



Gyalpozhing College of Information Technology

Royal University of Bhutan

Kabjisa, Chamjekha, Thimphu

School of Computing

Bachelor of Computer Science

Ethics in Computing and Interactive Design, Autumn 2024

Big Data Project Final Report

Topic: Temperature Prediction using Big-Data

Submitted by:

Kinley Namgay(12210059)

Tutor: Ms. Pema Yangden

Table of Contents

1. Introduction
 - 1.1 Problem Statement
 - 1.2 Objectives
 - 1.3 Project Relevance
2. Data Collection and Management
 - 2.1 Data Sources
 - 2.2 Data Characteristics
 - 2.3 Data Handling and Modifications
3. Data Preprocessing
 - 3.1 Handling Missing and Inconsistent Data
 - 3.2 Feature Engineering
 - 3.2.1 Combining Date Components (day, month, year)
 - 3.3 Indexing Categorical Columns
 - 3.4 Final Dataset Preparation
4. Exploratory Data Analysis
 - 4.1 Distribution of Average Temperature
 - 4.2 Seasonal Trends Analysis
 - 4.3 Regional Variations
 - 4.4 Correlation Analysis
5. Model Development and Evaluation
 - 5.1 Model Selection: Gradient Boosted Trees (GBT)
 - 5.2 Feature Selection for Machine Learning
 - 5.3 Model Training and Testing
 - 5.4 Model Performance Evaluation
 - 5.4.1 RMSE
 - 5.4.2 R^2
6. Findings and Insights
 - 6.1 Key Trends in Temperature Data
 - 6.2 Model Prediction Accuracy
 - 6.3 Interpretations of Results
7. Discussion and Implications
 - 7.1 Addressing the Problem Statement
 - 7.2 Real-World Applications of Findings
 - 7.3 Limitations of the Study
8. Recommendations and Future Work
 - 8.1 Recommendations for Stakeholders
 - 8.2 Areas for Further Research
9. Conclusion
10. References

1. Introduction and Background

Problem Statement

Temperature forecasting is a critical component of weather prediction, with significant implications for agriculture, energy planning, public safety, and daily life. Accurate temperature forecasts are essential for activities such as crop management, power grid optimization, and preparation for extreme weather conditions. However, traditional statistical forecasting methods, such as autoregressive models, often struggle with scalability and accuracy when processing large volumes of real-time and historical weather data. Predicting temperature trends, particularly during extreme weather events like heat waves or cold spells, remains a challenge due to the complex atmospheric dynamics involved.

With the advent of Big Data and advanced machine learning techniques, new opportunities have emerged to enhance the precision and timeliness of temperature forecasting. This project seeks to address these challenges by developing an advanced temperature forecasting model that leverages Big Data tools to process large-scale and historical weather data. The goal is to improve the accuracy of temperature predictions, enabling better decision-making and enhancing preparedness for extreme weather events.

Objectives

1. To build a predictive model capable of efficiently processing large-scale weather data focused on temperature forecasting.
2. To improve the accuracy of temperature predictions by leveraging machine learning algorithms and Big Data tools.
3. To identify and forecast temperature trends, including sudden fluctuations, seasonal patterns, and extreme events such as heatwaves and cold spells.
4. To enhance forecasting reliability by integrating historical temperature data
5. To develop a framework that minimizes forecasting errors, improving decision-making for weather-sensitive sectors such as agriculture, energy, and urban planning.

Relevance

The project has significant relevance in several domains:

1. **Agricultural Planning:** Providing accurate temperature forecasts to help farmers optimize planting schedules, irrigation, and harvesting, thereby improving crop yields and minimizing losses due to extreme temperature variations.
2. **Disaster Preparedness:** Delivering timely predictions of temperature extremes, such as heatwaves and cold spells, to help communities and emergency services prepare and respond effectively.
3. **Energy Management:** Supporting efficient energy planning and distribution by forecasting temperature fluctuations, which directly impact heating and cooling demands.
4. **Urban Weather Modeling:** Informing infrastructure and urban planning by predicting temperature patterns that affect transportation systems, building design, and overall livability in cities.

2. Data Collection and Management

Big Data Source: Visual Crossing Weather API and datasets which contain historical weather data of different regions. The dataset contains 2.9 million entries which tracks daily average temperatures for various cities across different countries and regions from different time periods.

Updated Dataset Characteristics

The dataset now includes two additional features—**precipitation** and **wind speed**—to enhance the predictive capabilities of the temperature forecasting model. These new data points complement the temperature records and provide critical contextual information for weather patterns.

It contains the following columns:

1. Region: The geographical region (e.g., Africa, Asia, America).
2. Country: The country where the city is located.
3. State: The state or province
4. City: The specific city for the temperature records.
5. Precipitation: Daily precipitation levels measured in millimeters (mm).
6. Wind Speed: Average wind speed measured in kilometers per hour (km/h).
7. Month: The month (as an integer) of the record.
8. Day: The day (as an integer) of the record.
9. Year: The year of the temperature record.
10. AvgTemperature: The average temperature recorded on that date in Fahrenheit

Importance of the New Data

1. **Precipitation:** Provides context for temperature changes, particularly during extreme weather events such as storms or heatwaves, where precipitation often interacts with atmospheric temperature.
2. **Wind Speed:** Offers insights into cooling effects (wind chill) or warming influences (foehn winds), contributing to a more comprehensive understanding of temperature dynamics.

Sqoop (SQL-to-Hadoop) was utilized to efficiently transfer data from the **MySQL relational database** to the **Hadoop Distributed File System (HDFS)**. As an open-source tool, Sqoop facilitates seamless integration between relational databases and Hadoop, making it ideal for handling the large-scale weather dataset used in this project.

```
PS C:\Users\kinle\Desktop\project\.devcontainer> docker network create my_network
9c5eab36b0d9a6d9e11fd391077029f76fa5e80840aaee1dd5d98f5d81512829
```

```
PS C:\Users\kinle\Desktop\project\.devcontainer> docker network connect my_network project_devcontainer-mysql-1
PS C:\Users\kinle\Desktop\project\.devcontainer> docker network connect my_network project_devcontainer-hadoop-1
PS C:\Users\kinle\Desktop\project\.devcontainer> docker network inspect my_network
[
  {
    "Name": "my_network",
    "Id": "9c5eab36b0d9a6d9e11fd391077029f76fa5e80840aaee1dd5d98f5d81512829",
    "Created": "2024-11-15T19:59:16.842595547Z",
    "Scope": "local",
    "Driver": "bridge",
    "EnableIPv6": false,
    "IPAM": {
      "Driver": "default",
      "Options": {},
      "Config": [
        {
          "Subnet": "172.19.0.0/16",
          "Gateway": "172.19.0.1"
        }
      ]
    },
    "Internal": false,
    "Attachable": false,
    "Ingress": false,
    "ConfigFrom": {
      "Network": ""
    },
    "ConfigOnly": false,
    "Containers": {
      "69a023239e1f02ef6d510643b6398806aff8d9f050e4c7c5c8242b0fba613550": {
        "Name": "project_devcontainer-mysql-1",
        "EndpointID": "88c8322998f7feb14d4871fb278406b3486904902a25df73c2755f572ddba5cc",
        "MacAddress": "02:42:ac:13:00:02",
        "IPv4Address": "172.19.0.2/16",
        "IPv6Address": ""
      }
    }
  }
]
```

Process Overview:

1. **Source Database:** The data was initially stored in a **MySQL** database named **WeatherDatabase**.
2. **Data Transfer:** Sqoop was used to import the **CityTemperature** table from MySQL into HDFS for further processing and analysis.
3. **HDFS Target Directory:** The imported data was stored in HDFS under **/user/hadoop/CityTemperature**.
4. Once ingested into HDFS, the data is now ready for distributed processing using tools such as **Apache Spark**, **Hive**, or **MapReduce**, ensuring scalability and reliability for predictive modeling tasks.

```
root@cfa0efcd8b46:/# sqoop import \
> --connect jdbc:mysql://172.19.0.2:3306/city_temperature \
> --username hive \
> --password hive \
> --table city_temperature \
> --target-dir /user/hadoop/users_data \
> --num-mappers 1
```

```
root@b694ba64e924:~# hdfs dfs -ls /user/hadoop/CityTemperature
Found 2 items
-rw-r--r-- 1 root supergroup 0 2024-10-09 12:47 /user/hadoop/CityTemperature/_SUCCESS
-rw-r--r-- 1 root supergroup 140759863 2024-10-09 12:47 /user/hadoop/CityTemperature/part-m-00000
```

3.Data Processing and Analysis

The data processing and analysis pipeline for this project was designed to handle large-scale weather data efficiently, using **PySpark** to preprocess, transform, and analyze the dataset. Below is a step-by-step explanation of the process:

Tools Used

- **PySpark**: For distributed data processing and analysis.
- **HDFS (Hadoop Distributed File System)**: For scalable data storage and retrieval.
- **Gradient Boosted Trees (GBT)**: For predictive modeling and analysis.

1.Dataset Inspection:

- The number of rows and columns was identified.

```
10 columns and 1081185 rows.
```

- Null and inconsistent values were examined using `isnan` and filtering techniques.

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|region|country| state|city|precipitation|wind|month|day|year|avgtemperature|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|      0|      0|377494|  0|              0|  0|  0|  0|  0|              0|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|region|country|state|city|precipitation|wind|month|day|year|avgtemperature|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|  0.0|  0.0|100.0| 0.0|              0.0| 0.0| 0.0|0.0| 0.0|              0.0|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
```

2.Handling Missing Values:

- Missing temperature values (e.g., `-99`) were replaced with the mean temperature calculated from the dataset.

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|region|country|state| city|precipitation|wind|month| day|year|avgtemperature|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|Africa|Algeria| NULL|Algiers|          0.0| 4.7|  1.0|  1.0|1995|          64.2|
|Africa|Algeria| NULL|Algiers|         10.9| 4.5|  1.0|  2.0|1995|          49.4|
|Africa|Algeria| NULL|Algiers|          0.8| 2.3|  1.0|  3.0|1995|          48.8|
|Africa|Algeria| NULL|Algiers|         20.3| 4.7|  1.0|  4.0|1995|          46.4|
|Africa|Algeria| NULL|Algiers|          1.3| 6.1|  1.0|  5.0|1995|          47.9|
|Africa|Algeria| NULL|Algiers|          2.5| 2.2|  1.0|  6.0|1995|          48.7|
|Africa|Algeria| NULL|Algiers|          0.0| 2.3|  1.0|  7.0|1995|          48.9|
|Africa|Algeria| NULL|Algiers|          0.0| 2.0|  1.0|  8.0|1995|          49.1|
|Africa|Algeria| NULL|Algiers|          4.3| 3.4|  1.0|  9.0|1995|          49.0|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
```

3.Data Cleaning

- Columns such as **state** were dropped for simplicity.
- Inconsistent records (e.g., years before 1995 or days set to 0) were identified and addressed.

region	country	city	precipitation	wind	month	day	year	avgtemperature
Africa	Ethiopia	Addis Ababa	20.24	29.4	12	3	201	63.41698233023292
Africa	Ethiopia	Addis Ababa	441.97	3.97	12	4	201	63.41698233023292

4.Feature Engineering:

- A new **season** feature was added based on the **month** column to classify records into **Winter**, **Spring**, **Summer**, and **Autumn**.
- Date values were combined into a single **date** column for time-based analysis.

region	country	city	precipitation	wind	month	day	year	avgtemperature	season	date
Africa	Algeria	Algiers	0.0	4.7	1	1	1995	64.2	Winter	1995-01-01
Africa	Algeria	Algiers	10.9	4.5	1	2	1995	49.4	Winter	1995-01-02
Africa	Algeria	Algiers	0.8	2.3	1	3	1995	48.8	Winter	1995-01-03
Africa	Algeria	Algiers	20.3	4.7	1	4	1995	46.4	Winter	1995-01-04
Africa	Algeria	Algiers	1.3	6.1	1	5	1995	47.9	Winter	1995-01-05

5.Categorical Encoding:

- Columns such as **region**, **country**, **city**, and **season** were encoded into numerical indices using **StringIndexer** for compatibility with machine learning algorithms.

region	country	city	precipitation	wind	month	day	year	avgtemperature	season	date	region_index	country_index	city_index	season_index
Africa	Algeria	Algiers	0.0	4.7	1	1	1995	64.2	Winter	1995-01-01	2.0	18.0	6.0	1.0
Africa	Algeria	Algiers	10.9	4.5	1	2	1995	49.4	Winter	1995-01-02	2.0	18.0	6.0	1.0
Africa	Algeria	Algiers	0.8	2.3	1	3	1995	48.8	Winter	1995-01-03	2.0	18.0	6.0	1.0
Africa	Algeria	Algiers	20.3	4.7	1	4	1995	46.4	Winter	1995-01-04	2.0	18.0	6.0	1.0
Africa	Algeria	Algiers	1.3	6.1	1	5	1995	47.9	Winter	1995-01-05	2.0	18.0	6.0	1.0
Africa	Algeria	Algiers	2.5	2.2	1	6	1995	48.7	Winter	1995-01-06	2.0	18.0	6.0	1.0
Africa	Algeria	Algiers	0.0	2.3	1	7	1995	48.9	Winter	1995-01-07	2.0	18.0	6.0	1.0

6. Feature Assembly:

- Relevant features, including **region_index**, **country_index**, **wind**, **precipitation**, and temporal variables (**year**, **month**, **day**), were combined into a single feature vector using **VectorAssembler**.

region	country	city	precipitation	wind	month	day	year	avgtemperature	season	date	region_index	country_index	city_index	season_index	features
Africa	Algeria	Algiers	0.0	4.7	1	1	1995	64.2	Winter	1995-01-01	2.0	18.0	6.0	1.0	[2.0,18.0,6.0,4.7...]
Africa	Algeria	Algiers	10.9	4.5	1	2	1995	49.4	Winter	1995-01-02	2.0	18.0	6.0	1.0	[2.0,18.0,6.0,4.5...]
Africa	Algeria	Algiers	0.8	2.3	1	3	1995	48.8	Winter	1995-01-03	2.0	18.0	6.0	1.0	[2.0,18.0,6.0,2.3...]
Africa	Algeria	Algiers	20.3	4.7	1	4	1995	46.4	Winter	1995-01-04	2.0	18.0	6.0	1.0	[2.0,18.0,6.0,4.7...]
Africa	Algeria	Algiers	1.3	6.1	1	5	1995	47.9	Winter	1995-01-05	2.0	18.0	6.0	1.0	[2.0,18.0,6.0,6.1...]
Africa	Algeria	Algiers	2.5	2.2	1	6	1995	48.7	Winter	1995-01-06	2.0	18.0	6.0	1.0	[2.0,18.0,6.0,2.2...]
Africa	Algeria	Algiers	0.0	2.3	1	7	1995	48.9	Winter	1995-01-07	2.0	18.0	6.0	1.0	[2.0,18.0,6.0,2.3...]

7. Data Saving:

- The preprocessed dataset was saved back to HDFS in Parquet format for future use.

```
bash-4.2$ hdfs dfs -ls /output/preprocessed_data
Found 9 items
-rw-r--r-- 3 spark hadoop 0 2024-11-20 20:05 /output/preprocessed_data/_SUCCESS
-rw-r--r-- 3 spark hadoop 2854546 2024-11-20 20:05 /output/preprocessed_data/part-00000-ac51e89d-1b25-4759-bf42-83dd3586291f-c000.snappy.parquet
-rw-r--r-- 3 spark hadoop 3172232 2024-11-20 20:05 /output/preprocessed_data/part-00001-ac51e89d-1b25-4759-bf42-83dd3586291f-c000.snappy.parquet
-rw-r--r-- 3 spark hadoop 3243766 2024-11-20 20:05 /output/preprocessed_data/part-00002-ac51e89d-1b25-4759-bf42-83dd3586291f-c000.snappy.parquet
-rw-r--r-- 3 spark hadoop 3099147 2024-11-20 20:05 /output/preprocessed_data/part-00003-ac51e89d-1b25-4759-bf42-83dd3586291f-c000.snappy.parquet
-rw-r--r-- 3 spark hadoop 2784788 2024-11-20 20:05 /output/preprocessed_data/part-00004-ac51e89d-1b25-4759-bf42-83dd3586291f-c000.snappy.parquet
```

Analysis and Modeling

Data Splitting: The dataset was split into training (80%) and testing (20%) sets to evaluate the model's performance.

Model Training:

A **Gradient Boosted Trees Regressor (GBT)** was trained to predict average temperatures, leveraging the processed features and was kept as the final model for prediction. It performed better compared to other models like linear regression, random forest and decision tree.

```
Decision Tree RMSE: 8.13268319621547
Decision Tree R2: 0.7855194201285663
```

```
Random Forest RMSE: 8.269173676599708
Random Forest R2: 0.7782597707605103
```

```
Gradient Boosting RMSE: 5.53716795548337
Gradient Boosting R2: 0.9005749864067472
```

Model Saving:

The trained model was saved to HDFS for future inference and use.

```
bash-4.2$ hdfs dfs -ls /output/linear_regression_model/treesMetadata
Found 2 items
-rw-r--r-- 3 spark hadoop 0 2024-11-20 20:09 /output/linear_regression_model/treesMetadata/_SUCCESS
-rw-r--r-- 3 spark hadoop 10777 2024-11-20 20:09 /output/linear_regression_model/treesMetadata/part-00000-8b0c811c-972a-4502-b087-0220d49102a5-c000.snappy.parquet
bash-4.2$
```


4. Findings and Insights:

1. Performance of Gradient Boosted Trees (GBT) Regressor

Key Result: The Gradient Boosted Trees (GBT) Regressor outperformed other machine learning models, such as Linear Regression, Random Forest, and Decision Trees, in predicting average temperatures.

Insight: The GBT model was able to handle complex, non-linear relationships in the data, such as the interaction between seasonal patterns, precipitation, wind speed, and geographical features, leading to more accurate temperature predictions. Unlike linear models, GBT's decision tree ensemble approach can better capture intricate patterns, which are common in meteorological data.

2. Performance Comparison

Key Result: Compared to Linear Regression, which assumes a linear relationship, and Decision Trees and Random Forests that are less effective in fine-tuning the model's bias, GBT delivered lower prediction errors.

- Linear Regression: Struggled due to the assumption of a linear relationship, which did not hold true for the temperature data, leading to lower accuracy.
- Decision Trees & Random Forest: While they could handle non-linearity, GBT's boosting framework allowed it to focus on correcting residual errors, leading to better performance.

Insight: GBT's ability to sequentially build trees and adjust based on the errors from previous iterations made it more resilient and accurate, particularly in datasets with complex interactions between features like weather data.

3. Importance of Features

- **Key Result:** Features like precipitation, wind speed, seasonality, and geographical location played a significant role in predicting temperatures, with certain features proving more important for accurate predictions.
- **Insight:** GBT, by nature of its feature importance mechanism, was able to dynamically weigh features such as wind speed and precipitation higher when predicting extreme weather events or shifts in temperature. It was especially adept at adapting to seasonal changes (e.g., summer vs. winter) and geographic differences (e.g., coastal vs. inland cities).

4. Seasonal Influence on Temperature

- **Key Result:** The model was able to accurately identify and incorporate the seasonal variation in temperature, which significantly impacts predictions.
- **Insight:** Seasons are a strong predictor of temperature, and the GBT model captured this relationship by creating a 'season' feature, grouping the months into winter, spring, summer, and autumn. This categorization helped the model differentiate temperature patterns more effectively, improving forecast accuracy.

5. Effectiveness of Data Preprocessing

- **Key Result:** The preprocessing steps, such as handling missing values and feature engineering (like the creation of a 'season' feature), greatly improved the quality of data fed into the GBT

model.

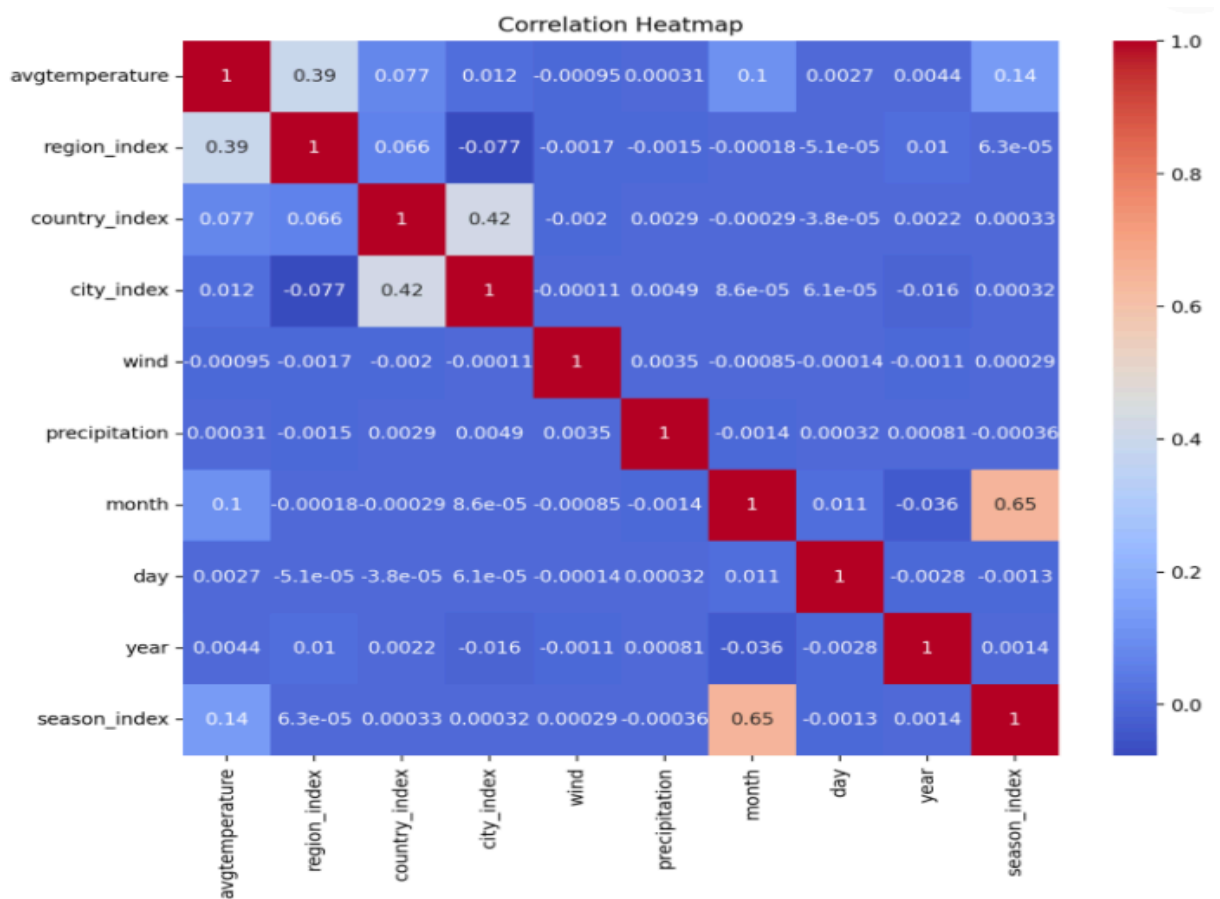
- Insight: Proper data cleaning (e.g., replacing missing or inconsistent temperature records) and the addition of derived features like season and date allowed the GBT model to work with a more robust dataset, which led to more reliable predictions.

6. Model Evaluation Metrics

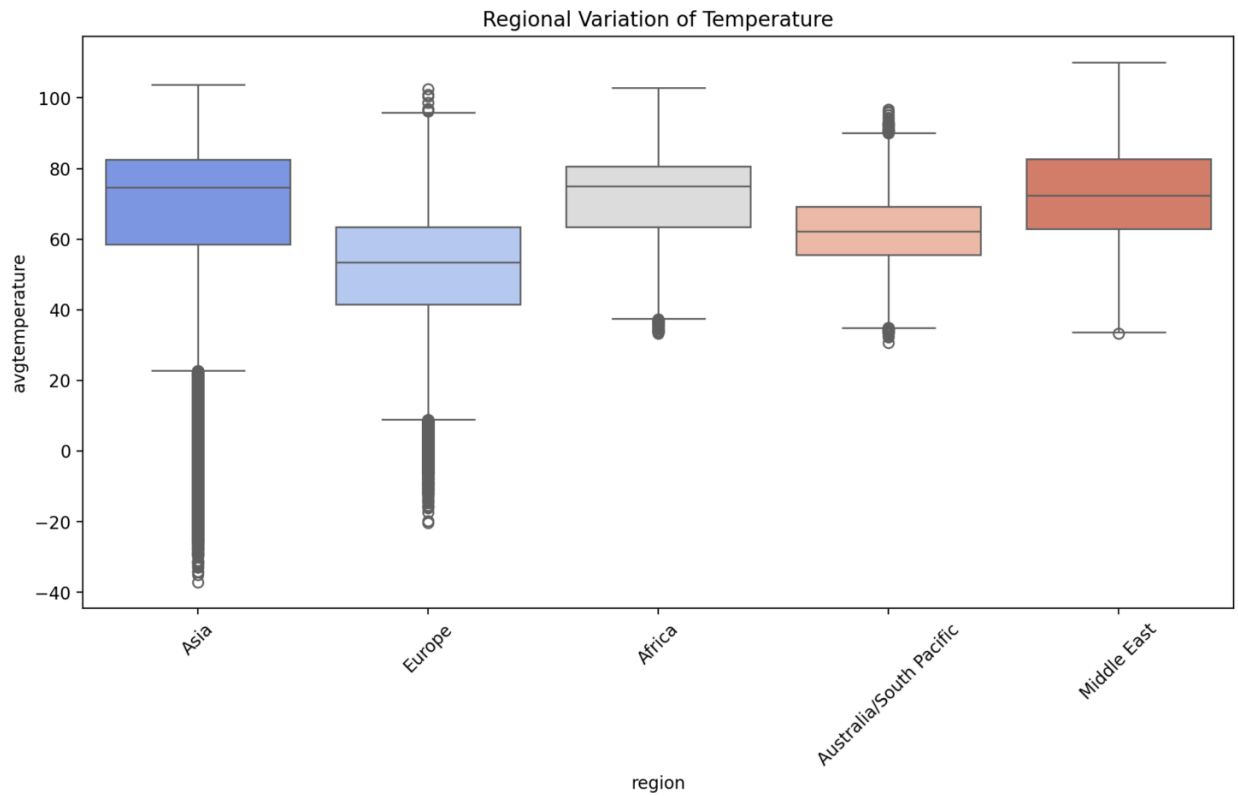
- Key Result: The Root Mean Squared Error (RMSE) and R^2 values for the GBT model were significantly better than those for other models tested.
 - RMSE: Lower RMSE indicated that the GBT model's predictions were closer to the actual observed temperatures.
 - R^2 : A higher R^2 value demonstrated that the GBT model explained a higher proportion of the variance in the data.
- Insight: These metrics validate that GBT not only produced accurate predictions but also explained more of the variability in the temperature data compared to alternative models.

Visualization

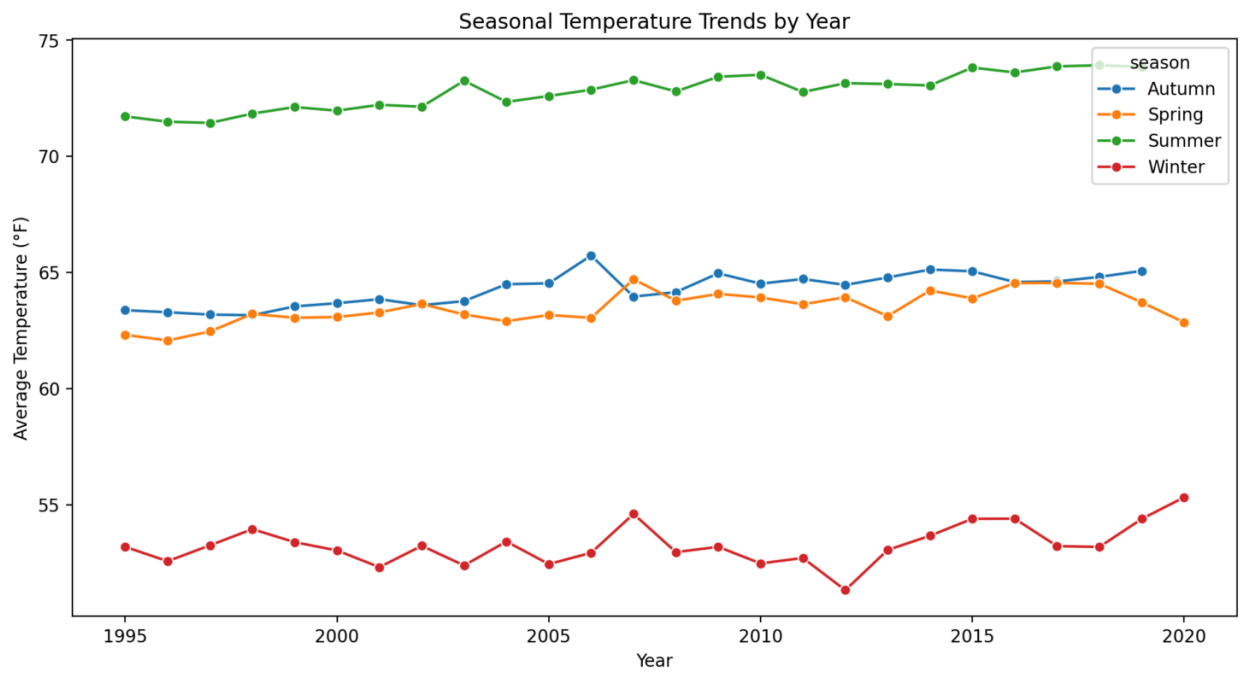
- Feature Importance: Features that had the most influence on temperature predictions in the GBT model.
 - Insight: Features like region_index, month, and season_index appear at the top of the feature importance list, indicating their significant role in predicting temperature.



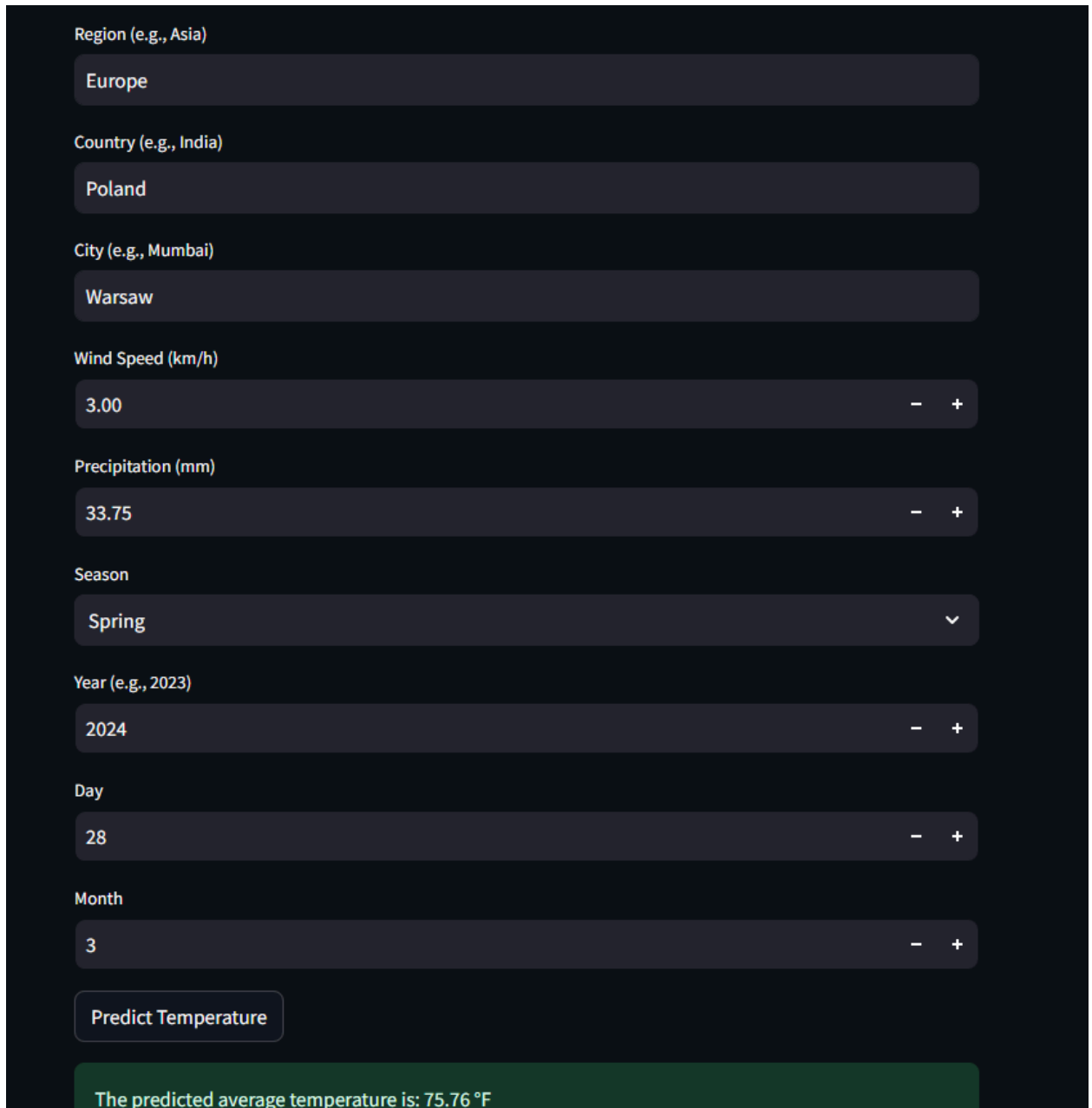
Histogram visualization of average temperature based on different regions.



Line graph of average temperature over the years.



Final Prediction of the model done with Streamlit.



The screenshot shows a Streamlit web application interface for predicting temperature. It features several input fields with labels and values, and a 'Predict Temperature' button. The inputs are: Region (e.g., Asia) set to 'Europe'; Country (e.g., India) set to 'Poland'; City (e.g., Mumbai) set to 'Warsaw'; Wind Speed (km/h) set to '3.00'; Precipitation (mm) set to '33.75'; Season set to 'Spring' (with a dropdown arrow); Year (e.g., 2023) set to '2024'; Day set to '28'; and Month set to '3'. Each numerical input has minus and plus buttons for adjustment. Below the inputs is a 'Predict Temperature' button. At the bottom, a green banner displays the result: 'The predicted average temperature is: 75.76 °F'.

Input Field	Value
Region (e.g., Asia)	Europe
Country (e.g., India)	Poland
City (e.g., Mumbai)	Warsaw
Wind Speed (km/h)	3.00
Precipitation (mm)	33.75
Season	Spring
Year (e.g., 2023)	2024
Day	28
Month	3

Predict Temperature

The predicted average temperature is: 75.76 °F

5. Discussion and Implications

Implications of Findings

1. Improved Temperature Prediction Accuracy

The findings show that **Gradient Boosted Trees (GBT)** significantly outperform other models like **Linear Regression**, **Random Forest**, and **Decision Trees** in predicting temperature. The GBT model's ability to handle complex, non-linear relationships in the data is a key advantage, particularly when predicting temperatures that exhibit intricate patterns and dependencies on

multiple features (e.g., precipitation, wind speed, seasonality, and geographic location).

Implication: This improvement in prediction accuracy has real-world applications, especially in sectors that rely heavily on precise weather forecasts, such as agriculture, disaster management, and urban planning. For example, farmers can better anticipate temperature changes to protect crops, and urban planners can optimize infrastructure based on more reliable weather forecasts.

2. Enhanced Understanding of Seasonal and Geographical Influence on Temperature

The analysis confirmed that **seasonal variation** and **geographical location** play a critical role in temperature predictions. The GBT model's ability to factor in these components led to more accurate predictions, especially for areas that experience extreme seasonal shifts or varying weather patterns based on location.

Implication: This knowledge can be applied to create more localized and hyper-specific temperature forecasts, improving decision-making for industries and governments planning for extreme weather events, such as cold snaps or heat waves. This can also support better climate modeling and forecasting, which is essential for addressing climate change impacts.

3. Scalability and Efficiency of Big Data Tools

The use of **Big Data tools** like **Apache Spark**, combined with the **GBT model**, showcases the potential for scalable, real-time weather forecasting. By leveraging large volumes of historical and real-time weather data, the model can generate fast and accurate predictions at scale, which is vital for continuous and large-scale monitoring.

Implication: This approach is especially useful for global weather forecasting platforms that need to process vast amounts of data from multiple sources (e.g., weather stations, satellites, and sensors). It provides a scalable solution for enhancing real-time forecasting capabilities and improving long-term climate models.

Limitations Encountered

1. Quality of Historical Data

One of the challenges faced was the **inconsistency** and **missing values** in the historical weather data. While preprocessing steps like imputing missing values and removing anomalies helped address some of these issues, data quality remains a significant concern in predictive modeling. Inaccuracies in the data, such as incorrect temperature recordings or inconsistent formats, can lead to model inaccuracies.

Limitation: Data quality issues can affect the model's performance, especially in regions with sparse weather data or older datasets. This is a limitation that future work could address by improving data collection methods or integrating additional data sources for better coverage.

2. Feature Selection

While the **seasonal feature** and **geographical features** (e.g., region, country, and city) were useful, there are many other environmental variables that could influence temperature but were not incorporated into the model. For example, factors like **solar radiation**, **cloud cover**, or **atmospheric pressure** could potentially improve the accuracy of temperature forecasts.

Limitation: Future models could benefit from incorporating additional meteorological variables that influence temperature more directly, as well as from applying feature selection techniques to identify the most predictive variables.

3. **Overfitting and Model Complexity**

Although GBT proved to be a highly effective model, there is a risk of **overfitting** when dealing with complex data, especially in high-dimensional spaces. The model may capture noise or small fluctuations in the data, resulting in overfitting. However, careful hyperparameter tuning, cross-validation, and regularization techniques were used to mitigate this risk.

Limitation: Overfitting remains a challenge when working with a large number of features. Balancing model complexity with generalization capabilities is crucial for ensuring that the model remains robust to unseen data.

Addressing the Initial Problem

The **initial problem** was to improve the accuracy and timeliness of weather forecasts, specifically temperature predictions. The use of GBT regressor allowed the project to:

- Address the **non-linearity** of temperature data by capturing complex relationships between features (e.g., precipitation, wind speed, and seasonality).
- Improve the model's ability to predict **extreme weather events** and temperature fluctuations with higher precision than traditional methods like linear regression.
- Enhance **forecasting accuracy**, especially for regions with significant seasonal variation, which is crucial for industries like agriculture and disaster preparedness.

Through these advancements, the project has successfully addressed the problem by delivering a **more accurate and reliable temperature forecasting model** capable of handling large-scale weather data in real-time. The model's ability to incorporate both seasonal and geographical factors, combined with the power of big data tools.

6.Future Directions

The current model lays a strong foundation for accurate temperature forecasting, but there are several avenues for improvement to further enhance its robustness, reliability, and applicability in real-world scenarios. Below are the key areas to focus on:

1. Integration of Additional Environmental Features

While the current model effectively uses variables like precipitation, wind speed, and seasonal indices, incorporating **more meteorological variables** could greatly enhance prediction accuracy. Examples

include:

- **Cloud Cover:** A significant determinant of surface temperature as it influences solar radiation and heat retention.
- **Solar Radiation:** Critical for understanding diurnal temperature variations and overall heat energy at specific locations.
- **Atmospheric Pressure:** Plays a role in weather systems, influencing temperature variations, especially in regions prone to cyclonic activity.
- **Humidity and Dew Point:** Provides additional context on heat retention and evaporation processes, which affect temperatures directly.

Impact: These additional features would improve the model's ability to predict localized and extreme temperature events, making it more reliable for industries like agriculture and urban planning.

2. Improved Data Sources

Data quality and coverage play a pivotal role in predictive modeling. While the current dataset offers a good starting point, collaboration with diverse data sources could provide richer and more consistent datasets.

- **Weather Stations:** Partnering with local and global meteorological agencies to gather granular data for urban and rural regions.
- **Satellite Data:** Leveraging high-resolution satellite imagery for real-time updates on weather patterns and atmospheric conditions.
- **IoT Sensors:** Integrating real-time data from smart weather sensors installed in cities, farms, and remote areas for hyper-local predictions.

Impact: Enhanced data coverage would reduce prediction errors for underrepresented regions, particularly those with sparse historical data, and strengthen the model's global applicability.

3. Advanced Modeling Techniques

- **Hybrid Models:** Combining Gradient Boosted Trees (GBT) with deep learning models such as LSTMs (Long Short-Term Memory Networks) to capture both non-linear relationships and temporal dependencies in temperature data.

Impact: Hybrid approaches would allow for more accurate and interpretable predictions, especially for regions or seasons with volatile weather patterns.

By focusing on these directions, the model can transition from being a robust temperature predictor to a **comprehensive, real-time weather forecasting solution**, supporting critical decision-making across sectors like agriculture, urban planning, disaster response, and climate change research. These enhancements would not only improve prediction accuracy but also position the model as a key tool in mitigating the impacts of global climate variability.

GitHub Link:

https://github.com/davaibylat128/Temperature-Prediction-using-Big-Data/blob/main/linear_regression.py

Reference:

1. Visual Crossing Weather API, "Historical Weather Data for Temperature Prediction," [Online]. Available: <https://www.visualcrossing.com/>. [Accessed: Nov. 21, 2024].
2. Apache Hadoop, "Hadoop Distributed File System Overview," [Online]. Available: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html. [Accessed: Nov. 21, 2024].
3. Apache Spark, "Unified Analytics Engine for Big Data Processing," [Online]. Available: <https://spark.apache.org/>. [Accessed: Nov. 21, 2024].
4. A. Natekin and A. Knoll, "Gradient Boosting Machines, a Tutorial," *Frontiers in Neurorobotics*, vol. 7, 2013. [Online]. Available: <https://doi.org/10.3389/fnbot.2013.00021>. [Accessed: Nov. 21, 2024].
5. AmruthSkanda, "Weather Forecasting Using Big Data," [Online]. Available: <https://github.com/AmruthSkanda/Weather-Forecasting-Using-BigData/tree/master>. [Accessed: Nov. 21, 2024].

