Final Project – Real-world problem

Developing real-life projects is the best way to sharpen your data science skills and materialize your theoretical knowledge into practical experience. Forest fire is a disaster that causes economic and ecological damage and is a threat to human lives. Thus, predicting such critical environmental issues is essential to mitigate this threat. In this final project, you will develop machine learning models to predict forest fires based on multiple features related to fire weather indices. The used dataset "Algerian_forest_fires_dataset.csv" is uploaded for you on Blackboard. The dataset includes 244 observations on two regions of Algeria, namely the Bejaia and Sidi Bel-abbes, located in the northeast and the northwest of the country respectively. The different features with their respective attribute values are described on the last page of this document.

It is very important to conduct the exploratory data analysis of the dataset.

Be sure to fit the models on a training set and to evaluate their performance on a test set. Use cross-validation in order to determine the optimal parameters for the different algorithms and assess their impacts on the performance of the obtained model. approach (Do not forget to set a random seed before beginning your analysis).

You must focus in your report on the below machine learning methods and techniques:

- Logistic regression, LDA and QDA.
- KNN with several values of K. Which value of K seems to perform the best on this data set? You can present your results graphically
- Tree-Based Methods (bagging, random forest, boosting, pruned vs unpruned trees).
- Support vector machines (using different kernels, with different values of gamma, degree and cost).

Report the confusion matrix and compute the various performance metrics. Summarize and compare the performance of the various classifiers in a table and by drawing their ROC curves and computing their AUC values.

The deadline for this project is **Sunday December 11, 2022** by midnight and it must be respected. Students failing to meet the deadline will get 20% of the grade deducted per day late.

This is a teamwork of 2-3 students max (do not forget to write your names on the top) and a **single R Notebook per team** (containing the scripts, the graphs, the analyses and comparison of the different methods' results, the interpretation of your findings and finally a conclusion) with the **knitted html file** must be uploaded on Blackboard. Your script must be very well commented.

Dataset Description

1. Date: (DD/MM/YYYY) Day, month ('june' to 'september'), year (2012)

2. Temp: temperature noon (temperature max) in Celsius degrees: 22 to 42

3. RH: Relative Humidity in %: 21 to 90

4. Ws: Wind speed in km/h: 6 to 29

5. Rain: total day in mm: 0 to 16.8

FWI Components (check this LINK for more information)

6. Fine Fuel Moisture Code (FFMC) index from the FWI system: 28.6 to 92.5

7. Duff Moisture Code (DMC) index from the FWI system: 1.1 to 65.9

8. Drought Code (DC) index from the FWI system: 7 to 220.4

9. Initial Spread Index (ISI) index from the FWI system: 0 to 18.5

10. Build-up Index (BUI) index from the FWI system: 1.1 to 68

11. Fire Weather Index (FWI) Index: 0 to 31.1

12. Classes: two classes, namely "fire" and "not fire"