Project: K-NN and Naive Bayes

Goal

Porting MapReduce-based Classification Algorithms to Spark
(*K-Nearest Neighbors & Naive Bayes*) based on these papers :
(dropbox.com/scl/fi/3wwph2l0b8a2k1scrvclq/A-MapReduce-based-k-Nearest-Neighbor.pdf)
(dropbox.com/scl/fi/04pv23898iqxukbib3doy/Na-ve-Bayes-Classifier-A-MapReduce-Approach.pdf)

Context

The two papers introduce distributed versions of K-NN and Naive Bayes using MapReduce.
In this project, you are asked to translate the logic of these algorithms into Spark (Python or Scala), using both RDDs and DataFrames, and study their scalability and efficiency.

Instructions

- Understand the MapReduce workflow for K-NN and Naive Bayes as presented in the papers
- Re-implement both algorithms in Spark using both RDD and DataFrames
- Use one of the datasets mentioned in the papers, or any **classification dataset** of your choice
- Apply Spark ML and compare performance between your own implementation and Spark ML models
- Compare **RDD vs. DataFrame** performance (execution time, clarity, flexibility)

Deliverables

- Clean and documented code in both RDD and DataFrame versions
- Final report including:
  - Algorithm understanding and Spark translation
  - Description of pipeline steps
  - Experimental setup and results
  - Comparative analysis between RDD and DataFrame
  - Discussion of results and implementation choices

Guidelines and Evaluation

Group Work

- Maximum of 4 students per group

Evaluation Criteria

1. Description of the adopted solution – 4 points
2. Algorithm design and explanation – 4 points
3. Key code fragments with comments – 3 points
4. Experimental analysis with a focus on scalability – 3 points
5. Discussion of results, highlighting strengths and weaknesses – 4 points
6. Appendix including all source code – 2 points

Deadline

The code along with a pdf version of the report have to be sent via email (dario.colazzo@lamsade.dauphine.fr) by August 31, 2025.