Data Science Skills Evaluation Project

Please read the instructions carefully.

The dataset used for this challenge is a modified version of the *"Pump it Up: Data Mining the Water Table"* dataset. Although there may be some online articles describing this dataset and how to analyze it – it will not help with this challenge due to changes made to the dataset for the purposes of this evaluation, and specific tasks we are giving.

Note that these changes are relatively minor in nature – for example individual rows and fields have been corrupted with invalid information. The purpose of making these changes is to evaluate your ability to clean the data. Cleaning involves identifying data that is incorrect (or missing) and determining how best to handle it.

For each of the below tasks (1-8) it is expected the applicant will produce documentation within a working Jupyter notebook demonstrating their understanding of steps taken. It is important that the code is documented so that we can understand *why* you did what you did. One or two sentences per block of code is sufficient (i.e. you shouldn't need to document each individual line of code).

It will be important that the applicant can answer *contextual* questions about the dataset – e.g. understanding why predicting status_group is important.

Tasks:

1. Carry out exploratory analysis on the dataset. Create an appropriate visualization of each feature.
2. The file (data.csv) contains a target value which we will use to build a classification model. Join the two files using the ID column. Be prepared to describe any pitfalls in joining.
3. Determine whether any missing or corrupt values need to be handled. Be prepared to discuss how you identified missing or corrupt values and how you handled them.
4. Determine whether any of the numeric features are correlated. What impact might this have for modeling and how can it be mitigated?
5. Some of the variables are categoric. Come up with a scheme for handling categorical variables
6. Given the above steps – prepare a final dataset for predicting the status_group target
7. Carry out a machine learning exercise to predict status_group. Be prepared to discuss the following:
   - What is the class balance?
   - What ML algorithm(s) was/were chosen and why?
   - How was the data split for training / testing /validation?
   - What was the performance on validation data? And performance on cross-validation data – was there a significant difference between these two?
   - What were the important features?
   - How did you determine the accuracy of the model?
   - How did you compare models?
8. What else what you have done with this dataset given more time?

Data Dictionary:

- amount_tsh - Total static head (amount water available to waterpoint)
- date_recorded - The date the row was entered
- funder - Who funded the well
- gps_height - Altitude of the well
- installer - Organization that installed the well
- longitude - GPS coordinate
- latitude - GPS coordinate
- wpt_name - Name of the waterpoint if there is one
- num_private -
- basin - Geographic water basin
- subvillage - Geographic location
- region - Geographic location
- region_code - Geographic location (coded)
- district_code - Geographic location (coded)
- lga - Geographic location
- ward - Geographic location
- population - Population around the well
- public_meeting - True/False
- recorded_by - Group entering this row of data
- scheme_management - Who operates the waterpoint
- scheme_name - Who operates the waterpoint
- permit - If the waterpoint is permitted
- construction_year - Year the waterpoint was constructed
- extraction_type - The kind of extraction the waterpoint uses
- extraction_type_group - The kind of extraction the waterpoint uses
- extraction_type_class - The kind of extraction the waterpoint uses
- management - How the waterpoint is managed
- management_group - How the waterpoint is managed
- payment - What the water costs
- payment_type - What the water costs
- water_quality - The quality of the water
- quality_group - The quality of the water
- quantity - The quantity of water
- quantity_group - The quantity of water
- source - The source of the water
- source_type - The source of the water
- source_class - The source of the water
- waterpoint_type - The kind of waterpoint
- waterpoint_type_group - The kind of waterpoint