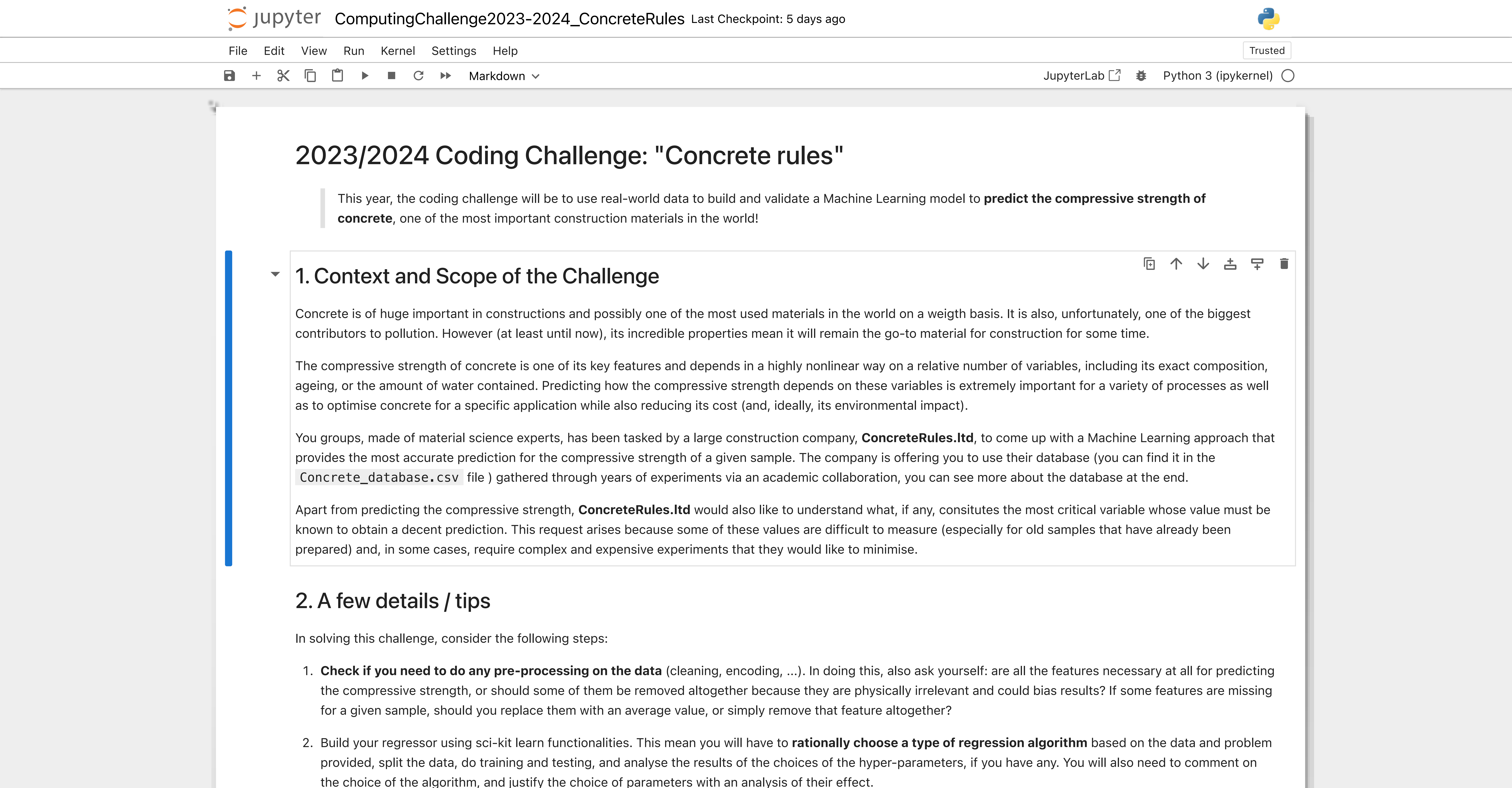

2023/2024 Coding Challenge: "Concrete rules"

> This year, the coding challenge will be to use real-world data to build and validate a Machine Learning model to **predict the compressive strength of concrete**, one of the most important construction materials in the world!

1. Context and Scope of the Challenge

Concrete is of huge important in constructions and possibly one of the most used materials in the world on a weigth basis. It is also, unfortunately, one of the biggest contributors to pollution. However (at least until now), its incredible properties mean it will remain the go-to material for construction for some time.

The compressive strength of concrete is one of its key features and depends in a highly nonlinear way on a relative number of variables, including its exact composition, ageing, or the amount of water contained. Predicting how the compressive strength depends on these variables is extremely important for a variety of processes as well as to optimise concrete for a specific application while also reducing its cost (and, ideally, its environmental impact).

You groups, made of material science experts, has been tasked by a large construction company, **ConcreteRules.ltd**, to come up with a Machine Learning approach that provides the most accurate prediction for the compressive strength of a given sample. The company is offering you to use their database (you can find it in the `Concrete_database.csv` file) gathered through years of experiments via an academic collaboration, you can see more about the database at the end.

Apart from predicting the compressive strength, **ConcreteRules.ltd** would also like to understand what, if any, consitutes the most critical variable whose value must be known to obtain a decent prediction. This request arises because some of these values are difficult to measure (especially for old samples that have already been prepared) and, in some cases, require complex and expensive experiments that they would like to minimise.

2. A few details / tips

In solving this challenge, consider the following steps:

1. **Check if you need to do any pre-processing on the data** (cleaning, encoding, ...). In doing this, also ask yourself: are all the features necessary at all for predicting the compressive strength, or should some of them be removed altogether because they are physically irrelevant and could bias results? If some features are missing for a given sample, should you replace them with an average value, or simply remove that feature altogether?
2. Build your regressor using sci-kit learn functionalities. This mean you will have to **rationally choose a type of regression algorithm** based on the data and problem provided, split the data, do training and testing, and analyse the results of the choices of the hyper-parameters, if you have any. You will also need to comment on the choice of the algorithm, and justify the choice of parameters with an analysis of their effect.