Project 2: Machine Learning on Weather Data

The dataset `2024.tar.gz`, available at `https://www.ncei.noaa.gov/data/global-hourly/archive/csv/`, contains hourly surface weather observations collected from stations around the world. The objective of this project is to build a machine learning model to predict the air temperature (column `TMP`) using other variables. Please refer to `https://www.ncei.noaa.gov/data/global-hourly/doc/` for a detailed description of the dataset.

This project involves model training and testing using **Apache Spark**. You may use any programming language supported by Spark, such as Python, Scala, or R. The main requirements are as follows:

- Preprocess the dataset by removing invalid values, standardizing the data, and performing any other necessary cleaning steps.
- Split the dataset into a training set (70%) and a test set (30%).
- Evaluate at least two types of machine learning models.
- Use an appropriate validation method to select the best combination of model parameters.

You should submit a report summarizing your methods and results. The report must include:

- A clear description of your approach, including data cleaning, feature selection, and the learning methods you used.
- The computing environment used for training (e.g., local machine, Google Cloud, Databricks, etc.).
- Results of model training and evaluation, including model parameters, figures showing training performance, RMSE values, and other relevant metrics.
- Any additional efforts you made to improve performance.

Please upload the following materials to Canvas no later than **5 November 2025**:

- Your complete source code and a `README` file explaining how to run it.
- The trained model.
- A report in PDF format.
- A record of your communication with AI tools (if any).

At least one member from each group should submit the above documents via Canvas. The report must clearly list the names and student IDs of all group members.