VIETNAM NATIONAL UNIVERSITY
HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

Assignment - Data Mining

BIG DATA FOR WEATHER FORECASTING (USING PYSPARK)

Instructors: Le Hong Trang

Students: Dao Quoc Khanh – 2013452
Nguyen Ngoc Hung – 2013368
Nguyen Thanh Tuan – 2014931
Nguyen Doan Hoang – 2111238

Ho Chi Minh City, 2023

Mục lục

1 Introduction

1.1 Problem statement

Weather forecasting plays a crucial role in mitigating the impact of natural disasters and supporting various sectors such as agriculture, transportation, and emergency management. However, the inherent complexity of atmospheric systems poses challenges to achieving high accuracy in weather predictions. Traditional forecasting models often struggle to capture the dynamic and nonlinear nature of weather patterns.

The advent of Big Data technologies opens new avenues for revolutionizing weather forecasting by harnessing vast amounts of diverse and real-time data sources. Despite this potential, there exists a gap in leveraging Big Data analytics to its fullest extent for enhancing weather prediction accuracy. Current forecasting models often face limitations in assimilating and processing the sheer volume and variety of data generated by satellites, weather stations, sensors, and other sources.

1.2 Applications

Improved Disaster Preparedness and Response: By leveraging Big Data analytics, meteorological agencies can provide more accurate and timely predictions of extreme weather events such as hurricanes, tornadoes, and floods. This enables better preparedness and more effective response strategies, potentially reducing the impact of disasters on communities.

Precision Agriculture:Big Data analytics in weather forecasting can benefit the agricultural sector by providing farmers with detailed and accurate information about weather patterns. This allows for optimized crop planning, irrigation management, and pest control, leading to increased crop yields and resource efficiency.

Renewable Energy Optimization: The variability of weather conditions significantly affects renewable energy sources like wind and solar power. Big Data analytics can help forecast energy production based on weather patterns, allowing energy grid operators to manage and optimize the integration of renewable energy sources into the power grid more efficiently.

Transportation and Logistics Planning: Accurate weather predictions are crucial for the transportation and logistics industry. Big Data can be used to forecast adverse weather conditions, enabling better route planning, scheduling, and resource allocation for airlines, shipping companies, and ground transportation providers.

Urban Planning and Infrastructure Management: City planners can benefit from advanced weather forecasting to anticipate and mitigate the impact of severe weather on urban infrastructure. This includes better stormwater management, flood prevention strategies, and the design of resilient buildings and transportation systems.

Epidemiological Studies and Disease Control: Big Data in weather forecasting can contribute to predicting the spread of vector-borne diseases influenced by climate conditions. This information is valuable for public health agencies in planning and implementing preventive measures, especially in regions prone to diseases like malaria or dengue.

Insurance and Risk Management: Insurance companies can use accurate weather forecasts to assess and manage risks associated with weather-related events. This includes pricing insurance policies, assessing claims, and developing risk mitigation strategies based on the probability of specific weather events occurring in a given region.

Tourism and Event Planning: The tourism industry can benefit from weather forecasts to optimize travel plans and enhance the overall tourist experience. Event planners can also use weather predictions to make informed decisions regarding outdoor events, reducing the risk of weather-related disruptions.

Scientific Research and Climate Studies: Meteorologists and climate scientists can use Big Data analytics to analyze vast datasets and gain insights into long-term climate trends. This contributes to a better understanding of climate change and supports the development of more accurate climate models.

Consumer Applications: Mobile apps and online platforms can provide personalized weather forecasts based on user preferences and location data. These applications can offer real-time updates, alerts, and recommendations for outdoor activities, contributing to improved user experiences.

1.3 How the problem has been solving together with their pros and cons

1.3.1 Standardization of Data Formats:

Solution: Implementation of standardized data formats to simplify integration across different datasets.

Pros: Streamlined integration processes, reducing complexity and improving interoperability.

Cons: Initial standardization efforts may require significant coordination and may not cover all data sources.

1.3.2 Lack of data

Solution: Crawl more data

Pros: Have more data

Cons: Cost

1.4 Why we choose our method to study

The choice to use Big Data for weather forecasting is driven by several compelling reasons, each contributing to the improvement of prediction accuracy, efficiency, and the overall effectiveness of weather forecasting systems. Here are key reasons why Big Data is chosen for weather forecasting:

- Big Data technologies handle massive volumes of diverse data, including satellite imagery, ground-based observations, and sensor networks.
- Big Data analytics can integrate diverse types of data, such as temperature, humidity, wind speed, and atmospheric pressure, from various sources.
- Big Data technologies allow for the integration of high-resolution data from a wide range of sources.
- Big Data analytics, including machine learning models, can identify complex patterns and relationships within the data.
- Big Data technologies enable real-time processing of vast datasets.
- Big Data analytics can adapt to evolving weather patterns and changing data sources.
- Big Data technologies are designed to scale horizontally, handling growing datasets and computational demands.
- Big Data platforms facilitate data sharing and collaboration among meteorological agencies and research institutions.
- Big Data analytics enables the development and implementation of innovative forecasting techniques.

In summary, the adoption of Big Data for weather forecasting is driven by the desire to leverage the vast potential of diverse, high-volume, and real-time data sources. This choice aims to overcome the limitations of traditional forecasting methods and usher in a new era of more accurate, localized, and responsive weather predictions.

2 Approaches

2.1 Overview description

In the landscape of weather forecasting, the integration of big data technologies like PySpark has ushered in a transformative era, significantly enhancing our capacity to comprehend and predict weather patterns with unprecedented precision and accuracy. PySpark, a robust distributed computing framework built upon Python, stands as a pivotal tool empowering meteorologists and data scientists to process and analyze colossal volumes of weather-related data. This technological advancement has not only streamlined the handling of vast datasets but has also accelerated the analysis of intricate atmospheric conditions. Leveraging PySpark's distributed architecture, weather data sourced from diverse outlets including historical records, satellite imagery, and real-time observations are assimilated and stored in distributed systems such as Hadoop Distributed File System (HDFS) or cloud-based platforms. Subsequently, a meticulously designed series of preprocessing steps within PySpark ensues, encompassing data cleaning, feature extraction, and transformation to prepare the collected data for sophisticated analysis. PySpark's parallel computing capabilities play a pivotal role in this phase, enabling efficient data processing on distributed clusters and ensuring scalability to manage the sheer magnitude of weather data. This streamlined process serves as a foundation for deploying various analytical techniques and machine learning algorithms to decipher patterns, correlations, and trends embedded within the vast array of weather data. The resultant predictive models, constructed utilizing PySpark's scalable infrastructure, leverage historical weather data to forecast future weather conditions with increased precision. This amalgamation of big data technologies, particularly PySpark, has redefined the landscape of weather forecasting, fostering improved preparedness for weather-related events and catering to diverse industries reliant on accurate weather predictions.

2.2 Architecture/method pipeline

The architecture for weather forecasting using PySpark involves a multi-stage pipeline that encompasses data collection, preprocessing, analysis, and model development:

- Data Collection: We collect large volumes of weather data, including historical records that are gathered from diverse resources and stored in a CSV file.
- Data Preprocessing: We conduct a series of preprocessing steps on the collected weather data before analyzing and processing it within PySpark. This invole cleaning the data, extracting pertinent features and tranforming it to suitable format for analysis and model process.

- Data Analysis
- Model Development: Machine Learning model are constructed using historical weather data. These models can range from traditional statistical models to sophisticated deep learning architectures, aiming to predict future weather conditions.

2.3 Main steps

The main steps involved in using PySpark for weather forecasting can be outlined as follows:

1. Load the historical weather data as a DataFrame from a CSV file.
2. Explore and preprocess the data, such as handling missing values, outliers, or noise.
3. Extract relevant features from the data, such as date, time, location, humidity, pressure, wind speed, etc.
4. Split the data into training and testing sets.
5. Train the model on training set.
6. Evaluate the model on the testing set.
7. Use the model to make predictions on new or unseen data.

In conclusion, PySpark's role in weather forecasting showcases its ability to handle vast amounts of data and complex computations, ultimately contributing to more precise and reliable weather predictions. This technology continues to drive innovation in meteorology, enabling better preparedness for weather-related events and benefiting various industries dependent on weather information.

3 Experiments

3.1 Dataset

3.1.1 Description

The original dataset comes from historical-hourly-weather-data (contain data from 2012 to 2017). Besides, we extend the dataset by adding weather measurements data of Ho Chi Minh city as well as data from 2017 - 2023 from worldweatheronline.com.

The dataset contains hourly weather measurements data of 37 cities (e.g. temperature, humidity, air pressure,...), collected from 2012 to 2023, composed by 7 different csv files:

- *1 csv file containing geographical information about the different cities.*
- *1 csv file containing the textual description of the weather conditions, where each column refers to a different city and each row refers to a specific datetime in which the weather condition occurred.*
- *5 csv files, for each weather measurement type (humidity, pressure, temperature, wind direction, wind speed), where each column refers to a different city and each row refers to the specific datetime of the measurement.*

After the data cleaning and integration process, we obtained **107** csv files containing weather measurement data from 2012 to 2023, with totally **1.193.632** samples, where each column refers to a weather measurement type or geographical information, and each row refers to a sample, corresponding to a specific datetime of the measurement.

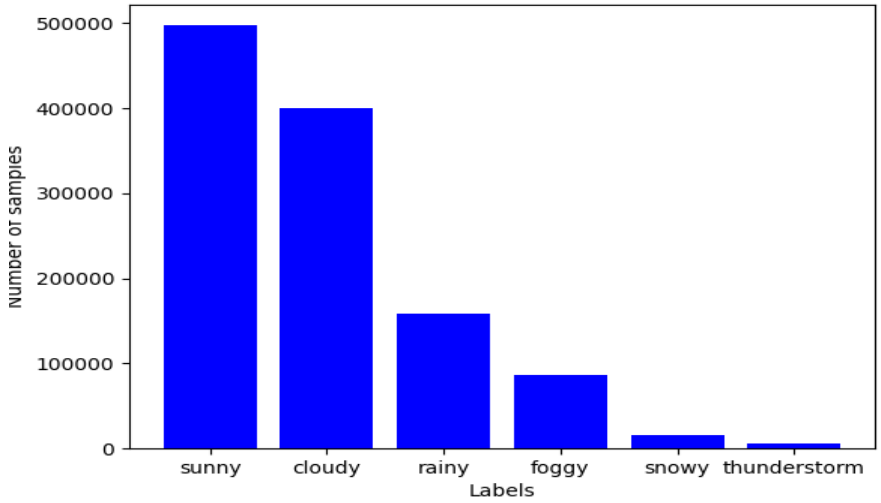
1	datetime	humidity	pressure	temperatu	wind_dire	wind_spee	weather_c	city	country	latitude	longitude
2	########	62	1014	287.1611	290	2	cloudy	Philadelph	United Sta	39.95234	-75.1638
3	########	49	1015	289.69	230	3	cloudy	Philadelph	United Sta	39.95234	-75.1638
4	########	88	1014	292	40	3	foggy	Philadelph	United Sta	39.95234	-75.1638
5	########	100	1015	291.15	40	2	foggy	Philadelph	United Sta	39.95234	-75.1638
6	########	100	1018	294.25	0	0	foggy	Philadelph	United Sta	39.95234	-75.1638
7	########	88	1019	288.05	250	3	foggy	Philadelph	United Sta	39.95234	-75.1638
8	########	100	1019	287.41	270	1	foggy	Philadelph	United Sta	39.95234	-75.1638
9	########	81	1020	282.22	270	3	foggy	Philadelph	United Sta	39.95234	-75.1638
10	########	87	1022	282.36	360	4	foggy	Philadelph	United Sta	39.95234	-75.1638

3.1.2 Weather condition processing

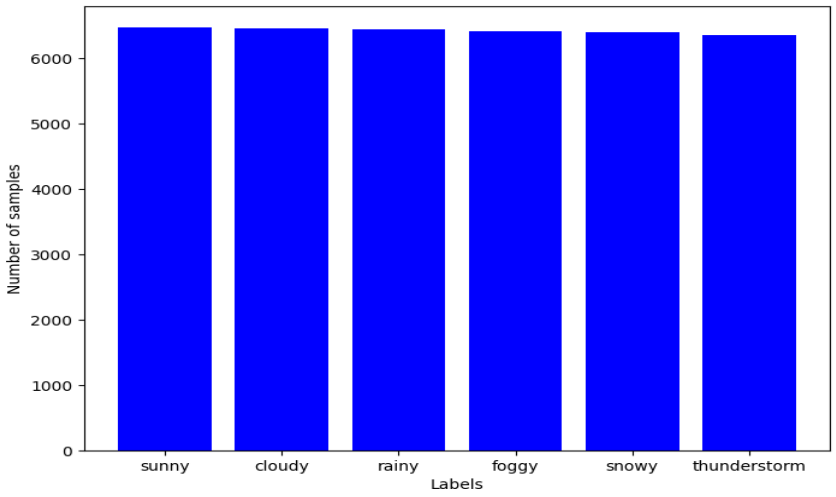
We consider *weather_condition* column as the target column for classification task.

There are more than 90 different weather conditions. But the target classes are to sparse in the original dataset, some occur few times and a lot of them are really similar between each other.

Therefore, we decided to aggregate them into 6 common weather condition classes: *thunderstorm, cloudy, rainy, foggy, snowy, sunny*.



The classes aggregation led to a huge class imbalance. To avoid bias in the classification output, we decided to undersample the dataset. The result we get after this process is about **6000** samples for each class.



3.2 Machine Learning pipeline

3.2.1 Train and test split

In this assignment, we use **80%** of the dataset for the train set, and **20%** for the test set.

3.2.2 Data encoding pipeline

We use the functions available in PySpark sequentially to encode data for machine learning pipeline:

- *StringIndexer:* Convert categorical label to numerical.
- *VectorAssembler:* Assemble features into a vector.
- *StandardScaler:* Normalize the features with mean and standard deviations.

3.2.3 Machine Learning models

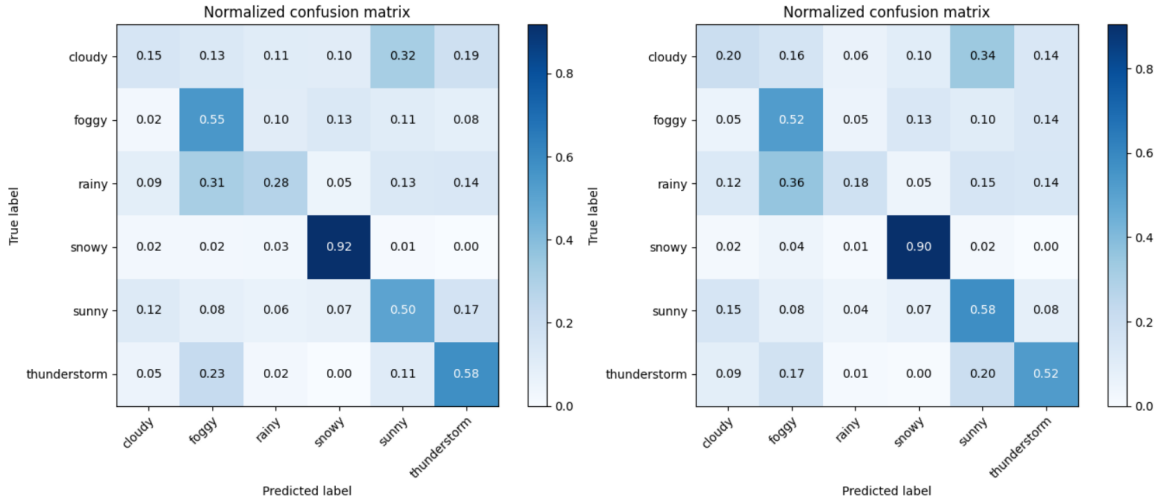
In this assignment, we construct and evaluate some machine learning models in order to predict weather condition based on several features (*humidity*, *pressure*, *temperature*, *wind_direction*, *wind_speed*, *latitude*, *longtitude*):

- *Random Forest;*
- *Decision Tree;*
- *Logistic Regression;*
- *Multilayer Perceptron.*

3.2.4 Evaluate

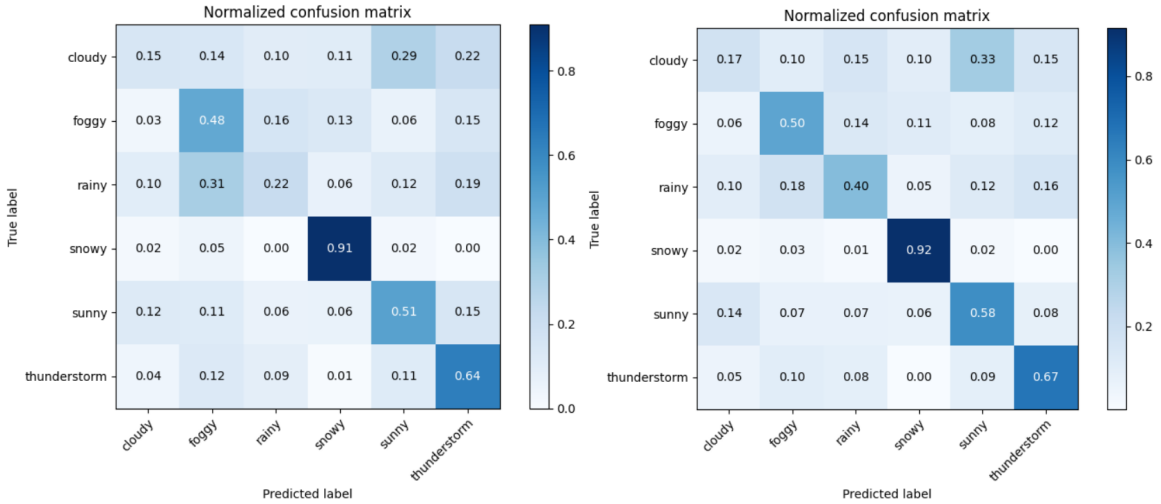
After training process, we evaluate those models with some kind of metrics: ***accuracy***, ***precision***, ***recall*** and ***F1-score***.

Model	Parameters	accuracy	precision	recall	F1-score
Random Forest	numTrees = 10 maxDepth = 5	49.96	47.95	49.96	47.63
Decision Tree	maxDepth = 5	48.67	48.16	48.67	46.34
Logistic Regression	maxIter = 500 regParam = 0.0 elasticNetParam = 0.0	48.52	45.53	48.52	45.82
Multilayer Perceptron	layers = [input, 16, 32, 64, 128, output] maxIter = 2000 blockSize = 128	54.42	51.94	53.55	52.53



(a)

(b)



(c)

(d)

Hình 1: *(a) Random Forest, (b) Decision Tree, (c) Logistic Regression, (d) Multilayer Perceptron*

In this experiment, we achieved the best results on *Multilayer Perceptron* with:

- **accuracy:** 54.52
- **precision:** 51.94
- **recall:** 53.55
- **F1-score:** 52.53

3.3 OpenWeather comparison

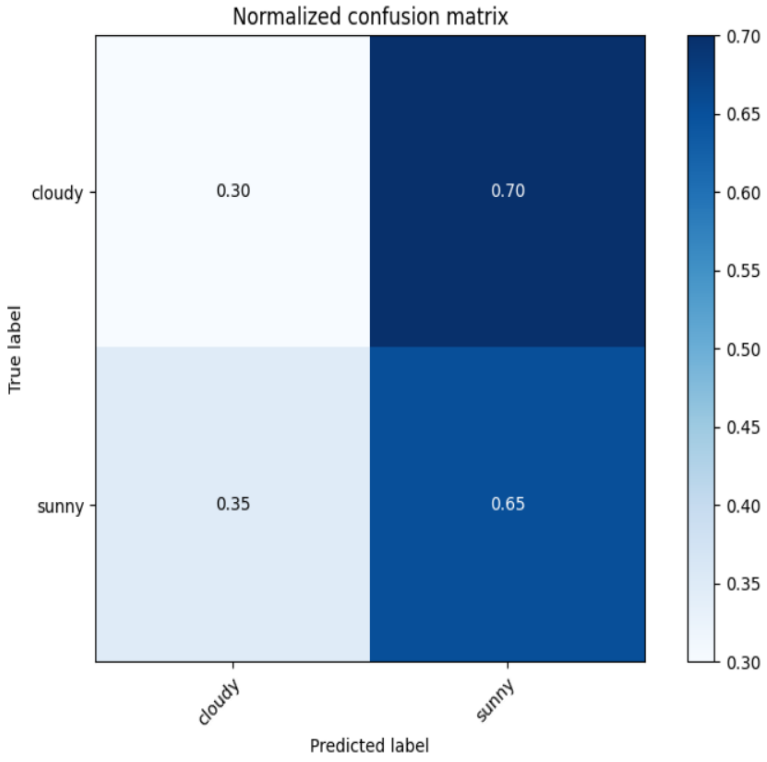
We test and evaluate our Multilayer Perceptron model by taking weather data information for the next 7 days in Phoenix, Arizona (from November 19, 2023), each recording is 3 hours apart.

Result:

- **accuracy:** 58.92

- **precision:** 69.42
- **recall:** 58.93
- **F1-score:** 63.07

Confuse matrix:



As we can see, the trained model still has a bias towards the ***"sunny"*** label. This is due to ambiguity, imbalance in labeling, geographical location, as well as limited number of records in our data.

4 Conclusion

Based on the original dataset, we observed a significant imbalance in both geographical locations and the time period of recordings (2012-2017). This could greatly impact the model's predictive ability for different geographical locations at the current time. To address this issue, we expanded and enhanced the original dataset by collecting additional weather data recorded from 2017-2023. We also included data from different geographical locations around the Earth, such as Ho Chi Minh City.

However, due to limitations in finances, hardware, and time, the augmented dataset has not yielded satisfactory results in the training and evaluation of machine learning models, as well as testing on real-world data. This is attributed to the ongoing imbalance in the data, specifically in the number of records for each label (most data collection locations being cities in the United States and Canada), inconsistency in time (the majority of our records are distributed within the time range of 2012-2017), and the inherent fuzziness and limitations in the number of records we obtained, which significantly affects the results. While the model performed relatively well for weather data from 2019 and earlier, it did not effectively generalize to recently recorded weather data (showing bias towards certain labels).

We also attempted an approach based on Recurrent Neural Networks (RNN). However, due to constraints in time, hardware, and data, the result achieved did not yield promising performance.

References

[1] andrea-gasparini , *big-data-weather-forecasting*
https://github.com/andrea-gasparini/big-data-weather-forecasting, 2021

[2] Apache Spark , *PySpark Overview*
https://spark.apache.org/docs/latest/api/python/index.html, 2023

[3] Neelam Tyagi , *Weather Forecasting: How Does Big Data Analytics Magnify it?*, 2021.

[4] World Weather Online
https://www.worldweatheronline.com