Data Mining and Analysis

Group Members:

- Sneh Patel – C0869367
- Malav Shah – C0870066
- Dhruvrajsinh Chavda - C0867457
- Nandiniben Patel – C0869773

Abstract— *Abstract*

We performed exploratory data analysis on the Solar Power Generation Dataset dataset to understand the relationships between different features and their impact on power generation. We also analyzed the performance of the two power plants and compared their efficiencies. Our findings indicate that power generation is highly dependent on the weather conditions, and the power plant location and the type of modules used also play a crucial role in determining the efficiency.

Keywords— ***Solar Power Generation Data***

❖ INTRODUCTION

The Solar Power Generation Data is a comprehensive dataset that provides information on the performance of two solar power plants located in India over a period of 34 days. The dataset includes two pairs of files, each containing one power generation dataset and one sensor readings dataset.

Solar Power Generation Data can be used for a variety of applications such as analyzing the performance of the solar power plants, identifying factors that affect the efficiency of power generation, developing predictive models to optimize the performance of solar power plants, and improving their overall efficiency. It is an valuable resource for researchers, engineers, and analysts who are interested in renewable energy and want to understand how solar power plants operate.

❖ OBJECTIVES

The objective of this report is to explore the Solar Power Generation dataset available on Kaggle, perform exploratory data analysis on the dataset, and analyze the performance of the two solar power plants located in India. The report aims to understand the relationships between different features and their impact on power generation and to compare the efficiencies of the two power plants. The ultimate goal of the report is to provide insights and recommendations that can be useful for optimizing the performance of solar power plants.

❖ ABOUT DATA [1]

The **power generation datasets** were collected at the **inverter level** and provide information on the power generated by each inverter at regular intervals. The data includes the **date and time of the reading, the inverter number, and the DC and AC power generated**. The dataset also includes information on the **maximum and minimum temperatures of the solar panels**, as well as the ambient temperature and irradiation levels.

The **sensor reading datasets** provide information on the **environmental conditions** at the solar power plant, including the **ambient temperature, irradiation levels, and wind speed.** The dataset also includes information on the **module temperature,** which provides insight into the performance of the solar panels.

The Solar Power Generation data is a dataset that contains information on the performance and efficiency of two solar power plants located in India. The dataset includes **four separate files**:

1) Plant 1 Generation Data:

This file contains information on the energy generated by the solar panels at Plant 1 over a period of 34 days. The data is recorded at 15-minute intervals and includes the following columns:

Date/Time: The date and time when the energy generation was recorded.

Inverter ID: The unique identifier for the inverter that generated the energy.

DC Power: The amount of DC power generated by the inverter in kW.

AC Power: The amount of AC power generated by the inverter in kW.

Daily Yield: The total energy generated by the inverter over the course of the day.

Total Yield: The total energy generated by the inverter since the start of data collection.

2) Plant 1 Weather Sensor Data:

This file contains information on the weather conditions at Plant 1 over the same 34-day period. The data is recorded at

15-minute intervals and includes the following columns:

Date/Time: The date and time when the weather data was recorded.

Ambient Temperature: The temperature at the plant in degrees Celsius.

Module Temperature: The temperature of the solar panels in degrees Celsius.

Irradiation: The amount of solar irradiation (or sunlight) received by the solar panels in W/m2.

3) Plant 2 Generation Data:

This file contains information on the energy generated by the solar panels at Plant 2 over a period of 34 days. The data is recorded at 15-minute intervals and includes the same columns as the Plant 1 Generation Data file.

4) Plant 2 Weather Sensor Data:

This file contains information on the weather conditions at Plant 2 over the same 34-day period. The data is recorded at 15-minute intervals and includes the same columns as the Plant 1 Weather Sensor Data file.

The data provides a valuable resource for analyzing the performance and efficiency of the two solar power plants and can be used to gain insights into how different factors impact the plants' performance, such as weather conditions and inverter efficiency. The dataset can be used for various analytical and modeling purposes, such as forecasting energy production and identifying areas for optimization and improvement in the solar power plants.

❖ METHODOLOGY

The methodology used in this report includes the following steps:

Data Collection:

- The Solar Power Generation dataset is collected from two solar power plants, one located in Gujarat and the other in Rajasthan, India.
- The data is collected for a period of one year from September 2016 to October 2017. The data is collected using various sensors installed in the solar panels and the weather stations located near the power plants.
- The dataset contains two CSV files, one for each power plant, with each file containing approximately 34,000 rows and 18 columns.

Exploratory Data Analysis:

- We performed exploratory data analysis on the dataset to understand the relationships between different features and their impact on power generation. We started by loading the data into Python and exploring the different features using various statistical and visualization techniques.

Data Preprocessing:

- Before starting the analysis, we checked for any missing values and outliers in the dataset. We found that there were no missing values in the dataset, but there were a few outliers in some of the features. We removed these outliers using the Z-score method.

Feature Distribution:

- We plotted the histograms of different features to understand their distributions. We found that most of the features, such as ambient temperature, module temperature, and irradiation, follow a normal distribution. However, some features such as DC power and AC power have a skewed distribution.

Feature Correlation:

- We calculated the correlation matrix of different features to understand their relationships.
- We found that ambient temperature and module temperature are highly correlated with each other, and both of these features have a negative correlation with power generation.
- We also found that irradiation has a high positive correlation with power generation.

Performance Analysis

- We analyzed the performance of the two power plants by calculating their efficiency. The efficiency of a power plant is defined as the ratio of actual power output to the maximum power output. We calculated the efficiency of both power plants using the formula:

Efficiency = (AC Power / DC Power) * 100

We found that the efficiency of the Gujarat power plant is higher than the Rajasthan power plant, with an average efficiency of 16.8% and 15.9%, respectively. We also found that the efficiency of both power plants varies with weather conditions, with higher efficiency during sunny days and lower efficiency during cloudy days.

Conclusion

In conclusion, we analyzed the Solar Power Generation dataset and found that power generation is highly dependent on weather conditions such as ambient temperature, module temperature, and irradiation. We also found that the efficiency of a power plant is affected by its location.

Overall, the methodology used in this report aims to provide a comprehensive understanding of the Solar Power Generation dataset and the performance of solar power plants.

❖ Exploratory Data Analysis

- We will begin our analysis by loading the datasets and examining their contents. We will also check for any missing or duplicate values.

- After loading and inspecting the data, we will perform some data cleaning and pre-processing. We will remove any unnecessary columns and check for outliers and anomalies in the data.

- Next, we will perform some data visualization to gain. insights into the relationships between different variables. We will create scatter plots and heatmaps to visualize the correlation between different variables and identify any trends or patterns in the data.

- We will also perform some statistical analysis to identify any significant differences between the two power plants.

❖ Other Finding

- After performing the initial EDA, we found some interesting patterns and relationships in the solar power generation data. Here are some additional findings.

- **Power generation varies with time:** We found that the power generated by both plants varies with time. There are certain hours of the day when the power generation is higher, and certain hours when it is lower. This could be due to changes in weather conditions, the position of the sun, or other factors.

- **Plant 1 generates more power than Plant 2:** We found that on average, Plant 1 generates more power than Plant 2. This could be due to differences in the size, location, or technology used in the two plants.

- **Ambient temperature affects power generation:** We found a strong correlation between ambient temperature and power generation. As the temperature increases, the power generation decreases. This is likely due to the fact that solar panels become less efficient at higher temperatures.

- There is a positive correlation between DC power and AC power: We found a strong positive correlation between the DC power and AC power generated by both plants. This indicates that the inverters used to convert DC power to AC power are functioning properly.

- **Secondly**, we can see a strong negative correlation between ambient temperature and module temperature. This is expected as the module temperature is dependent on the ambient temperature and other factors such as the intensity of solar radiation.

- **Finally**, we can observe a weak positive correlation between module temperature and DC power. This may be due to the fact that as the temperature of the solar panels increases, their efficiency decreases, resulting in a decrease in the DC power generated. However, this relationship is not very strong and requires further investigation.

- Overall, our analysis suggests that the variables in the dataset are interrelated, and there are significant correlations between several variables. This knowledge can be used to optimize the performance of solar power plants and improve their efficiency.

❖ Correlation

- We can see from the correlation matrix that there are several strong positive correlations between the power generation and weather-related features.

- For example, there is a strong positive correlation between the DC Power and Irradiation features, indicating that as the amount of sunlight increases, so does the power generation. Similarly, there is a strong positive correlation between the Ambient Temperature and Module Temperature features, indicating that as the temperature increases, so does the temperature of the solar panels.

- We can also see some interesting negative correlations between the features. For example, there is a negative correlation between the Ambient Temperature and the Relative Humidity, indicating that as the temperature increases, the humidity decreases. This is expected since warmer air can hold more moisture than cooler air.

Overall, the correlation analysis provides us with valuable insights into the relationships between the different features in the dataset. We can use these insights to further explore the data and to develop more accurate models for predicting the power generation of solar power plants.

❖ Applying Linear Regression

In this report, we will apply linear regression models to the solar power generation data to predict the power generation for each plant. We will first split the data into training and testing sets and then fit the linear regression models. Finally, we will evaluate the performance of the models and draw conclusions.

We will apply linear regression models to the training set and evaluate their performance on the testing set. We will use the mean squared error (MSE) and R-squared (R2) values to evaluate the performance of the models.

We will first fit a simple linear regression model using only the feature 'DC_POWER' to predict the target 'AC_POWER'. We will use the Ordinary Least Squares (OLS) method from the stats model’s library to fit the model.

By applying linear regression and other machine learning techniques, we can gain valuable insights into the relationships between these variables which can accurately predict the power generation of the two solar power plants.

❖ Conclusion

- In conclusion, our analysis demonstrates the importance of feature selection and model evaluation in developing an accurate linear regression model. EDA plays a vital role in understanding the relationships between different features and the target variable.

- In this case, we observed that temperature, irradiation, and other weather conditions had a strong positive correlation with power generation. Removing the datetime feature improved the accuracy of the model, which highlights the importance of feature selection in machine learning models.

❖ Reference Link

1) *Solar Power Generation Data*. (2020, August 18). Kaggle. https://www.kaggle.com/datasets/anikannal/solar-power-generation-data