Project - Machine Learning

Will it rain tomorrow?

Main Objective

The main objective of the current project is to answer the question:

Will it rain tomorrow?

Given some information about the weather, try to provide an answer for that question.

Specific Objectives

The idea is for you to train (got it?) how the process of a machine learning model is structured and get a sense for what machine learning models can do for us.

Specific Objectives

1. You are required to compare at least 3 machine learning methods. Try to understand their differences.

2. Remember you are required to specifically answer whether it is going to rain tomorrow or not (yes or no). You can simulate a "tomorrow" in the dataset.

Deliverables

1. Who's your customer? Describe your customers.

- You are required to provide a discussion about how costly your errors are for two customers:

- The first one is a person who wakes up in the morning and requires your prediction to go to work. What types of error impacts this customer the most? Which metric should you optimize?

- The second one, you'll have to create. Imagine another customer and specify which errors should be optimized for this case.

Deliverables

2. Provide an estimate for your error. Which metrics did you choose? How did you estimate it? Do you feel comfortable with your results? Can you translate it for a more business-like language (other than precision, recall, accuracy)?

Extra hints and ideas

1. You can use clustering methods to group similar observations together. You can use dimensionality reduction algorithms to visualize them.

2. Take care with the way you estimate your error. Remember this is a time-dependent problem

3. Take care with the way you input missing values. Again, remember which variables you'll have if you are predicting tomorrow's weather

Extra hints and ideas

4. There are some weird variables in the dataset. Don't trust 100% on the documentation. Try to find at least <u>two</u> of them and tell how did you find it. There are many ways to find it, some are more manual and some are more automatic

5. Try to figure out which variables are the most important and which are the least. Interpret your results.

6. There's a variable called 'modelo_vigente'. Think of what you could use that for. Could you use it as a predictive variable? Discuss the pros and cons for that.