**TOPIC SUMMARY**

| Research Topic: Research and Implement Clustering and Classification algorithms for Weather Prediction Using PySpark |
| --- |
| Instructor: Nguyen Ho Duy Tri |
| Working time: from 08/05/24 to 01/06/24 |
| Short Description:  The topic focuses on implementing K-Means, KNN, SMA algorithm using PySpark for weather prediction, emphasizing the efficient use of distributed computing capabilities of Apache Spark. Initially, the data is preprocessed using techniques like StringIndexer, OneHotEncoder,… to convert categorical data into numerical format and assemble feature vectors. The dataset is then split into training and testing sets, and machine learning algorithms are implemented manually without using machine learning libraries like MLLib or Scikit-learn.. After predicting the test data labels, the results are compared with actual labels to calculate the accuracy of the model. This approach demonstrates how to leverage Spark’s parallel processing power for machine learning tasks, ensuring scalability and efficiency in handling big data. The topic highlights practical aspects of data preprocessing, model training, and evaluation in a distributed environment, showcasing the potential of Spark in building robust and scalable machine learning solutions. |