Tab 1

**Project plan**

Project name: Solar plant data analysis

Project members: Lauri Lopp, **Gustav Nikopensius**, Martin Leissoo

Problem statement:

Evaluating the operation of solar energy equipment on-site is a time-consuming and complex task due to the diverse range of devices spread across large areas. That's why analyzing the incoming energy from solar panels is essential for identifying performance issues. However, it is equally important to consider the amount of solar radiation available at any given time. When sunlight is insufficient, panels may produce minimal energy, making it difficult to determine whether any observed performance issues stem from electronic malfunctions or simply from a lack of sunlight. Therefore, an integrated analysis of solar power output and solar radiation data is crucial for effective error detection and optimizing solar energy system performance.

Objectives:

**Fault Detection**: Analyze the solar energy equipment data to identify and diagnose faulty components or systems. This involves detecting patterns of outliers and missing values in the dataset and identifying the associated source devices.

**Maintenance Insights**: Determine which solar panels may require cleaning based on their performance metrics and solar radiation data. Since the data lacks labels for panel cleanliness, our goal is to identify underperforming devices relative to the entire fleet. These findings will serve as cues for further investigation in a business context.

**Energy Generation Prediction**: Utilize weather data to develop predictive models for energy generation, facilitating grid optimization and improving overall efficiency. Given the small size of the dataset, we would consider a Mean Absolute Error (MAE) of 20% or less as a resounding success.

Data: We have a Kaggle dataset.

1. Form a preliminary understanding of the data
   1. Read documentation
   2. Examine and visualize datasets
2. Identify outliers and missing values
   1. Identify patterns for **Fault Detection**
3. Cleanse data
   1. Imputation and filtering
4. Engineer new features
5. Visualize findings
   1. Communicate **Maintenance Insights**
6. Prepare dataset for prediction task
   1. Propose models
   2. Go through the ML processing pipeline
   3. Train and evaluate
7. Compile a report
   1. Fault detection report
      1. Describe patterns in data: Explain detected patterns such as outliers, inconsistencies, or performance drops.
      2. Explain implications: Discuss how these patterns indicate faults.
      3. Point out faulty devices: List the devices identified as potential sources of failure, based on the data analysis.
   2. Maintenance Insights
      1. Visualize device performance: Use graphs or visual aids to show which devices are underperforming relative to the fleet.
      2. Propose a basis for maintenance decisions: Suggest performance thresholds on key metrics that should trigger further investigation or maintenance.
   3. ML model proposition
      1. Performance evaluation: Present the results of the machine learning model, discussing how accurate the predictions were.
      2. Suitability analysis: Analyze the practicality of implementing the model in a real-world setting for grid optimization.

Methodology:

To **predict energy generation**, we will compare:

* Linear Regression,
* a Gradient Boosting Machine (XGBoost),
* an LSTM network.

To gather **maintenance insights**, we will compare:

* K-means clustering,
* Hierarchical clustering.

Evaluation:

For the predictor we will use Mean Absolute Error as it provides error in the same units as our predicted variable, which makes it intuitive and easy to interpret.

To evaluate the clustering algorithms for maintenance insights, we will use Silhouette Score.

Expected challenges:

* *Data quality issues*

***Challenge****: The dataset may contain missing or erroneous values, especially in fields related to weather and solar panel outputs. This could hinder accurate fault detection and energy prediction.*

* *Distinguishing faults from normal variability*

***Challenge****: Solar panel performance is heavily influenced by environmental factors like weather conditions and time of day. Poor performance due to cloudy weather might be misinterpreted as equipment failure.*

* *Overfitting or not having good enough generalizability due to a small dataset*

Resources and tools:

Pandas for data transformations,

Numpy for vector calculations,

Sklearn for ML models and preprocessing (except for deep learning),

TensorFlow/Keras for deep learning,

Matplotlib/seaborn for visualizations,

Google Colab environment.

Questions for further guidance:

Is MAE a good evaluation metric for the power generation model? Should we consider other metrics, like RMSE, for example?

Milestones and timeline:

| **WEEK** | **DATE** | **MILESTONES OR PLANNED ACTIVITIES** |
| --- | --- | --- |
| Week 4 | September 30 |  |
| Week 5 | October 7 *(teams formed)* | **Continuous reporting throughout project** |
| Week 6 | October 14 | Preliminary data understanding |
| Week 7 | October 21  *(project plan deadline)* | Project plan submitted, faults detected |
| Week 8 | October 28 | Data analyzed and cleansed |
| Week 9 | November 4 | Feature engineering complete |
| Week 10 | November 11 | Insightful visualizations, preliminary report |
| Week 11 | November 18 - 20  *(intermediate presentations)* | Presentation complete |
| Week 12 | November 25 | Machine learning environment setup |
| Week 13 | December 2 | Machine learning done, evaluated |
| Week 14 | December 9 | Compile report and presentation |
| Week 15 | December 16 - 18  *(final presentations and team evaluation)* | Present |