**Problem statement**

The scope of this project is to conduct a detailed evaluation of the predictive value and business value contained in the power generation data from photovoltaic panels. At the end of the project the students should be able to provide insight and suggestions of possible warnings that can be implemented in the power plants in order to ensure early correction of potential inefficiencies and reduce the losses due to these inefficiencies.

The conclusion will take into consideration not only the evaluation of technical feasibility, but also detailed considerations about the practical and business aspects (time-scales, practical feasibility, costs, etc. ) that need to be considered in order to judge about the actual utility of the solution.

**Data**

The data for this project is taken from open-source datasets available on Kaggle (not the competitions). The scope of the project, however, is unique and ensures that the students work does not intertwine with what is available on the website. In fact, the project is a feasibility study and the choice of the data used for this type of a project is not as crucial as the approach that will be used by the students for the argument and use case construction.

**Work plan**

**Stage 1**: Basic data evaluation and exploration

* Technical overview and understanding of the subject matter
* Exploration and evaluation of the available data and information
* Preliminary exploration of the inverter behaviour within plants
  + Study the power generation behaviour of single inverter and its evolution/changes in time
  + Study the power generation behaviour of a single inverter relative to weather conditions or relative to the conditions of the inverter itself
  + Study and compare the behaviour between different inverters in the same plant

Main notions used at this stage:

Data visualisation

Simple statistical calculations

Exploration of correlations and variable dependences

**Stage 2**: Identification of failures/inefficient behaviour/degradation in the data using visualisation or simple statistical tools

* Exploration of failure signs and their interpretation using the information available in the data (if possible)
* Exploration of signs of inefficient behaviour of inverters relative to the given weather or inverter conditions
* Exploration of signs of long-term degradation among inverters

Main notions used at this stage:

Data visualisation

Simple statistical calculations

Exploration of correlations and variable dependencies

**Mid-project report**: At this stage the student/the team can prepare a presentation summarising their findings so far and providing suggestions of the most promising future directions to take. For example, outlining the type of anomalies that seem to be most promising to detect using automised techniques.

The students will also be able to revisit their plan of work at this stage. Based on the remaining time and their preferences a range of directions from the Stages described below can be selected for deep exploration and implementation.

**Stage 3:** Exploration of possible methodologies for automatic detection of selected problematics. Below is a selection of possible ways, but the exact activities will depend on the results of the previous stages

* Detection of failures
  + e.g. signal processing methods to detect abrupt changes in time series
* Detection of short-term inverter inefficiencies
  + Algorithms based on statistical KPIs to compare inverter behaviour and detect anomalies
  + Algorithms based on curve similarity measures
  + ML algorithms such as regression to estimate the expected behaviour and compare it to real
  + Autoencoders
* Detection of inverter/cell long-term degradation. In this case the algorithm will have to detect signs of the inverter being consistently performing worse than expected. Thus, a variety of the methods mentioned in the previous point are applicable and will need to be adapted to ensure the long-term character of the detected behaviour.

Main notions used at this stage:

Signal processing techniques

Statistical description of data

Curve similarity

Anomaly detection

Regression

Autoencoders

**Stage 4:** Evaluation

* Conclusion on the feasibility of accurate detection of the anomalies explored in the previous stage. Evaluation of the performance of the algorithms.
* Evaluation of the practical applicability of the algorithms in a real setting (here practical considerations will be done, such as 1. Are we sure the behaviour studied in the work and the conclusions made are applicable on a larger set of data and in more general settings? 2. Are we sure the resolution of the selected inefficiencies is at all possible and the time-scales of for doing this are compatible with detection time-scales or acceptable at all?, Other questions allowing to determine the practical utility of the result)
* Evaluation of business value – estimation of costs, potential earnings due to inefficiency improvements, etc.

**End of project report**: This report will contain an overview of the insights gained in the first stage and the details on the evaluation results of the stages 3 and 4. It will give the stakeholder a final conclusion on which seem to be the most promising (or the least promising) directions in practical and financial senses for the implementation of a final product.

**Deliverables**

1. Mid project report/presentation – designed for a non-technical minded user
2. End project report/presentation – Final evaluation and conclusions of the DS experts designed for the non-technical minded user (in case of limited time to prepare material this report could be merged into the project report)
3. Project report for the university – detailed technical description of the work.

Approximate structure of the project report

* 1. Introduction to the subject matter
  2. Introduction to the mathematical concepts used during the project
  3. Description of the methodology (particularities of data processing, types of anomalies explored, details on the models used, the considerations made for the final evaluation)
  4. Results

1. Code repository to be evaluated for code writing skills. The repository will contain the organised code that was used for exploration and any final algorithm implementations.
2. Other deliverables that could be required by the faculty.

**Notions that could be used and practised by the students during the project**

* Data visualisation
* Extensive use of statistical concepts and development of data intuition
* Signal processing techniques
* Curve similarity methods
* Regression algorithms, autoencoders
* Trend analysis
* Experience in software writing and version control for data science projects
* Experience considering the business aspects of applications
* Experience of preparing presentations and communicating with non-technical minded stakeholders

**Tools used**

* Python/R and the respective data analysis and data visualisation packages (e.g. pandas, matplotlib, seaborn, etc for Python). The choice of the language to work with is up to the student or the team to select.
* Git (or any other version control system selected by the students or the team)
* (Optional) PowerBI. Depending on the available time, the data and the resulting algorithms can be packaged into a dashboard in order to provide a user-friendly application interface