The Forest Buddy: Summary

Project Design

The project aim is two-fold:

- Build a predictor for the areas in mountain forests with a high risk for the **occurrence of wildfires**.
    - Binary classification: high risk (10 % of the observations) and low risk. The efforts to make a model for this case can be found in ***/FIRE/fire_high.*** The promising results obtained results motivated a multiclass classification.
    - Multiclass classification (***/FIRE/fire_multi***): high risk (10 % of obs.) moderate risk (40 % of obs.) and low risk were considered. Several methods were used and hyperparameter tunning was carried out for the ones giving the best results.
- Build a predictor for the **vegetation coverage** that would will be found in a given area:
    - Two preliminary binary classification for the most common (class 2) and the rarest (class 4) cover types were explored. Used notebooks can be found in ***/TREES/common*** and ***/TREES/rare.***
    - Multiclass classification: seven cover types were considered.

Two notebooks (Forest_clean_data.ipynb and Fire_clean_data.ipynb) were used for data preprocessing and the generation of the data necessary for the different problems.

Tools

The **recall and precision** for the high risk class were used as the deciding indicators in the wildfire risk predictor. High recall minimize the risk of missing high risk areas and high precision avoids the occurrence of false positives that would lead to unnecessary waste of resources for fire prevention. These quantities for each class were also use for the vegetation cover classifier as they give a more detailed description of the goodness of the different fits.

Data

Cartographic data for more than half a million observations were considered. Only four features ("Wilderness_Areax") were discarded from the beginning as they are only relevant for the particular mountain forest (Roosevelt National Forest) and can not be used for generalization.

The classification for the coverage type is directly given by the "Cover_Type" column but the assignment for the fire risk classes was decided considering the "Horizontal_Distance_To_Fire_Points" variable. High risk zones were assigned to values < 600 meters.

All features were considered in a first step but three of them ("Hillshade_X") were demonstrated to be irrelevant by Chi^2 test and calculating feature importance with Random Forest Method.

Algorithms

An stratified test/train split of 30% was carried out for all used algorithms. The methods described below were first tried with default settings of sklearn and hyperparameter tunning using five-fold cross validation.

- **Wildfires (Binary)**
  - K Nearest Neighbours.
  - Random Forest.
- **Wildfires (Multiclass**)
  - Gaussian Naive Bayes.
  - K Nearest Neighbours.
  - Bagging Classifier with Decision Trees.
  - Support Vector Machines.
  - Random Forest.
- **Vegetation Cover (Binary rare)**
  - K Nearest Neighbours.
- **Vegetation Cover (Binary rcommon)**
  - K Nearest Neighbours.
  - Logistic Regression
  - Decision Trees.
  - Gaussian Naive Bayes.
  - Random Forest.

Best results for the both multiclass classifications were obtained with the Random Fores methods although K Nearest Neighbours and Decision Trees gave quite good results when used (accuracies > 90%).