**1. Project Title and Team Members:**

**Solar Energy Prediction using Machine Learning**

**Shrinivas Gollalappa Kadaganchi**

**Chaitanya Phani Kumar Akula**

**Vamsi Inampudi**

**2. Goals and Objectives:**

* Objectives:
  + Constructing machine learning models to predict the generated and exported energy using weather parameters and derived features of energy at a particular Photovoltaic power plant.
  + Observing the variability in the data and taking measures to reduce the variability.
  + Applying various ensemble models like bagging regression models, boosting regression models and stacking regression models to reduce the variability of the data and to improve the efficiency of the predictions.
  + Comparing the performance of the bagging, boosting and stacking ensemble models.
  + Analyzing the best fit line of predictions of the models using scatter plot, feature importance plot and permutation importance test.
* Significance:
  + Predicting the solar energy (exported and generated) is a challenging task due to high variance in data. Ensemble models like bagging, boosting and stacking models help to reduce the variance in the data. This can be achieved through aggregation of predictions and by incorporating the diverse models.
  + Achieved r2 is 0.98 on both exported and generated energy prediction tasks.
* Motivation:
  + The motivation for this idea derives from the growing use of renewable energy plants, such as solar power plants.With the increased use of power plant setup, there is a great need to create cost-effective setups. To design a cost effective set up knowing output or generated and exported energy is important. Existing systems failed to capture the data complexity. Therefore building an effective machine learning model to understand and predict the real time data is important.
* Features:
  + Main features of the project are inducing diversity to improve model generalization on new data. In the bagging method multiple individual models or similar models are composed and trained on randomly sampled subset of data and in the end the predictions are aggregated to produce accurate predictions. In boosting methods weak learners are trained sequentially improving the predictions of each step and in the end the predictions are produced based on weightage. In the stacking method multiple weak learners or base models are trained and the collected predictions are passed to the meta model for final predictions.

**Dataset :**

Data in this project is collected from the source PV-Output.org. The original data is maintained by solar energy department Queensland, Australia. The dataset contains the 21 files which is the consolidation of different profiles who installed the Photovoltaic plants at their business or home. This is made available through github[https://github.com/gomesramos/PV-Output-Datasets](https://github.com/gomesramos/PV-Output-Datasets.)

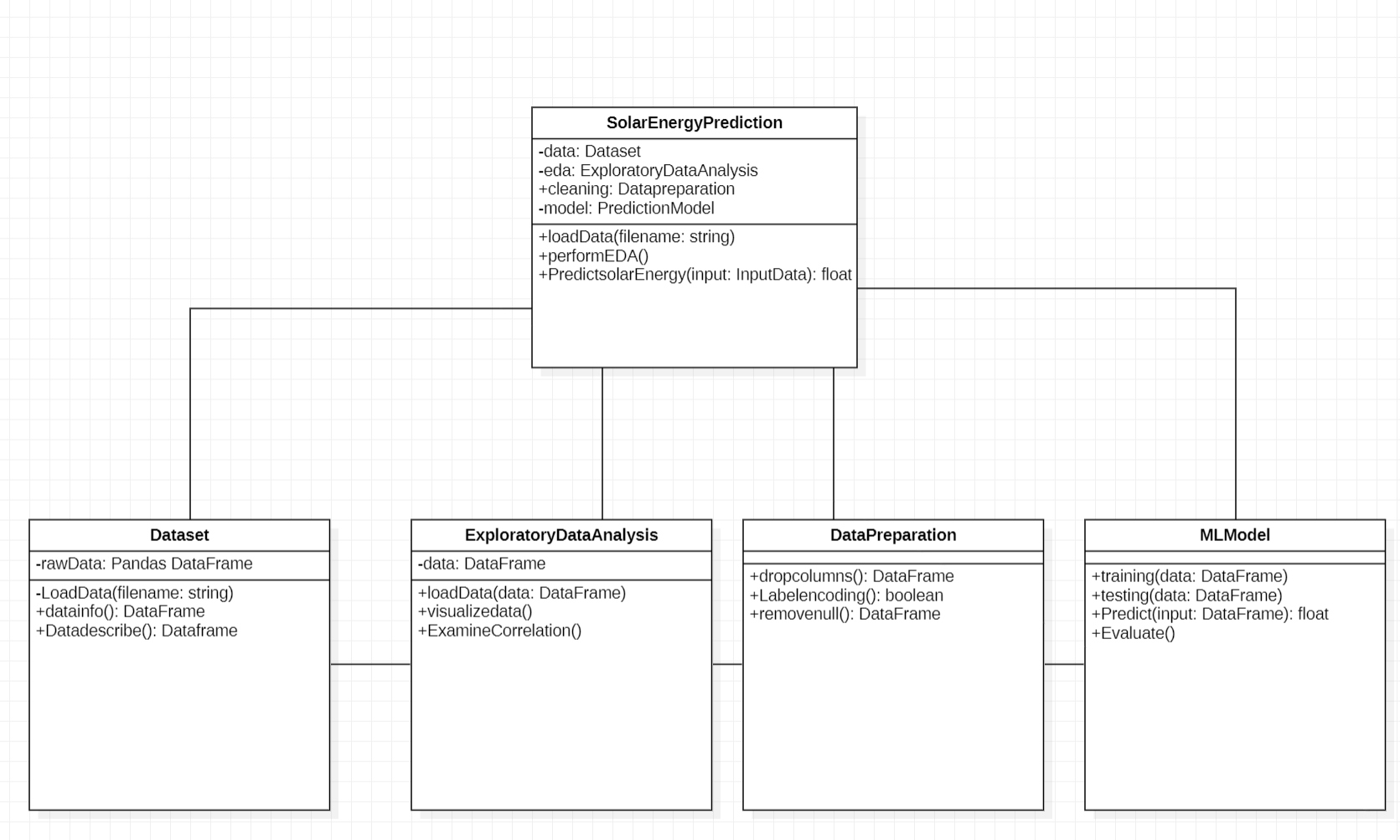
The features in the dataset are: Generated energy, Peak Generation, Exported energy, minimum temperature, maximum temperature and generate/exp ratio.

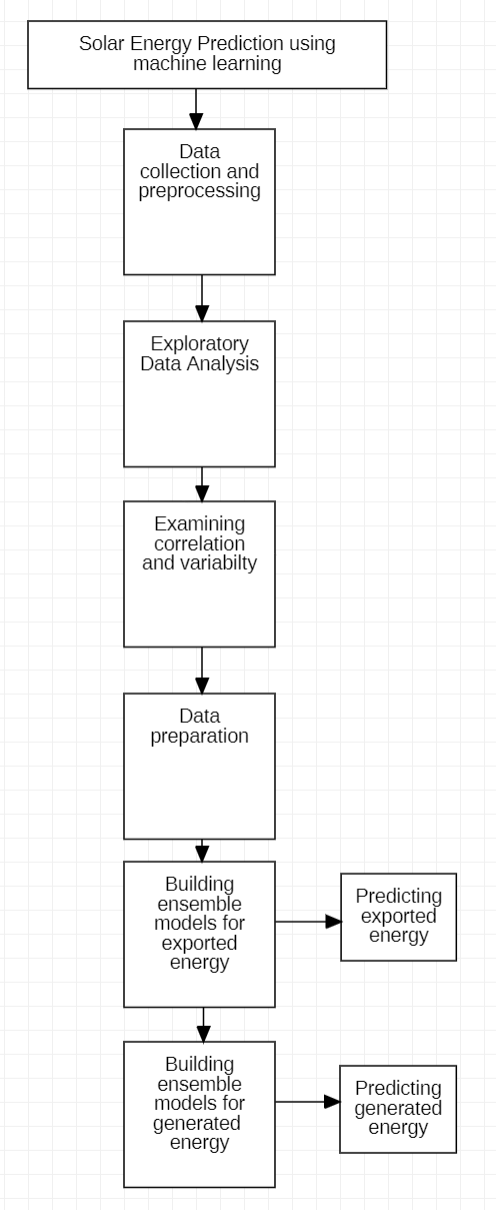
In this project we have conducted the experiment on 6 danny.csv file. This file contains 1280 records with 15 columns. The dataset contains annual and daily statistics.

| **Feature** | **Description** |
| --- | --- |
| Date | Time stamp |
| Mes | Month |
| Ano | Year |
| Gen | Generated solar energy in kWh |
| SubGen | SubGen |
| mmGen | - |
| Exp | Exported solar energy kWh |
| subExp | Sub exported solar energy kWh |
| mmExp | - |
| PP | Peak Power |
| Cond | Weather condition: Fine, partly cloudy, mostly cloudy, cloudy and showers |
| Temp min | Minimum temperature in a day |
| Temp max | Maximum temperature in a day |
| Temp med | Median temperature in a day |
| exp/gen | Derived feature exported vs generated energy ratio |

*table.1.Dataset description*

**Detail design of Features :**

*fig.1.Class diagram*

****

*fig.2.Data Flow Diagram*

**Analysis**:

Because of its environmental and economic benefits, solar energy is increasingly being integrated into smart grids and numerous utilities. However, the uncertainty of accessible solar energy poses issues in terms of power generating reliability and, as a result, consistency in production-consumption balance.In an ensemble, support vector regressors are employed as base predictors, and Multi-layer Perceptrons, Decision Trees, and K-Nearest Neighbour Regressors are utilized as meta-learners to combine [1].

An ensemble learning strategy is proposed to improve the forecasting of solar radiation strength on horizontals. To forecast solar radiation, two types of machine learning models are used: recurrent neural networks and support vector regressors. A multi-layer perceptron model serving as an ensemble learning technique is used to combine the forecasts of ensemble models using a stacking strategy. The combiner performs automatic weighted averaging of the forecasters' results. The proposed method aids in enhancing the accuracy of one-day-ahead solar energy predictions. The performance of the combining technique is tested over the course of a year using meteorological data. The trials demonstrated that the learning-based combinatory model outperformed single models and alternative combining strategies [2].

**Implementation:**

The project implementation is divided into 4 steps:

1.Data collection and preparation: In machine project life cycle data collection and preparation is the first step. After collecting the raw data data is cleaned by removing the null values or missing values. In the data preparation step categorical variables are encoded using Label encoder. Data is split into training and testing.

2. Exploratory Data Analysis:

In this step the distribution of the features is analyzed. Few observations from the EDA are:

* Majority of the samples are Fine weather condition-based samples. Least number of samples are showers data.
* There are no outliers in exported data and outliers are observed in generated energy samples.
* Outliers are present in Peak Power and the variance is very low
* Distribution of generated energy shows that there is dip in the month of June and peaks are observed in the months of March, August, October and December. Similarly, for exported energy also

|  |  |
| --- | --- |
|  |  |
|  |  |

*fig.3.EDA results*

3. Machine learning models:

In this step machine models ensemble regression models are constructed. To predict the results bagging regressor, gradient boosting regressor and stacked regressor are trained. Model performance is evaluated using the test predictions and actual values. For each model r2 score, RMSE (Root Mean Squared Error),MSE(Mean Squared Error), MAE(Mean Absolute Error), MAPE(Mean Absolute Percentage Error) are evaluated.

**Preliminary Results:**

To explore the data nature correlation strength is calculated for features. To perform this task heatmap is constructed using seaborn with correlated features. After the analysis it is observed that there features which are highly correlated. The following pairs have strongly correlated features

Gen - mmGen : 0.83

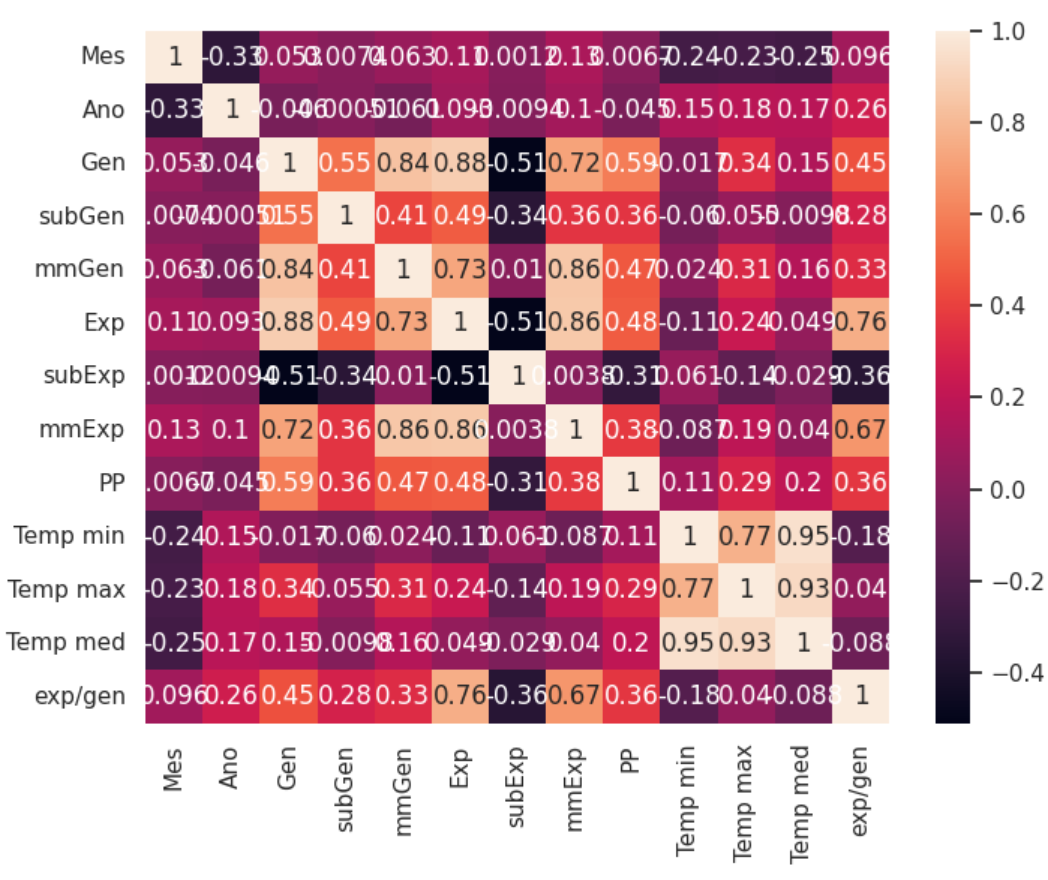
Gen - Exp : 0.88

mmGen - mmExp : 0.85

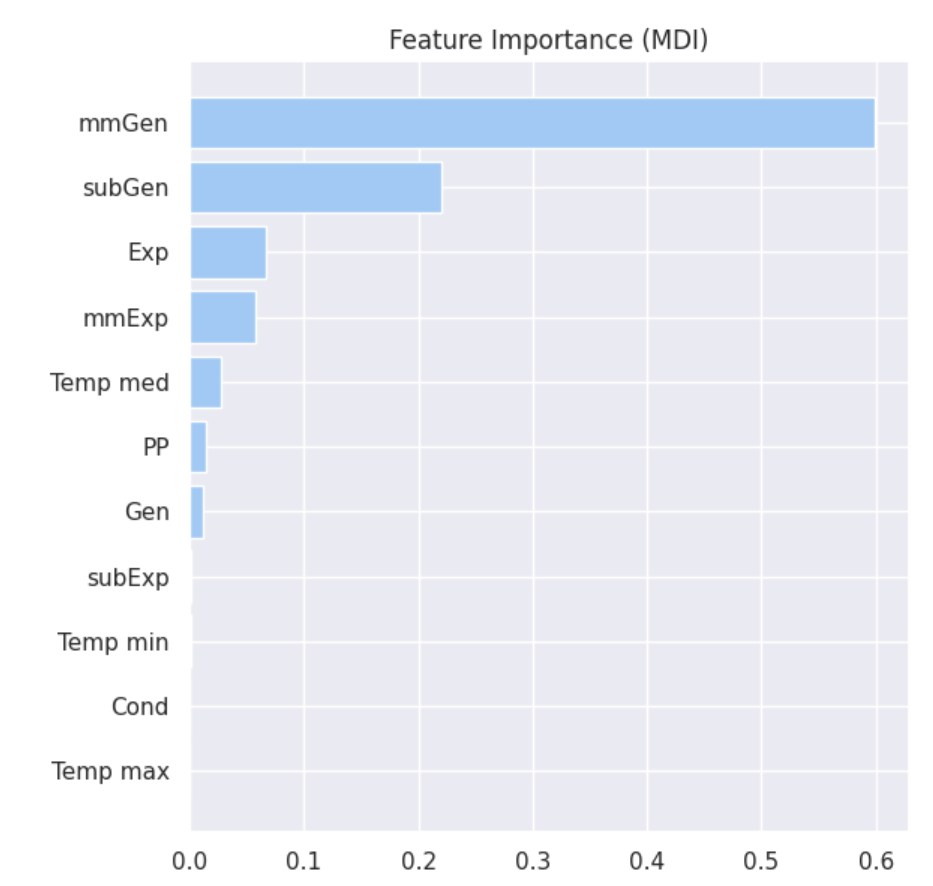
Exp - mmExp : 0.85

Temp min - Temp med : 0.95

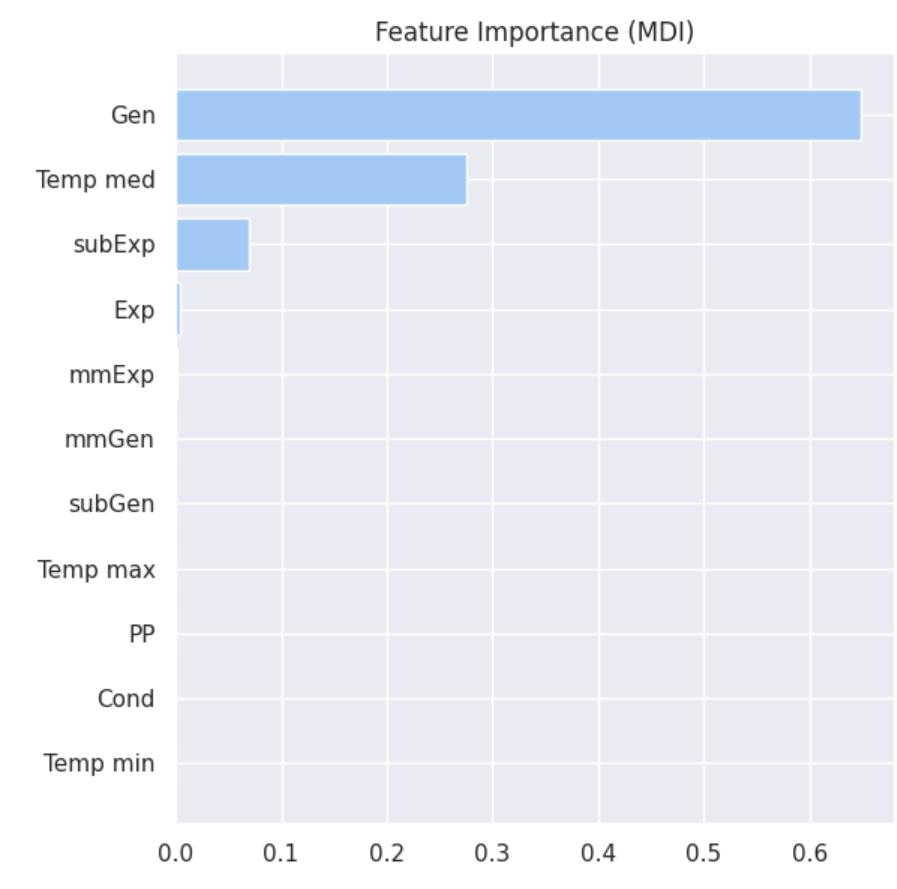
Temp max - Temp med : 0.92



*fig.4.Correlation matrix*



*fig.3.Feature importance score of generated energy*

****

*fig.4.Feature importance score of exported energy*

**Project Management:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Project implementation report | | | | | | |
|  |  |  |  |  |  |  |
| **Project title:** Solar Energy Prediction using machine learning | | | | | | |
|  |  |  |  |  |  |  |
| Status of project activities | | | | | | |
| [S.No](http://s.no/). | Activity | | Description | | Status | |
|
| 1 | Data cleaning and preparation | | Removing the null values and categorical encoding | | 100% | |
| 2 | Exploratory Data Analysis | | Visualizing the data | | 100% | |
| 3 | Examining correlation | | Examining the high correlation of features using heatmap | | 100% | |
| 4 | ML model construction | | Constructing bagging, boosting and stacking regressor models | | 100% | |
| 5 | Training and testing the models | | Training and testing regression models | | 100% | |
| 6 | Comparative analysis | | Comparing the results of each model using evaluation metrics r2,mse etc., | | 100% | |
|  |  |  |  |  |  |  |
| Roles and responsibilities | | | | | | |
| [S.No](http://s.no/). | Activity | | Team member | | Contribution | |
|
| 1 | Data cleaning and preparation | | Chaitanya, Vamsi | | 100% | |
| 2 | Exploratory Data Analysis | | Vamsi, Chaitanya, Shrinivas | | 100% | |
| 3 | Examining correlation | | Vamsi, Chaitanya | | 100% | |
| 4 | ML model construction | | Shrinivas, Chaitanya | | 100% | |
| 5 | Training and testing the models | | Shrinivas, Vamsi | | 100% | |
| 6 | Comparative analysis | | Shrinivas, Chaitanya, Vamsi | | 100% | |

Issues and challenges:

* There are several challenges we faced during the implementation of the project are:
  + Visualizing the model performance, we tried to visualize the best fit using other than scatterplot
  + Few features in the data are anonymous and the units of the features also unknown to better understand the data.

REFERENCES:

[1].R. Al-Hajj, A. Assi and M. M. Fouad, "Forecasting Solar Radiation Strength Using Machine Learning Ensemble," 2018 7th International Conference on Renewable Energy Research and Applications (ICRERA), Paris, France, 2018, pp. 184-188, doi: 10.1109/ICRERA.2018.8567020.

[2].R. Al-Hajj, A. Assi and M. M. Fouad, "Stacking-Based Ensemble of Support Vector Regressors for One-Day Ahead Solar Irradiance Prediction," 2019 8th International Conference on Renewable Energy Research and Applications (ICRERA), Brasov, Romania, 2019, pp. 428-433, doi: 10.1109/ICRERA47325.2019.8996629.