MACHINE LEARNING PROJECT SMARTIVE CASE GUIDELINES

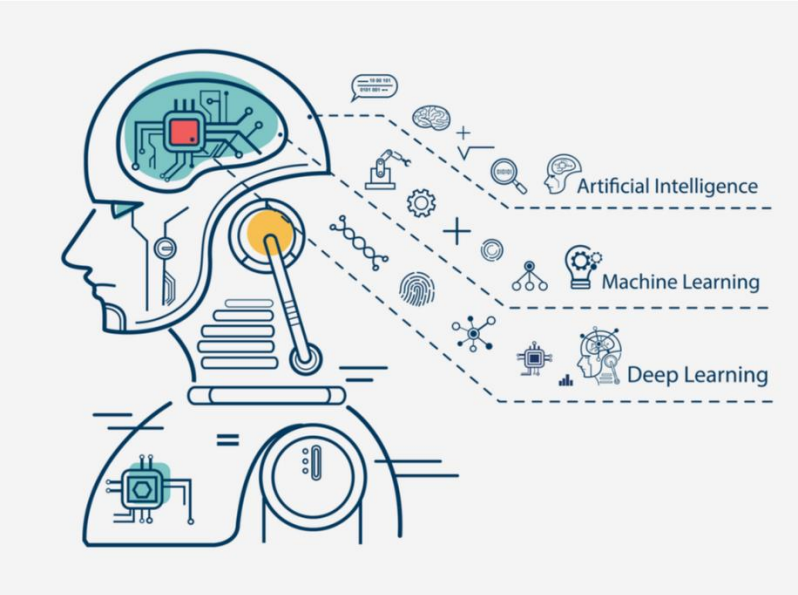

Image from : https://actualiteinformatique.fr/intelligence-artificielle/definition-deep-learning

Responsible: Khaoula TIDRIRI
Associate Professor at Grenoble INP (Ense3)
Researcher at GIPSA Lab (SAFE Team)

E-mail : khaoula.tidriri@grenoble-inp.fr

Machine Learning: Project Guidelines

"This project was initially developed by Gianluca Mauro at AI Academy with support from InnoEnergy."

In this project you'll play the role of a data scientist working for an energy company: Smartive. Smartive is a Spanish company that builds advanced AI algorithms to improve the efficiency of wind turbines. The core product of Smartive is a technology capable of **predicting with a few weeks in advance when a wind turbine is about to fail.** This piece of information can allow Smartive's customers to take preventive actions that can maximize uptime and minimize costs. This case describes the challenges that the company went through in commercializing a technology that in its infancy stage, and in defining a successful business model. It also offers a view onto the strong ties between the technical performance of the company's algorithms with the economic return on investment for Smartive's clients, and its repercussions for the company's business model.

You are asked by your managers to answer some questions that have direct business impact, and you'll need to do so by looking through the data and producing insightful metrics and graphics.

Learning outcomes:

1. Implement the different elements of data-preprocessing
2. Select and implement Machine Learning algorithms to answer real-world business problem
3. Determine which model is the most suitable for a regression task
4. Estimate how the model will perform.
5. Present a well-reasoned solution to the business problem for which you have analyzed energy datasets and developed ML algorithms

Available Material:

- Machine Learning Project.pdf
- Smartive Case.zip

I. Session 1: Business understanding and data cleaning

- Objectives: Understand the business problem
  Debate with your group
  Get the data
  Clean the data

- Material: Smartive_Case.pdf
  Jupyter Notebook: data_cleaning.ipynb
  Data EDPR folder

1) Understand the business problem

Smartive is a Spanish company that builds advanced AI algorithms to improve the efficiency of wind turbines. The company started from the research that its founder Jordi Cusido has focused his PhD on.

The core product of Smartive is a technology capable of predicting with a few weeks in advance when a wind turbine is about to fail. This piece of information can allow Smartive's customers to take preventive actions that can maximize uptime and minimize costs.

A clear and compelling value proposition hasn't been enough to spare the company from the challenges of selling their technology to customers. Convincing traditional organizations such as utilities to use a fairly new technology that has just left the lab isn't an easy task. Jordi had to take advantage of external funding, investment funds and creative business models. In its six years of existence, Jordi managed to set his company on a sustainable track that allows him to grow steadily, establishing itself as a key player in the European market.

This case describes the challenges that the company went through in commercializing a technology that in its infancy stage, and in defining a successful business model. It also offers a view onto the strong ties between the technical performance of the company's algorithms with the economic return on investment for Smartive's clients, and its repercussions for the company's business model.

- Read the document: **Smartive_Case.pdf** to understand the company's story, its ML business model and its technology.

2) Debate with your group: assignment information

Try to answer these questions by discussing them with your colleagues:

1. What defines a good use-case for ML?
2. What are the factors that can influence Smartive's business success?
3. Do you need to own data to start an ML company?
4. How can you build customers' trust for ML algorithms?
5. What are the challenges in pricing an ML solution?

3) Get the data

In the **Data-EDPR** folder, you'll find different datasets coming from EDPR's wind turbines. These are the different files:

- data_wind_prod.csv: wind turbine data raw
- wind-farm-1-metmast-2016.csv: meteorological data raw
- wind-farm-1-failures-training.csv: turbines' failures

You will also find some documentation about the data:

- Wind Farm - Metmast Variables.pdf
- Wind Farm - Signals Variables.pdf

4) Clean the data

Jupyter Notebook: data_cleaning.ipynb

II. Session 2-3: Wind descriptive analytics

- Objectives: Compare the energy produced by four different turbines
  Manage demand side
  Fault analysis

- Material: Session 2-3_ Descriptive Analytics.ipynb
  Data clean folder

In the **data_clean** folder, you'll find two cleaned datasets that you are required to use:

- turbines_df.p: wind turbine data cleaned - pandas multiindex DataFrame
- mast.p: meteorological data cleaned

1. Turbines' comparison

The dataset has data from four different turbines. They are all the same model and in the same location, so they should all produce the same energy. Are we sure though? Here's your first task: compare the energy produced by the four different turbines and find out whether there's any turbine that is not giving the same results as the others.
Hint: comparing the turbines along the full year may not be enough: check their performances in different months.

2. Demand side management

One of the challenges the energy system is facing is demand side management: it's increasingly difficult to forecast how much energy needs to be produced by traditional thermal plants to match demand. Renewables are making the problem harder. Before the rise of renewable resources, an energy company had to forecast demand and match that.

Demand = Fossil fuel production

Now, renewable resources are adding another source of volatility to the equation, and energy companies also need to forecast how much energy will be added to the grid by these new sources.

Demand = Fossil fuel production + Renewable resources production

Your managers are asking you to check the impact of the four wind turbines at different times of the day, so that they can foresee the impact on demand management of investing in more turbines. For this task, you need to look at a typical day and check the energy produced by the four wind turbines at different times of the day.
Hint: look at differences across different months.

3. Fault analysis

In the wind-farm-1-failures-training.csv dataset you can find different failures of the four turbines. Your managers want to know whether the faults had any impact on the data collected by the turbines' sensors. In this exercise, you need to produce some graphs that show what happened to key parameters of the turbines before the faults occurred, and try to spot any irregularities.
Hint: you may have to compare broken turbines with the others that are regularly functioning.

III. Session 4: Wind Predictive analytics

- Objectives: Model the power output of a turbine
- Material: Session 4_ Prescriptive Analytics.ipynb

The objective of this part is to model the power output of a turbine as a function of the wind speed. This is a very important problem, because it has a direct impact on the financial feasibility of the wind farm. As a result, it's been widely studied in the literature. Betz's theory establishes an upper limit to the efficiency of a wind turbine, based purely on the conservation of kinetic energy within the air stream. More sophisticated models also account for aerodynamic effects on the blade and conversion efficiency.
While these theoretical approaches offer a solid ground for the design of a wind farm, nothing beats the actual data that's collected from the operating turbines. In this exercise, we're going to try to model the performance of the turbines from collected data. Throughout the exercise, you'll learn about recognizing the negative impact of outliers, and different types of regression models.

IV. Session 5: Present your solution

Now, it is time for you to present your solution. You are also required to provide a **2-page report**, highlighting your main findings!