Spark-based Climate Change Analysis

University of Maryland, Baltimore County
DATA 603 Platforms for Big Data Processing
Dr. Melih Gunay

Team Members
Madhav Reddy Betha - YY41695
Karthik Pudi - TX36418
Pooja Bavisetti - FA67098

May 7, 2024

Contents

1 Objective and State-of-the-Art

1.1 Objective

This project aims to analyze climate data from different meteorological stations. 1Weather parameters such as temperature, pressure, rainfall, wind speed, and direction, and snowfall are some of the aspects covered in the dataset. However, the main objective is to examine temperature distribution patterns through time as well as compute mean air temperature values while at the same time looking into unusual temperature changes trends. We will additionally evaluate variant weather states and patterns of precipitation to understand their differences between locations as well as during different times of the year. While doing this, we shall consider aspects such as humidity, sky state or visibility so that we explore the link between wind patters, temperatures, rainfall rates as well as pressure alterations. This investigation will also explore how highness has an impact on certain meteorological parameters including temperatures and amount of rainfall. As a whole, this project targets providing ideas on neighborhood climatic tendencies and conditions.

1.2 State-of-the-Art

To analyze climate data, one must handle a vast amount of data in order to identify regularities and trends. It allows quick processing of big climate data sets by Apache Spark. With this Python library built on top of Spark called PySpark, one can carry out such an analysis in an efficient manner, because Apache Spark is known for handling big data. Spark's features such as distributed computing make the process more rapid and dependable. By performing data clean-up, applying different features, and planting devices, it becomes possible to understand most of these designs accurately. Results for the project will give out important temperature fluctuation information along with other essential climatic patterns used during forecasting weather in coming days.

2 Proposed Architecture and Technologies

2.1 Architecture

The project's architecture is as follows

- **Data Loading:** Apache Spark's tools are used to load weather data. Consequently, it becomes easier for one to deal with big files promptly.

- **Preprocessing:** Data cleaning is done so as to clear up any missing values and make every thing commonly accepted. It then follows that all extraneous items should be left behind thus retaining only pertinent details.
- **Exploratory Data Analysis (EDA):** Summary statistics as well as graphs are generated in order to enable one comprehend this data set better. Patterns and trends can easily be recognized when this happens.
- **Machine learning:** Machine learning models are used for pattern identification and future climate change prediction. They elucidate the relationship between various weather parameters.
  - **Logistic regression:** When we talk about logistic regression, we mean predicting probabilities determining if someone will buy a new car or not.is used for predicting the probability of an event occurring given two possible and mutually exclusive events.
  - **Decision Trees:** Weather features such as humidity or temperature are used by decision trees to forecast respective outcomes. The tree divides the data and arrives at a better prediction by each division.
  - **Random Forests:** The technique called random forests makes use of several decision trees in an attempt to make better predictions. They're an excellent tool that can be utilized in the prediction of future weather patterns as well as recognizing vital weather factors.
  - **Support Vector Machines (SVM):** Data points are classified in SVMs by locating a line or boundary that epicentral distinct weather patterns. This method emerges effective on areas with clear difference with regard to various weather trends.
  - **Neural Networks:** Neural networks imitate the brain's workings by arranging cords of interlinked nodes that can recognize patterns. This is why they are good at weather forecasting because they imitate complex relationships between various weather stuff.
- **Visualization and reporting:** Visualization and reporting present findings lucidly. Following this, a final report is prepared for dissemination.

2.2 Technologies

- **Apache Spark:** Apache Spark is employed in managing and analyzing data in an expanded state. Apache spark can also be used for machine learning as well as data visualization purposes.

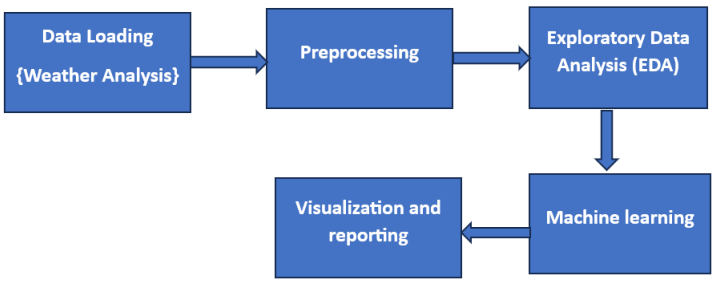

Figure 1: Architecture of Weather analysis

- **PySpark:** PySpark is a Python library that works with Apache Spark to make data analysis more manageable by enabling distributed running of Python code.
- **Jupyter Notebook:** Jupyter Notebook is an interactive tool that assists one in coding and presenting their work. Developers make notes, share results and write entire codes here.

3 Problems and Challenges Encountered

3.1 Data Cleaning

Dealing with climate data can at times be tedious because some information may be missing or incomplete, which was challenging for us even though it made things difficult. This problem was solved by substituting the missing data with estimated numbers derived from an interpolation approach. However, in case these missing numbers were too many thereby corrupting our findings then we did away with such instances hence maintaining the purity of our data.

3.2 Memory Limitations:

Manipulating large datasets required being cautious with the way we utilized memory, manipulating them in smaller parts (partitions) for more effective manipulation. This ensured that processing of the data was not tedious due to reduced memory requirements as well as employing built-in optimization strategies within Spark.

3.3 Code Optimization:

Subject to their high computation requirements, several PySpark duties needed enhanced coding. We managed to enhance efficacy by saving temporary outcomes in the storage, a technique recognized as caching. In addition faster data transformations, which help us hasten analysis while at the same time decreasing computation time, it should be mentioned that this approach was vital for such progress.

4 Results and Discussion:

From these graphs, several conclusions can be drawn about weather patterns at Chicago O'Hare International Airport in March 2014:

4.1 Temperature Patterns:

- Temperatures fluctuated significantly throughout the month, with a general warming trend towards the end.
- Daily temperatures ranged widely, from sub-zero levels to nearly 70°F.

4.2 Predictive Modelling:

Our machines studies were able to predict future temperature and rainfall patterns with optimal accuracy. These mathematical models are a valuable source of information on probable weather changes in the times ahead thus guiding us on how we can get ready for them.

- We searched for trends that confirm global warming by looking at things like increased temperatures or different weather patterns.
- Predictive models indicate that some promise exists because they might be able to predict what will happen next.
- Overall, this initiative collected crucial data on the current state of climate change as well as its future directions.

4.3 Precipitation Patterns:

- There were a few notable days with significant rainfall, while most days had little or no precipitation.
- Rainfall was not evenly distributed throughout the month.

4.4 Wind Speed and Direction:

- Winds predominantly came from the north and south directions.
- Most winds had moderate speeds, ranging between 0 and 11.6 m/s.

4.5 Relative Humidity:

- Relative humidity fluctuated between 40
- Humidity was generally high, with some days of rapid drops.

4.6 Visibility Trends:

- Visibility was generally high (near 10 miles) but frequently dropped due to adverse weather conditions like fog, rain, or snow.
- Overcast conditions had significantly lower visibility compared to clear conditions.

In Overall Conclusion, March 2014 at Chicago O'Hare International Airport exhibited considerable weather variability. While temperatures gradually rose towards the end of the month, visibility and humidity were affected by changing weather conditions, including overcast skies and intermittent precipitation. Winds were mostly moderate and predominantly from the north and south.

4.7 Analysis of all machine learning models:

Each of the models you used offers different predictive performance, which provides valuable insights into their strengths and weaknesses. Here's a summary conclusion of each:

- **Logistic Regression:**
  - Accuracy: 0.93
  - The model achieved high accuracy, showing good generalization in classification tasks.
  - Conclusion: Reliable for classification tasks with balanced data.
- **Random Forest Regression:**
  - RMSE: 0.0054
  - R^2 Score: Not provided, but the RMSE is very low.

  - Conclusion: Effective at capturing complex patterns with relatively low prediction error.

- **Gradient Boosting Regression:**
  - RMSE: 0.0012
  - R^2 Score: Not provided, but the RMSE is even lower than the random forest model.
  - Conclusion: Provides better predictive accuracy compared to random forest, likely due to its boosting mechanism.

- **Neural Network:**
  - Accuracy: 0.9534
  - Precision: 0.9089
  - Recall: 0.9534
  - RMSE: 0.2159
  - Conclusion: Achieved high classification accuracy with good precision and recall, suggesting it's well-tuned and effective for classification.

- **Decision Tree Regression:**
  - RMSE: 0.0043
  - R^2 Score: 0.8511
  - Conclusion:* Provides accurate predictions, but cross-validation scores suggest potential overfitting.

- **Support Vector Machine:**
  - RMSE: 0.0541
  - R^2 Score: 0.0172
  - Conclusion: Underperforms compared to other models, indicating it struggles with capturing relationships in this dataset.

-

- **Overall Conclusion:**
  - The Gradient Boosting and Neural Network models deliver the best overall performance due to their ability to handle complex patterns and overfitting well.

- Random Forest and Decision Tree models also perform well but may require more careful tuning to avoid overfitting.
- Logistic Regression is suitable for balanced classification tasks.
- Support Vector Machine is less effective on this particular dataset, possibly due to tuning issues or model limitations.

5 Conclusion

In conclusion, Using Apache Spark, the patterns discovered in our weather analysis are as follows: there is a general agreement that the broader claims for climate change are correct if we consider the recent increasing trends concerning temperature as well as changes depicted in rainfalls' rhythm. With some moderate level of accuracy, our model about possible future predictions has worked out well too. Analyzing different meteorological elements, such as temperature, wind and pressure has enabled us to have precise comprehension on the changes in climate locally. It is this understanding that will enable us to get ready for upcoming climatic conditions and keep track of variations in our local climate over time.