University of Abdelhamid Mehri-Constantine2
Faculty: NTIC
Department: IFA
Master's second year SDIA 2024/2025

Lab 4 BDPA
MLLIB in Spark

In this lab, you will use Spark MLlib to build a model capable of predictingthe temperature based on meteorological features. In this regard, We'll work with a **weather dataset** containing historical meteorological records, including the following columns:

- Time : String
- Summary : String
- PrecipType : String
- Temperature (Target) : Float
- Apparent Temperature : Float
- Humidity : Float
- WindSpeed : Float
- WindBearing : Float
- Visibility : Float
- LoudCover : Float
- Pressure : Float
- Daily Summary : String

1. Data preparation

Import the dataset in your colab environment or past this code :

```
import kagglehub
path = kagglehub.dataset_download("budincsevity/szeged-weather")
df = spark.read.csv(path, header=True)
```

1. Convert all columns with categorical values into floats using `StringIndexer`
2. Cast the remaining columns into float
3. Create a vector assembler in which we regroup all the independent features in one output column named "`weather_features`"
4. perform a transformation into a final DataFrame containing the "`weather_features`" and the target columns

2. Training and evaluation

1. Create a Regression (which one is the best?) Object and perform training with the fit() function
2. Evaluate the resulting model and calculate the R2 metric using the `RegressionEvaluator`