**Project and Data Management (PDM) Plan**

**Project Overview**

**PROJECT TITLE:** SOLAR ENERGY PRODUCTION FORECASTING USING MACHINE LEARNING AND METEOROLOGICAL DATA

**SUMMARY OF THE PROJECT TOPIC AND BACKGROUND:** While solar energy is an important part of the transition to renewable energy sources, it is highly variable based on weather changes during the day. Forecasting Solar Power Production has become a popular source of sustainable energy, and accurately forecasting its production is crucial. This project focuses on developing machine learning models for predicting solar energy output based on historical solar power generation data and meteorological parameters. Using historical trends and weather conditions like solar irradiance, temperature, and humidity, the project aims to produce short-term energy forecasts to help solar farm operators and energy planners.

**RESEARCH QUESTION:** How accurately can machine learning models predict solar energy production based on historical meteorological and solar power generation data?

**PROJECT OBJECTIVES:**

- **Data Collection & Preprocessing:** Obtain and merge solar power generation and weather data from credible sources (as loaded in the notebook), clean and preprocess the data (handling missing values, formatting datetime, scaling features) for further analysis.
- **EDA:** Explore trends, correlation, and patterns between meteorological variables and solar energy output using visualizations and descriptive statistics.
- **Feature Engineering:** Extract and select relevant features impacting solar power production, potentially including solar irradiance, temperature, humidity, and wind speed, as used in the model inputs.
- **Model Implementation & Comparison:** Implement and train multiple machine-learning models, specifically **LSTM** and **Random Forest (Regressor and Classifier)**, to predict solar energy output.
- **Model Validation & Optimization:** Assess model performance using metrics such as RMSE, MAE, and $R^2$ (as calculated in the notebook), and potentially optimize hyperparameters to improve accuracy.
- **Deploy & Interpret:** Appropriately deploy the best model (though deployment is not shown in the notebook), interpret results, and deliver insights into how to use these predictions for real energy management in the field.

**REFERENCE LIST:**

1. Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de-Pison, F. and Antonanzas-Torres, F., 2016. Review of photovoltaic power forecasting. *Solar Energy*, 136, pp. 78-111. Available at: https://doi.org/10.1016/j.solener.2016.06.069 [Accessed 10 February 2025].

2. Voyant, C., Notton, G., Kalogirou, S., Nivet, M.L., Paoli, C., Motte, F. and Fouilloy, A., 2017. Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, 105, pp. 569-582. Available at: https://doi.org/10.1016/j.renene.2016.12.095 [Accessed 10 February 2025].

3. Wan, C., Zhao, J., Song, Y., Xu, Z., Lin, J. and Hu, Z., 2015. Photovoltaic and solar power forecasting using machine learning: A state-of-the-art review. *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 3(4), pp. 1474-1486. Available at: https://doi.org/10.1109/JESTPE.2015.2427378 [Accessed 10 February 2025].

**Project Plan: Task List and Timeline**

| Week | Task | Description |
|---|---|---|
| **1-2** | Project Proposal & Literature Review | Define project scope, research relevant literature, and finalize research question. |
| **2-3** | Data Collection & PDM Plan Submission | Gather solar power generation and meteorological datasets from sources such as NREL, Kaggle, and Open-Meteo API. |
| **3-4** | Data Cleaning & Preprocessing | Handle missing values, format data correctly, and normalize variables for analysis (as performed in the notebook). |
| **4-5** | Exploratory Data Analysis (EDA) | Visualize and analyze relationships between meteorological factors and solar power output (as performed in the notebook). |
| **5-6** | Feature Engineering | Select key features (solar irradiance, temperature, humidity, wind speed) that impact energy production (as used in the notebook models). |
| **6-7** | Model Development | Train and test various machine learning models (**LSTM, Random Forest Regressor, Random Forest Classifier**) for prediction. |
| **7-8** | Model Evaluation & Optimization | Use RMSE, MAE, and R² metrics to validate models and fine-tune hyperparameters (metrics calculated in the notebook). |
| **8-9** | Final Model Selection & Report Draft | Choose the best-performing model, analyze results, and prepare deployment strategy. Write methodology, findings, and conclusions. |
| **9-10** | Report Finalization & Presentation Preparation | Refine final report, create PowerPoint slides, and practice for presentation. |

**Data Management Plan**

**Overview of the Dataset:**

- **Dataset Source:** Kaggle (https://www.kaggle.com/datasets/anikannal/solar-power-generation-data)
- **Geographical Coverage:** Includes solar farms from multiple regions.
- **File Format:** CSV
- **Number of Records:** ~300,000 + data points
- **Expected File Size:** ~500MB-1GB after preprocessing
- **Key Variables:**
  - DATE_TIME: Timestamp of energy production
  - SOLAR_IRRADIANCE (W/m2): Amount of solar energy received
  - TEMPERATURE (∘C): Ambient temperature
  - HUMIDITY (%): Moisture content in the air
  - WIND SPEED (m/s): Wind impact on solar efficiency
  - ENERGY_OUTPUT (kW): Power generated

**Data Collection & Processing Tools:**

- **Primary Source:** Kaggle dataset.
- **Data Retrieval Format:** CSV files from Kaggle.
- **Tools Used (based on notebook):** Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, TensorFlow/Keras.
- **Processing Steps (based on notebook):** Handling missing values (dropna), datetime conversion, data scaling (StandardScaler, MinMaxScaler), sequence creation for LSTM.

**Version Control:**

- **Primary Storage:** GitHub repository for code (as mentioned in your previous plan).
- **Commit Frequency:** Weekly commits and updates are recommended.

**Storage & Security:**

- **Backup Storage:** OneDrive or Google Drive for dataset backups (as mentioned in your previous plan).
- **Backup Frequency:** Weekly updates and data storage backups are recommended.
- **Access:** Restricted to supervisor and student.

**Ethical Requirements:**

- **GDPR Compliance:** The dataset does not include any personal or sensitive data; it is all publicly available data.

- **UH Ethical Policies:** Project adheres to University of Hertfordshire ethical research policies.
- **Data Use Permission:** Open-Source Dataset for academic research.
- **Ethical Data Collection:** All data was initially sourced from reputable agencies and collected under regular research guidelines.