**Solar Power Generation**

**Project Description and Goals:**

This data analysis and machine learning project investigates relationships between solar panel performance and environmental conditions for 2 solar panel arrays based in India, using a dataset made available on Kaggle.

The goal of Solar Power Generation project is to allow us to learn data analysis and visualization. We will be using machine learning algorithms to correctly predict the power generation for a particular time in the future as well as identify the need for panel cleaning and maintenance and identify faulty or suboptimal performing equipment. The topics that we will be practicing are computer vision fundamentals which includes methods that deal with regression and classification KNN (K- nearest neighbors).

**Details about the Data:**

Data has been gathered at two solar power plants in India over a 34-day period. Power generation datasets are gathered at the inverter level, with each inverter connected to multiple lines of solar panels, whilst sensor data is gathered at the plant level, with a single array of sensors optimally placed within the plant.

Three questions have been asked to analyze:

1. Can we predict the power generation for next couple of days?
2. Can we identify the need for panel cleaning/ maintenance?
3. Can we identify faulty or suboptimal performing equipment?

**The developed methods, algorithms, and tools to address the project’s requirements:**

In order to visualize the data we implemented matploptlib.pyplot which is plotting library which produces quality figures in a variety of hard-copy formats and interactive environments to display the hand drawn image by using the pixel values from the features.

Other algorithms we used are random forest, decision tree, K-nearest neighbors, and logistic regression along with their appropriate classifiers. The KNN method of classification is one of the simplest methods in machine learning its most basic level of machine learning, it is essentially classification by finding the most similar data points in the training data and making an educated guess based on their classifications. Even if it seems simple to understand and implement, the KNN method has seen wide application. Random forest is a supervised learning algorithm, and it creates a forest and makes tries to make it as random as possible, which can be used for both classification and regression problems used widely in machine learning.

Decision Trees are very commonly used tool in data mining to achieve a particular goal or target at each node, widely used in machine learning. However, most of the time trained with the “bagging” method. The idea of “bagging" method is that of a combination of learning models to increase the overall result.

Logistic regression is a very powerful algorithm it looks at the relationship between a training variables and testing variables. Provided the right representation of the labels and features it can return very accurate results which makes it an accurate machine learning algorithm.

To start our project we begin by creating a github repository and adding all csv data from Kaggle competition so we can read the data and put it in data frames and then use it to train and test.

**The developed codes and final results:**

|  |  |
| --- | --- |
| **Method** | **Accuracy / Results** |
| Linear Regression | [30.00734641, 29.69616069, 29.2317384 , ..., 20.0969496 ,  19.97594227, 19.78807447]) |
| Mean absolute error plant1 and 2 Ambient Temperature | 4.86314920476783 |
| Mean Absolute error Module Temperature and Irradiation | 30.993863394682236 |
|  |  |
|  |  |
|  |  |

**Code and Visualization:Graphical user interface, application

Description automatically generated**

**Team Members:**

Cristina Munteanu

Chandel Buelna

Samantha Simpson

Dakota Townsend

David Melendez

**Responsibilities:**

Project Report – Cristina Munteanu

Can we identify the need for panel cleaning/maintenance? - Cristina Munteanu

Project Slides – Samantha Simpson

Can we predict the power generation for next couple of days? - this allows for better grid management – David Melendez

Can we identify faulty or suboptimal performing equipment?