BDA PROJECT PROPOSAL

STUDENTE: Grassi Riccardo MATRICOLA: 1045404
EMAIL: riccardo.grassi5@studio.unibo.it

AUSTRALIA WEATHER CLASSIFICATION
Link to dataset:
https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package

DATASET INFO

It consists of 145460 rows and 23 features (16 categorical *and* 7 numerical). This dataset contains about 10 years of daily weather observations from many locations across Australia. A list of the feature and a brief description of them is the following:

- Date, The date of observation
- Location ,The common name of the location of the weather station
- MinTemp, The minimum temperature in degrees celsius
- MaxTemp,The maximum temperature in degrees celsius
- Rainfall, The amount of rainfall recorded for the day in mm
- Evaporation ,The so-called Class A pan evaporation (mm) in the 24 hours to 9am
- Sunshine, The number of hours of bright sunshine in the day.
- WindGustDir ,The direction of the strongest wind gust in the 24 hours to midnight
- WindGustSpeed ,The speed (km/h) of the strongest wind gust in the 24 hours to midnight
- WindDir9am, Direction of the wind at 9am
- WindDir3pm, Direction of the wind at 3pm
- WindSpeed9am, Wind speed (km/hr) averaged over 10 minutes prior to 9am
- WindSpeed3pm, Wind speed (km/hr) averaged over 10 minutes prior to 3pm
- Humidity9am, Humidity (percent) at 9am
- Humidity3pm, Humidity (percent) at 3pm
- Pressure9am, Atmospheric pressure (hpa) reduced to mean sea level at 9am
- Pressure3pm ,Atmospheric pressure (hpa) reduced to mean sea level at 3pm
- Cloud9am, Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eigths. It records how many eigths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.
- Cloud3pm ,Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cload9am for a description of the values
- Temp9am, Temperature (degrees C) at 9am
- Temp3pm, Temperature (degrees C) at 3pm
- RainToday ,Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0
- RISK_MM, The amount of rain. A kind of measure of the "risk".
- RainTomorrow ,The target variable. Did it rain tomorrow?

CONTEXT

The goal of the project is to predict the variable RainTomorrow.
RainTomorrow is the target variable to predict. It means -- did it rain the next day, Yes or No? This column is Yes if the rain for that day was 1mm or more.

OBJECTIVE

In this project, I will use PySpark library to analyze the dataset and build a machine learning model to identify if will it rain tomorrow or not.

ENVIRONMENT SETUP

Cluster built with Docker containers: 1 node Master and 1 node Worker.

ANALYSIS STEPS

A list of ideal steps to carry out the analysis is the following :

- DATA CLEANING
- EDA
- FEATURE ENGINEERING
- FEATURE SELECTION
- MODELING
- HANDLING IMBALANCED CLASSES
- HYPERPARAMETER TUNING
- MODELS EVALUATION

MODELS

- Decision Tree
- Random Forest
- Logistic Regression
- GBT Classifier