Wildfire project - group 8.

Here we will explain what files we have submitted, and how to use our model in order to predict the test set.

<u>In this Google Drive folder we have the following files and directories:</u>

1. datasets.zip - a compress folder which includes all the dataset we use:

It includes 9 sub-directories, and one file named preprocessed_final.csv - which is
a file that includes all the data that our model will be trained on. This file is created by
the function 'preprocess' in the file project_part2.ipynb, more on that later.

2. project_part1.ipynb - includes all code relevant to data cleaning and exploration.
3. project_part2.ipynb - includes all code relevant to feature engineering.
4. project_part3.ipynb - includes all code relevant to model finetuning and model selection
5. project_part4.ipynb - includes all code relevant to features importance and outlier detection.
6. features_extractor - includes code for extracting temperature feature from the internet (part of our pre-processing on the given data)
7. for_test_set.ipynb - the file to run on test set. instructions for using it will come below, and also in the start of the file.
8. encoder.p - a pickle file for a trained sklearn one hot encoder (used in pre-processing of the test set)
9. Requirements.txt - all packages needed for all our code to run properly (please note that for the code the graders will use, not all those requirements are necessary).
10. README - this file

<u>How to use the model to predict the train set:</u>

1. Download all contents from the drive to a single folder: https://drive.google.com/drive/folders/1MHNCjcWI2vtClGCzg74K4ZLyEiDX9c7O?usp=sharing
2. Unzip the datasets.zip folder into the same folder you downloaded the content to.
3. Install all the packages needed as in the requirements.txt file
4. Open the file for_test_set.ipynb
5. On the second cell of the notebook, load the data you want to predict
6. Run the entire notebook.
7. Classification report will be printed at the end of the notebook
8. **Important Note: This notebook will train and predict using two different models. The first model is a Random Forest Classifier, which will finish the prediction relatively quickly.**
   The second model a custom made Confusion Matrix Mixture Of Experts model that performs better, but takes a while to predict. For us, it took 6 hours on 600,000 samples. We don't know if this is a legitimate time, so we gave a faster option as well.