 9   Humidity       21045 non-null  float64
 10  AmbientTemp    21045 non-null  float64
 11  PolyPwr        21045 non-null  float64
 12  Wind.Speed     21045 non-null  int64
 13  Visibility     21045 non-null  float64
 14  Pressure       21045 non-null  float64
 15  Cloud.Ceiling  21045 non-null  int64
dtypes: datetime64[ns](1), float64(7), int64(6), c
memory usage: 2.6+ MB
```

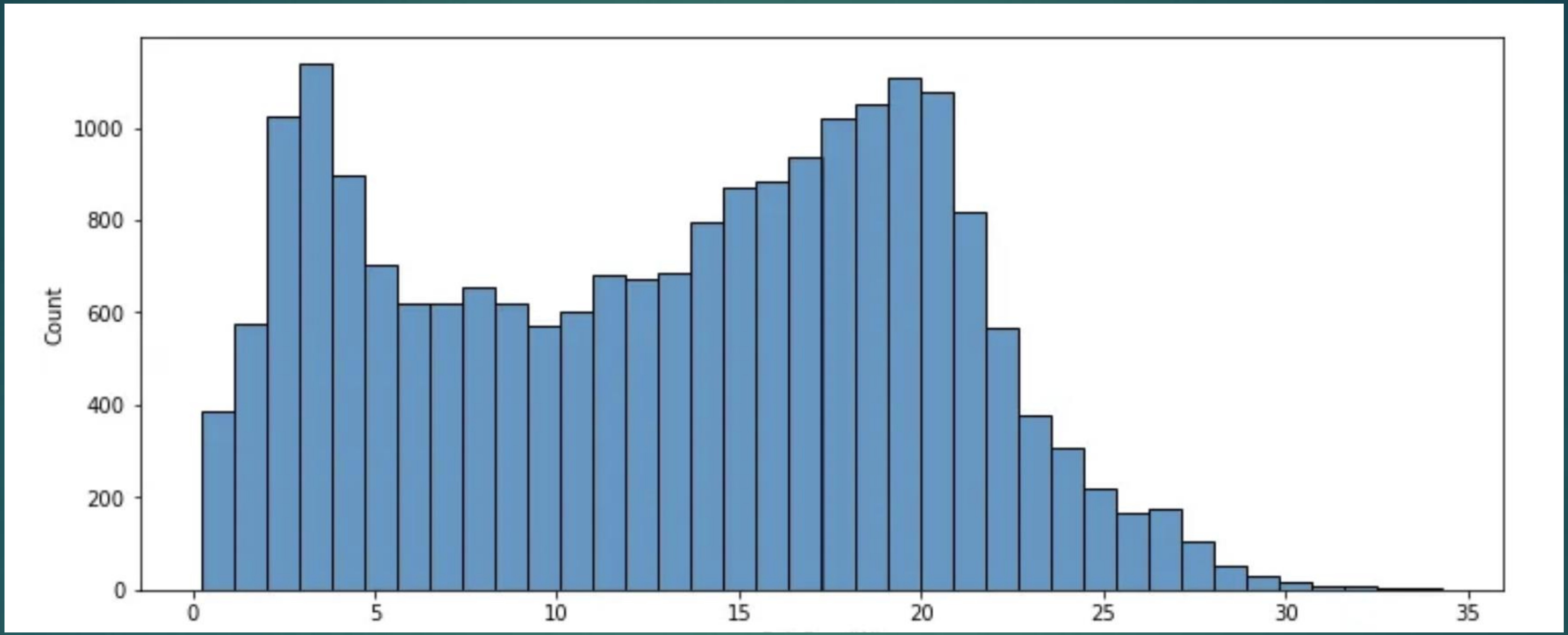
Top 5 rows of data :-



	Location	Date	Time	Latitude	Longitude	Altitude	Month	Hour	Season	Humidity	AmbientTemp	PolyPwr	Wind.Speed	Visibility	Pressure	Cloud.Ceiling
0	Camp Murray	2017-12-03	12:00	47.11	-122.57	84	12	11	Winter	81.71997	12.86919	2.42769	5	10.0	1010.6	722
1	Camp Murray	2017-12-03	13:00	47.11	-122.57	84	12	13	Winter	96.64917	9.66415	2.46273	0	10.0	1011.3	23
2	Camp Murray	2017-12-03	14:00	47.11	-122.57	84	12	13	Winter	93.61572	15.44983	4.46836	5	10.0	1011.6	32
3	Camp Murray	2017-12-04	15:00	47.11	-122.57	84	12	12	Winter	77.21558	10.36659	1.65364	5	2.0	1024.4	6
4	Camp Murray	2017-12-04	16:00	47.11	-122.57	84	12	14	Winter	54.80347	16.85471	6.57939	3	3.0	1023.7	9

We explored the available columns in the dataset using functions in pandas- Python data analysis library.

Distribution of target



PolyPower (MW)

Average Power - 16.32 MW

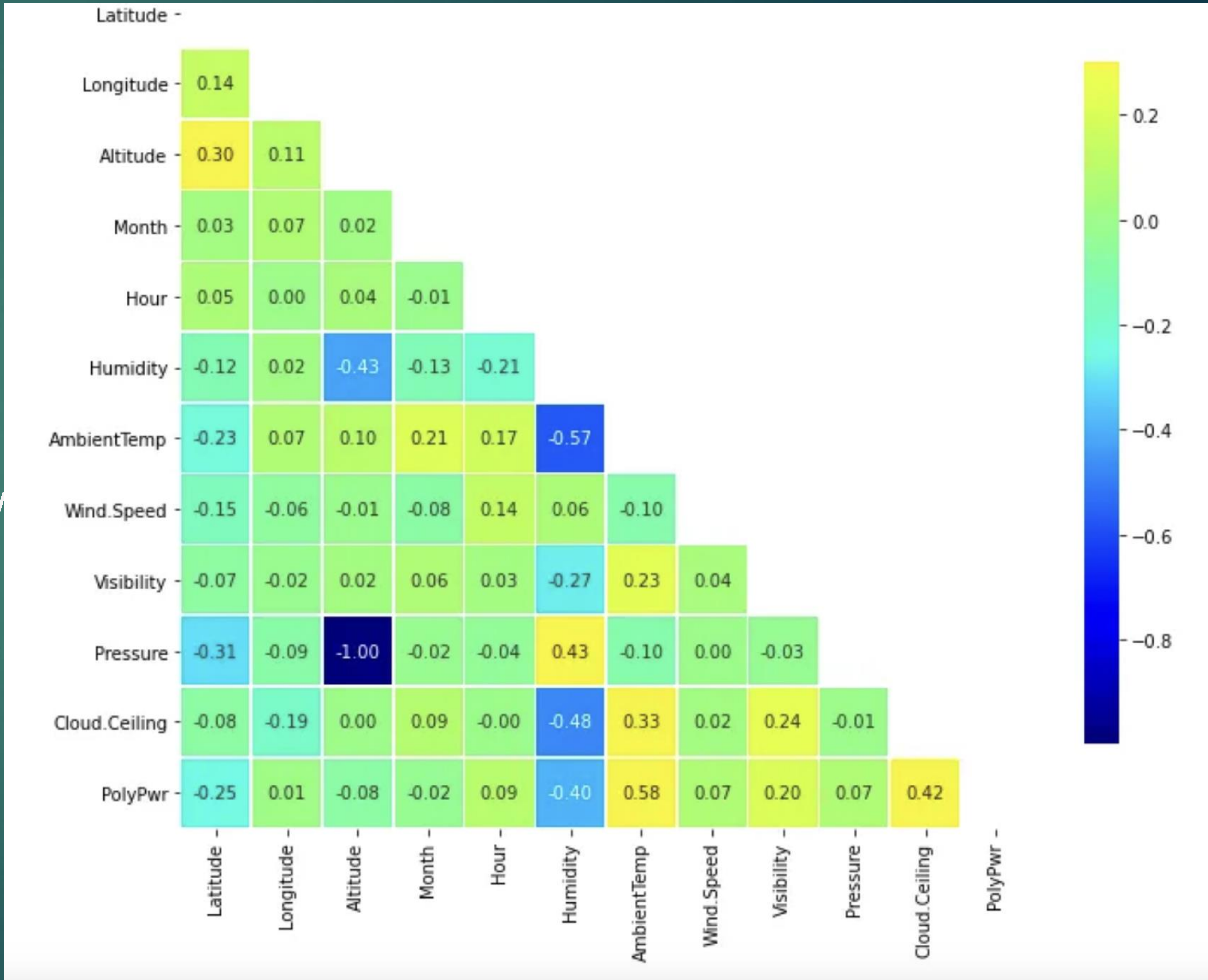
Correlation Matrix

Correlation matrix visualized the correlation between available features and power output.

From the correlation plot, ambient temperature, cloud ceiling, and humidity are the top three most correlated features with solar power output.

It should also be noted that latitude has a significant correlation with power output while longitude does not show the same behavior. Hence, longitude was dropped from the modeling process.

Altitude is also dropped because it has a perfect correlation with pressure but does not vary for a given location.



Data splitting

The entire dataset is split into 80% training data and 20% test data. The test data is held out and unseen throughout the hyper-parameter tuning and training of the different models.

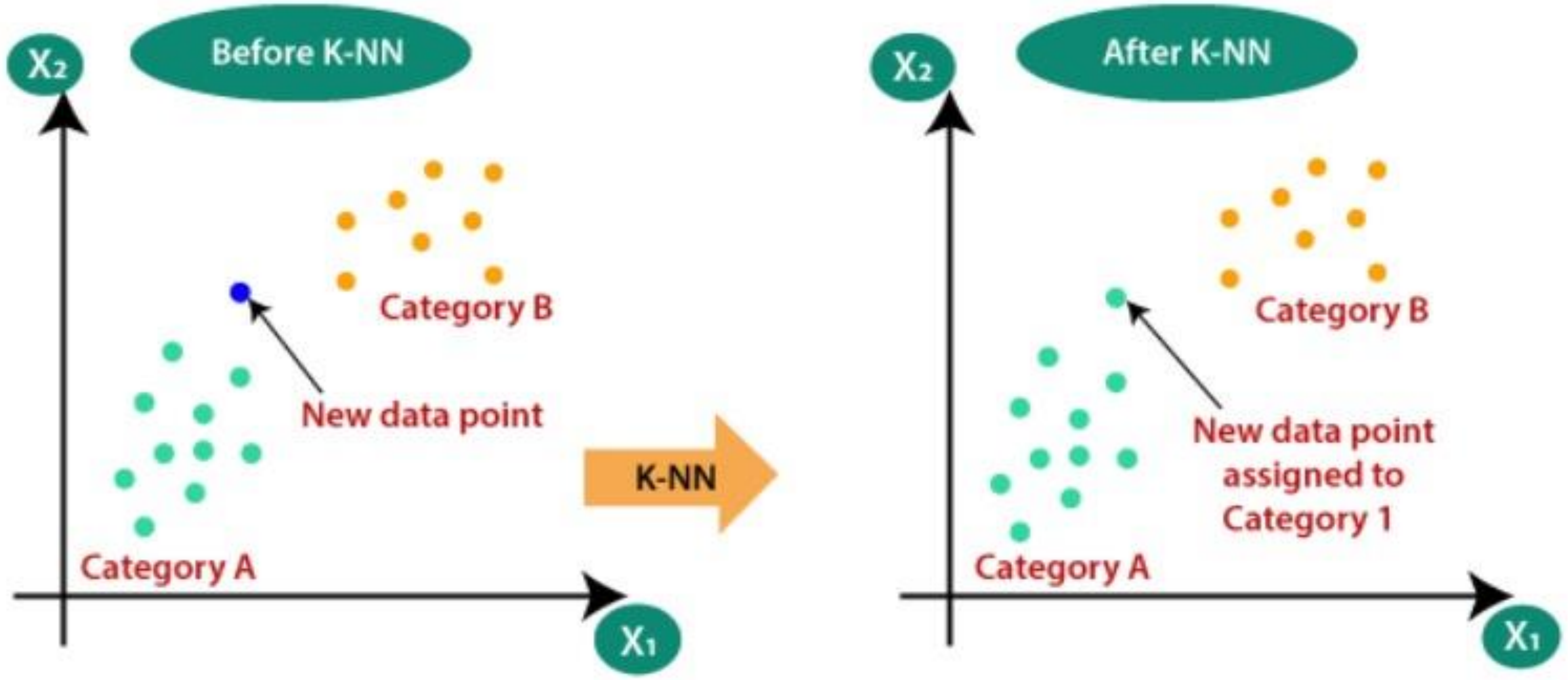
There was no missing data .

KNN

- K-Nearest Neighbours (KNN) is a supervised machine learning algorithm .
- It is a non-parametric and instance-based algorithm.
- The basic idea behind KNN is to find the K nearest neighbour's of a new data point in the training set based on a distance metric (usually Euclidean distance) and then predict value based on the majority or average of the K neighbours.

Advantages :

1. It is simple to implement .
2. It is robust to noisy data.

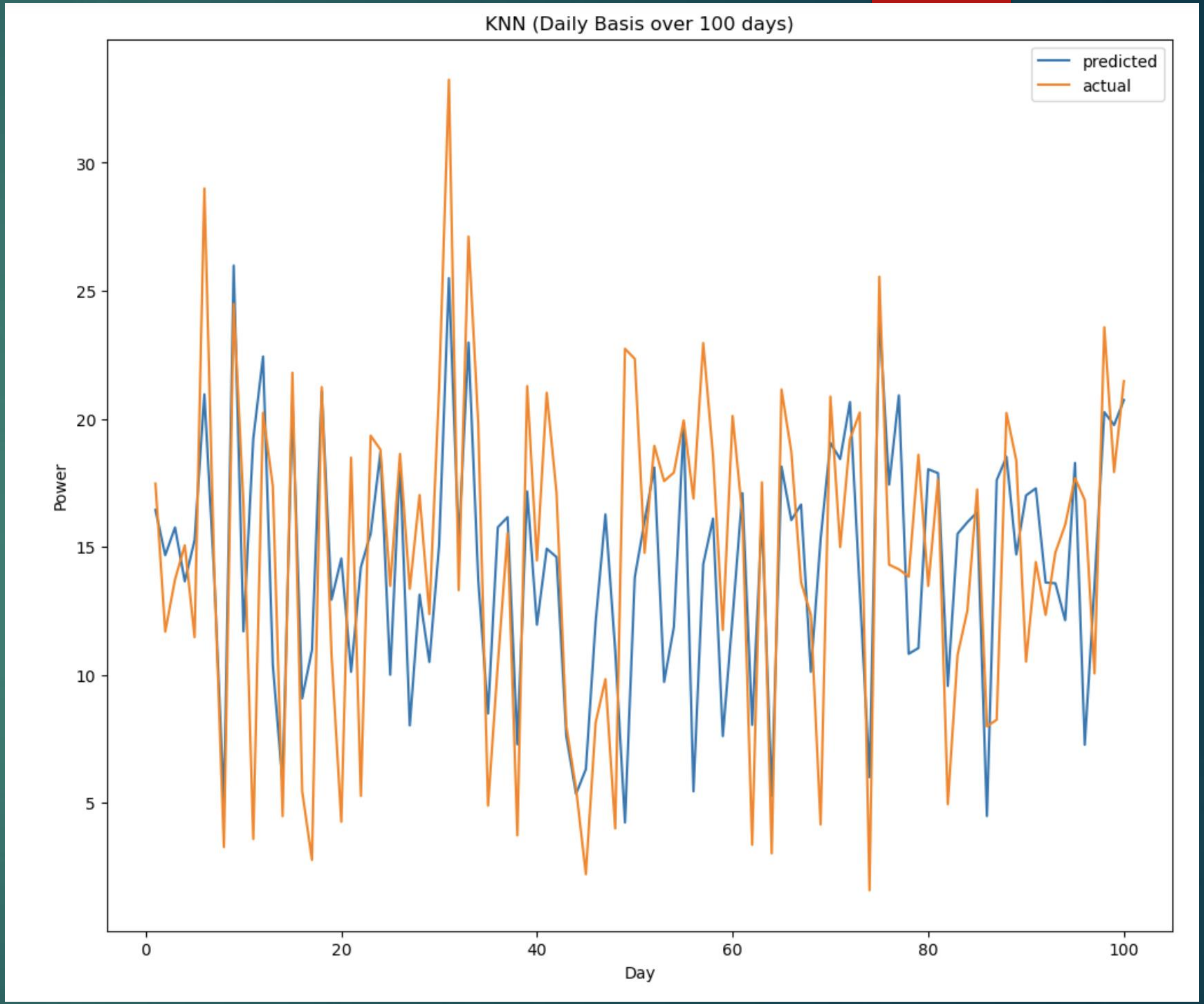


Limitations :

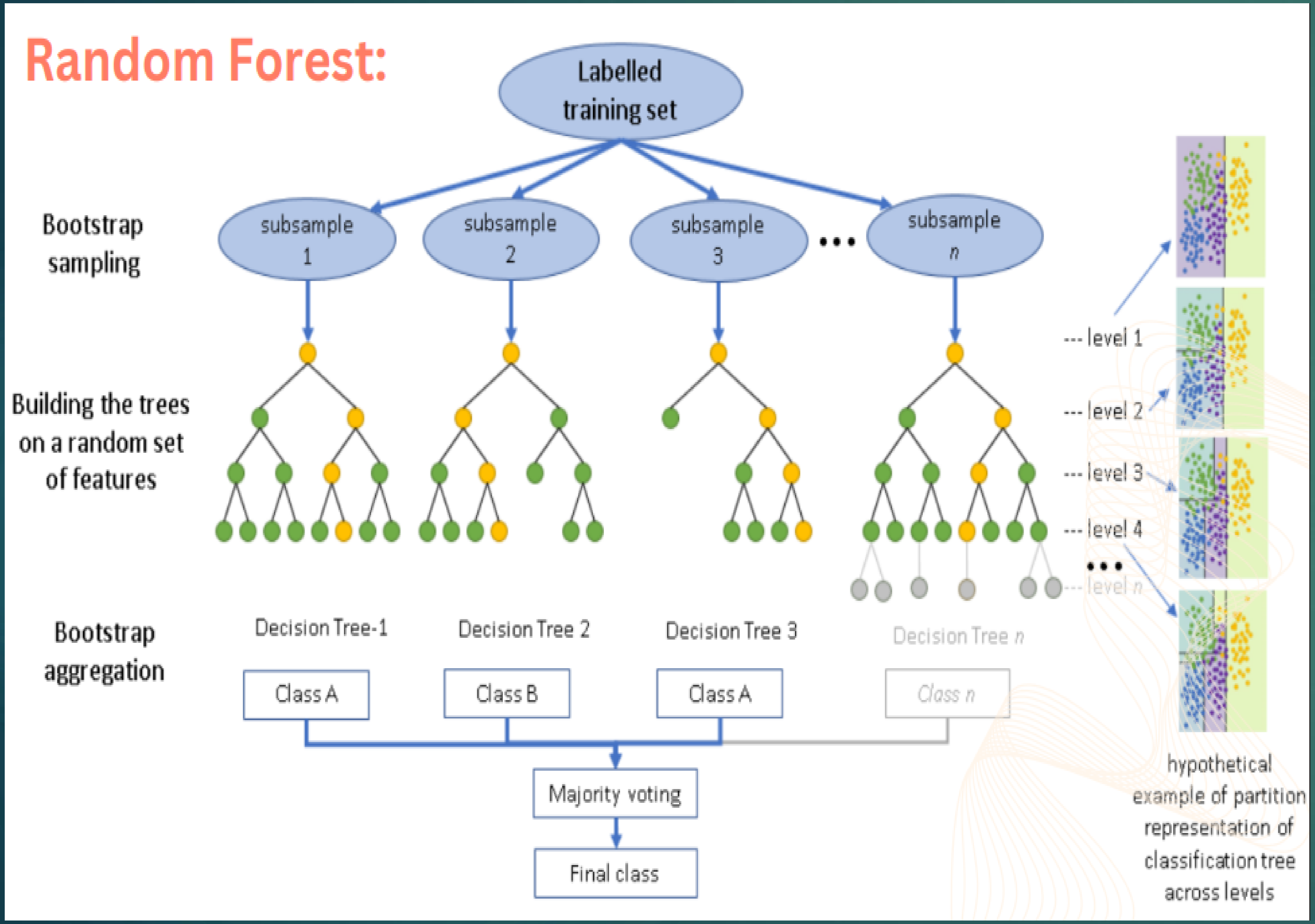
1. It is sensitive to choice of k and distance metrics.

2. The Computation Cost is high because of calculating the distance between the data points of training samples.

Overall, KNN is a useful algorithm for small to medium-sized datasets, especially when the decision boundary is highly irregular or when there is no underlying data distribution.



Random Forest

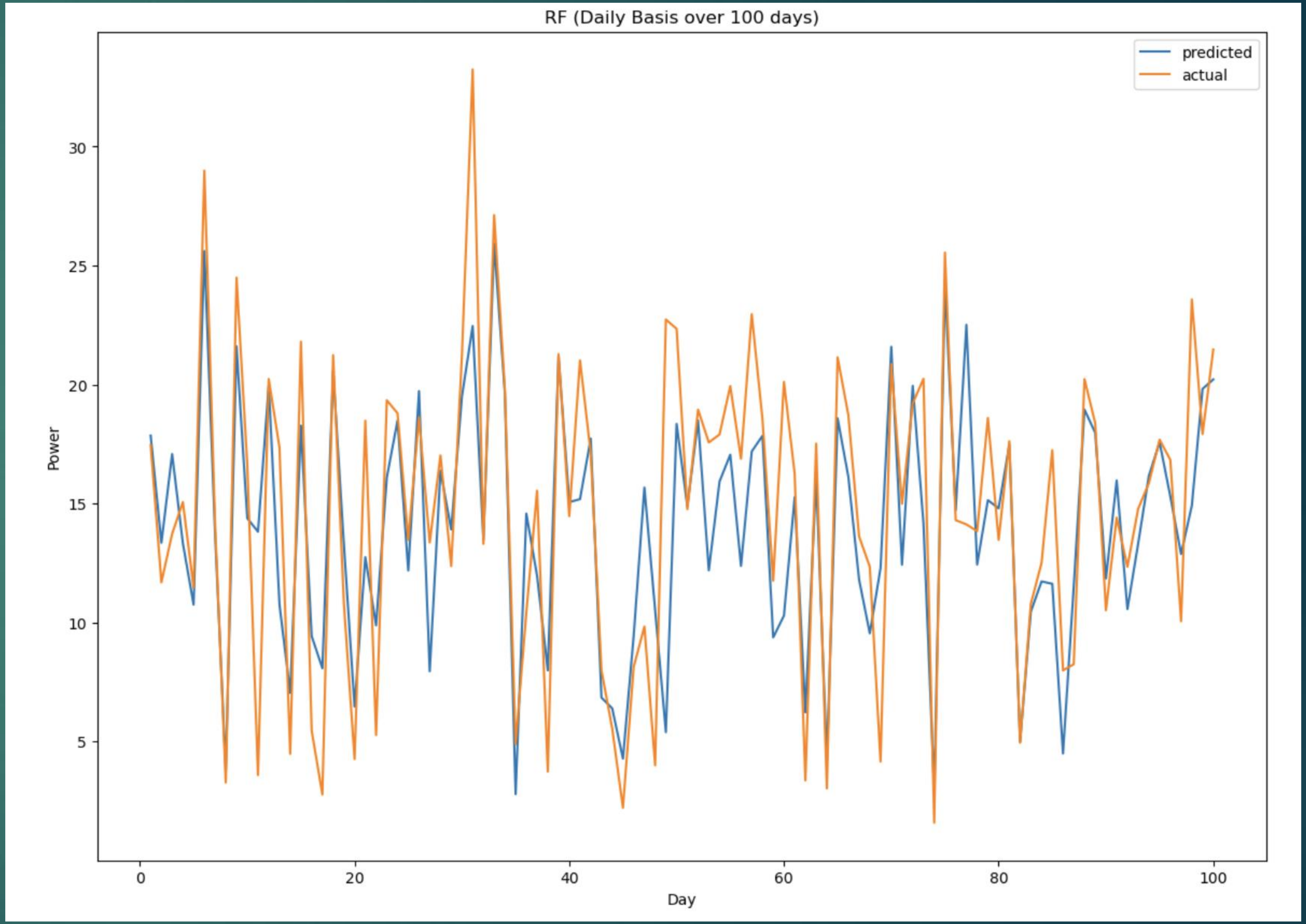


- Random forest is a supervised machine learning algorithm .
- It is an ensemble learning (BootStrap Aggregating) method that builds multiple decision trees and combines their predictions to improve accuracy.
- The basic idea behind random forest is to create a forest of decision trees, where each tree is trained on a random subset of the training data and a random subset of the features. This randomness ensures that the trees are diverse and not highly correlated with each other, which helps to reduce overfitting.

Advantages :

1. It is capable of handling large datasets with high dimensionality.

2. It enhances the accuracy of the model , being robust to overfitting .



Overall, random forest is a powerful and widely used algorithm for classification and regression tasks, especially when dealing with complex and high-dimensional datasets.

Evaluation

- R-Squared

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

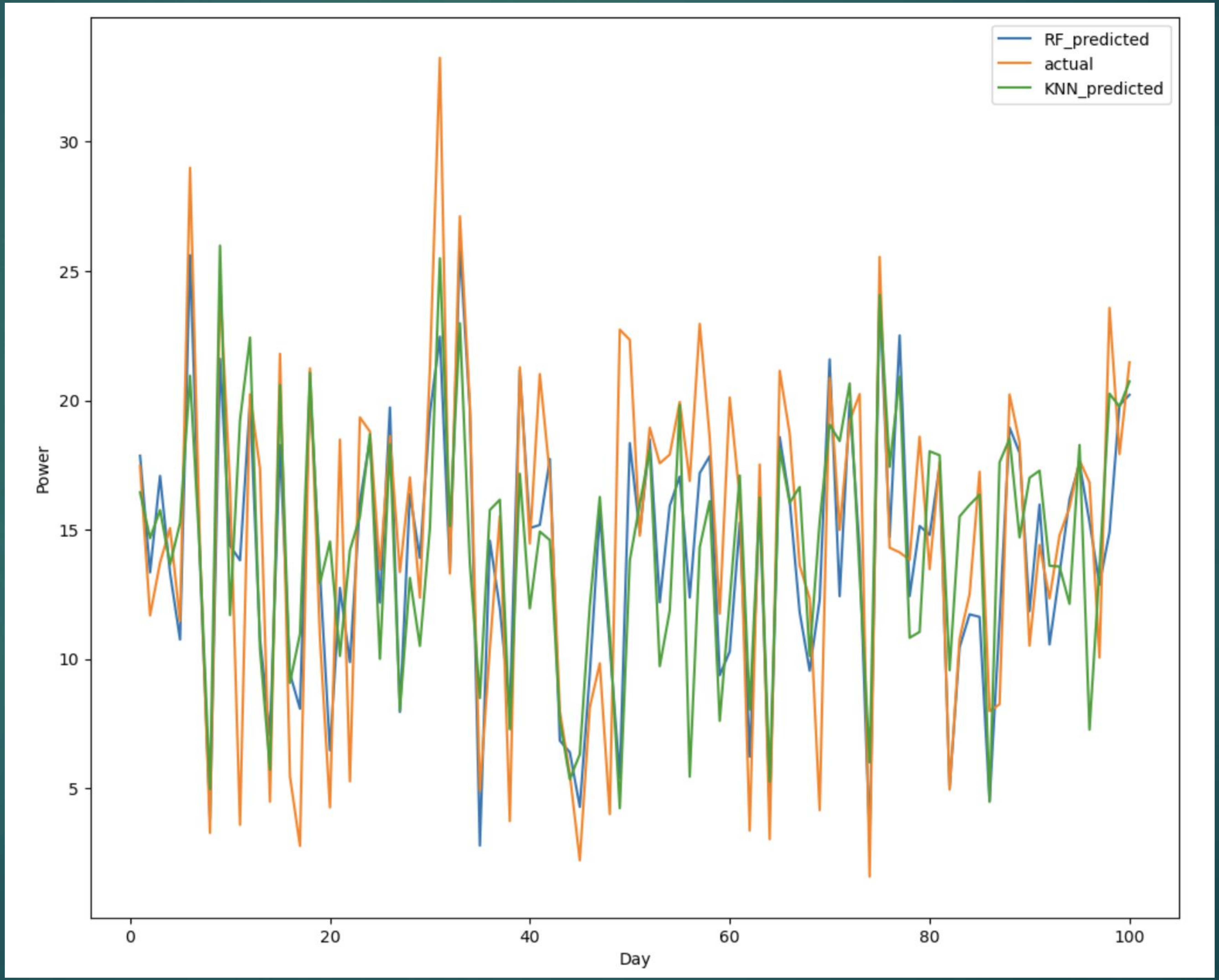
- Root-mean Square error.

$$RMSE = \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2}$$

- Mean Absolute error

$$MAE = \frac{1}{n}\sum_i \left|y_i - \hat{y}_i\right|$$

▶Comparison between RF and KNN



2. Test Data Scores :

The performance of each model is evaluated using the hold-out set which is 20% of the entire dataset.

The results are summarized below:

Model	R^2	RMSE	MAE
KNN	0.618	4.403	2.971
Random Forest	0.620	4.095	2.787

The RF model has the overall best performance with a 5.2% improvement compared with the KNN (baseline) model.

References

- Dynamic Forecasting of Solar Energy Microgrid Systems Using Feature Engineering

  Muamar Mohamed , Farhad E. Mahmood , Mehmmood A. Abd, Ambrish Chandra , Fellow, IEEE,

  https://ieeexplore.ieee.org/document/9684273

- A New Probabilistic Ensemble Method for an Enhanced Day-Ahead PV Power Forecast

  Silvia Pretto, Emanuele Ogliari , Member, IEEE, Alessandro Niccolai

  https://ieeexplore.ieee.org/document/9858626