INSTITUTE FOR ADVANCED COMPUTING

AND SOFTWARE DEVELOPMENT AKURDI,

PUNE

Documentation On

“Weather Data Analysis And Forecasting”

PG-DBDA SEPT 2023

Submitted By:

Group No: 11

Roll No.	**Name**
239533	**Prajwal Ambadkar**
239535	**Praveer Pratap**

Mrs. Priti Take
Project Guide

Mr. Rohit Puranik
Centre Coordinator

ABSTRACT

The "Weather Data Analysis and Forecasting" project aims to leverage advanced data analytics techniques to analyze historical weather data and develop an accurate forecasting model for predicting weather conditions in the upcoming days. Weather forecasting plays a pivotal role in various sectors, including agriculture, transportation, and disaster management. The project focuses on harnessing the power of machine learning and statistical methods to enhance the precision and reliability of weather predictions.

The project begins with the collection and preprocessing of extensive historical weather data, encompassing variables such as temperature, humidity, wind speed, and atmospheric pressure. Exploratory data analysis is conducted to uncover patterns, trends, and correlations within the dataset. Feature engineering is employed to extract relevant information, and missing data is addressed through appropriate imputation techniques.

A predictive model is then developed using state-of-the-art machine learning algorithms, including but not limited to time series analysis, regression, and ensemble methods. The model is trained on historical data to learn complex relationships and patterns, enabling it to make accurate predictions for future time points. Model evaluation is conducted using various metrics, ensuring its effectiveness in capturing the nuances of different weather conditions.

The forecasting system is integrated into a user-friendly interface, allowing end-users to access real-time weather predictions for specific locations. Additionally, the project incorporates visualization tools to present forecasted trends and anomalies, aiding in the interpretation of results.

The "Weather Data Analysis and Forecasting" project not only contributes to the advancement of meteorological science but also addresses the practical needs of diverse industries. The accuracy of weather forecasts can lead to informed decision-making, improved resource allocation, and better preparedness for weather-related events. The project's outcomes are expected to have a positive impact on sectors dependent on weather conditions, ultimately enhancing overall societal resilience in the face of dynamic weather patterns.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to everyone who has contributed to the completion of our project.

First and foremost, we would like to thank our project guide **Mrs. Priti Take** Mam for their constant guidance and support throughout the project. We extend our sincere thanks to our respected centre coordinator, **Mr. Rohit Puranik** for allowing us to use the facilities available.

We would also like to express our appreciation to the faculty members of our department for their constructive feedback and encouragement. Their insights and suggestions have helped us to refine our ideas and announce the quality of our work.

Furthermore, we would like to thank our families and friends for their unwavering support and encouragement throughout our academic journey, their love and support have been a constant source of motivation and inspiration for us.

Thank you for all your valuable contributions to our project.

Prajwal Ambadkar(239533)

Praveer Pratap (239535)

Table of Contents

1. INTRODUCTION

1.1 Problem Statement

Weather Data Analysis And Forecasting

1.2 Product Scope

1. Data Collection and Integration:
   - Gather and integrate diverse data sources, including customer profiles, credit histories, income details, employment stability, and macroeconomic indicators.
   - Implement mechanisms for real-time data updates to ensure model accuracy.
2. Predictive Modelling:
   - Develop machine learning algorithms to predict the likelihood of data.
   - Utilize various models such as elasticnet , lasso , ridge to achieve accurate predictions.
   - Consider ensemble methods for combining the strengths of multiple models.
3. Feature Selection and Engineering:
   - Identify and prioritize key features that significantly impact forecatsing decisions.
   - Implement feature engineering techniques to enhance model performance.
4. Model Interpretability and Explainability:
   - Ensure transparency in model predictions to build trust among users.
   - Implement methods for explaining model decisions to assist in understanding the forecasting.
   - Conduct rigorous validation processes using cross-validation techniques.

1.3 Aim & Objectives

The aim for weather analysis and forecasting is to leverage advanced analytics and machine learning techniques to enhance the efficiency and accuracy of the lending process within forecasting The primary objectives include:

1. Optimizing Forecasting Decisions:
   - Develop predictive models to assess the creditworthiness of forecasting.
   - Improve the accuracy of loan approval decisions by leveraging historical data and advanced analytics.

2. Reducing Risk and Enhancing forecasting Quality:
   - Identify and evaluate risk factors associated with weather.
   - Implement predictive models to proactively identify potential defaults or high-risk applicants, thereby improving the overall quality of the forecasting..

3. Increasing Operational Efficiency:
   - Streamline the loan approval process by automating and optimizing decision-making using forecasting models.
   - Reduce manual effort and processing time, leading to quicker and more efficient forecasting.

4. Enhancing User Experience:
   - Provide a faster and more responsive forecasting process, improving the overall experience for user.
   - Implement fair and transparent lending practices to build trust and loyalty among users.

By achieving these aims, and with the help of visualization, weather forecasting can create a more efficient, user-centric, and risk-aware lending environment, ultimately contributing to the growth and sustainability of forecast.

2. OVERALL DESCRIPTION

2.1 Workflow of Project

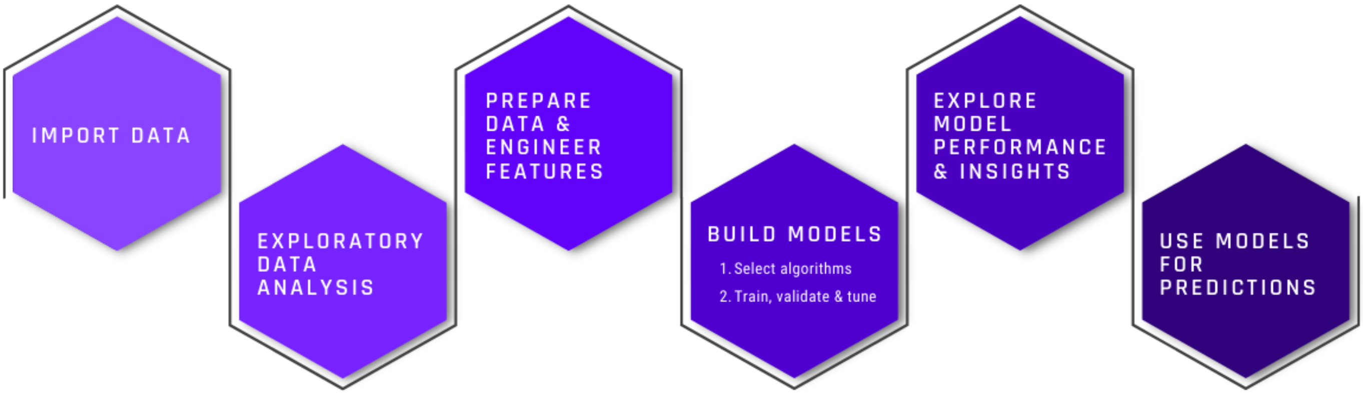

IMPORT DATA
EXPLORATORY DATA ANALYSIS
PREPARE DATA & ENGINEER FEATURES
BUILD MODELS
1. Select algorithms
2. Train, validate & tune
EXPLORE MODEL PERFORMANCE & INSIGHTS
USE MODELS FOR PREDICTIONS

Figure 1 Workflow of Project

2.2 Data Description

Weather dataset contain hourly weather data from different weather stations from different region in Brazil.

2.3 Importing Dataset.

1. Here we can see that we have categorical and continuous variables, we can also **Amount of Precipitation (mm) (last hour):**
   - Measurement of the amount of precipitation (rainfall) in millimeters recorded in the last hour.
2. **Atmospheric Pressure at Station Level (mb):**
   - Atmospheric pressure at the station level measured in millibars (mb).
3. **Maximum Air Pressure for the Last Hour (mb):**
   - The maximum air pressure recorded in the last hour, measured in millibars.
4. **Minimum Air Pressure for the Last Hour (mb):**
   - The minimum air pressure recorded in the last hour, measured in millibars.
5. **Solar Radiation (KJ/m2):**
   - Solar radiation measured in kilojoules per square meter (KJ/m2).

6. **Air Temperature (Instant) (°C):**
    - Instantaneous air temperature measured in degrees Celsius (°C).
7. **Dew Point Temperature (Instant) (°C):**
    - Instantaneous dew point temperature measured in degrees Celsius (°C).
8. **Maximum Temperature for the Last Hour (°C):**
    - The maximum air temperature recorded in the last hour, measured in degrees Celsius.
9. **Minimum Temperature for the Last Hour (°C):**
    - The minimum air temperature recorded in the last hour, measured in degrees Celsius.
10. **Maximum Dew Point Temperature for the Last Hour (°C):**
    - The maximum dew point temperature recorded in the last hour, measured in degrees Celsius.
11. **Minimum Dew Point Temperature for the Last Hour (°C):**
    - The minimum dew point temperature recorded in the last hour, measured in degrees Celsius.
12. **Maximum Relative Humid Temperature for the Last Hour (%):**
    - The maximum relative humidity recorded in the last hour, expressed as a percentage.
13. **Minimum Relative Humid Temperature for the Last Hour (%):**
    - The minimum relative humidity recorded in the last hour, expressed as a percentage.
14. **Relative Humidity (% Instant):**
    - Instantaneous relative humidity measured as a percentage.
15. **Wind Direction (Radius Degrees 0-360):**
    - Wind direction represented in degrees on a 0-360 scale.
16. **Wind Gust (m/s):**
    - Wind gust speed measured in meters per second.
17. **Wind Speed (m/s):**
    - Wind speed measured in meters per second.
18. **Brazilian Geopolitical Regions:**
    - The geopolitical region in Brazil where the weather station is located.
19. **State (Province):**
    - The state or province in Brazil where the weather station is situated.
20. **Station Name:**
    - The name or nickname of the weather station, usually indicating the city or specific location.

21. **Station Code (INMET Number):**
    - The unique identification code assigned to the weather station by the Brazilian National Institute of Meteorology (INMET).
22. **Latitude:**
    - The geographic latitude of the weather station.
23. **Longitude:**
    - The geographic longitude of the weather station.
24. **Elevation:**
    - The elevation or altitude of the weather station above sea level

2.4 Project Initiation:

- Define the project scope, objectives, and deliverables.
- Set up project management tools and resources.

2.5 Data Acquisition:

- Obtain the weather dataset from the appropriate source, ensuring compliance with data privacy and security regulations.
- Understand the structure and format of the dataset, including the features, target variable, and any data dictionaries or documentation provided.

2.6 Data Exploration and Understanding:

- Perform initial data exploration to gain insights into the dataset's characteristics, including the distribution of variables, missing values, outliers, and potential data quality issues.
- Visualize key features using tableau.

2.7 Data Pre-processing:

Weather Forecasting project based on temperature in the North region of Brazil, specifically focusing on the MANAUS weather station. Here's a summary of the key steps you've taken:

1. Data Extraction:

- Selected relevant columns: Date, Time, Total Precipitation, Atmospheric Pressure, Solar Radiation, Maximum Temperature for the Last Hour, Minimum Temperature for the Last Hour, Maximum Relative Humidity for the Last Hour, Minimum Relative Humidity for the Last Hour, Wind Speed in Metres per Second, Region, State, Station Name, Station Code, Latitude, Longitude, and Elevation.
- Focused on the North region of Brazil and the MANAUS weather station.

2. Temporal Consideration:

- Restricted the dataset to hourly data from 2000 to 2021.
- Maintained hourly timestamps for a detailed analysis.

3. Data Cleaning in PySpark:

- Filtered out rows with garbage values, ensuring data quality.
- Applied range-based filtering to remove irrelevant or erroneous data points.

4. Time-Based Filtering:

- Retained days in the dataset where the hour is greater than 5, potentially excluding nighttime data to focus on daytime observations.

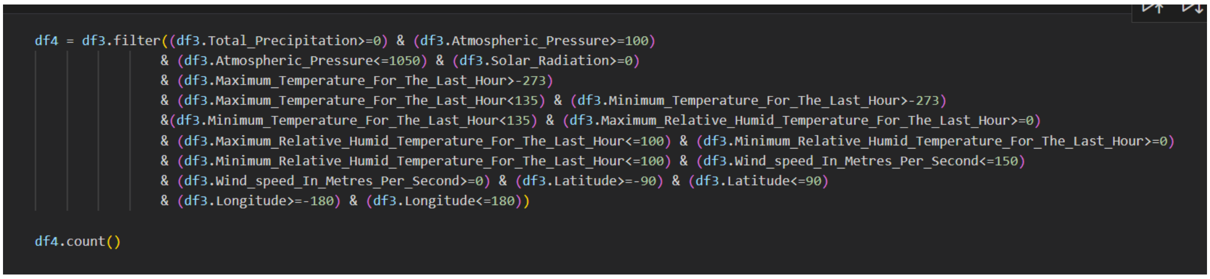
Next Steps:

- **Exploratory Data Analysis (EDA):** Conduct a thorough exploration of the cleaned dataset. Examine statistical summaries, distributions, and visualizations to gain insights into the temporal and spatial patterns of the selected weather variables.
- **Feature Engineering:** Consider deriving additional features that might enhance the forecasting model. For example, you could extract temporal features, such as day of the week or month, to capture seasonality effects.

- **Model Development:** Proceed with the development of the weather forecasting model based on temperature. Choose appropriate machine learning algorithms, considering time series characteristics, and split the data into training and testing sets.
- **Model Training and Evaluation:** Train the model on the training set and evaluate its performance on the testing set. Utilize metrics like Mean Squared Error (MSE) or Root Mean Squared Error (RMSE) to assess accuracy.
- **Hyperparameter Tuning:** Fine-tune model hyperparameters to optimize performance.
- **Validation:** Validate the model's predictions against actual weather observations to ensure its reliability and generalizability.
- **Deployment:** If the model performs satisfactorily, consider deploying it for real-time or future weather forecasting.

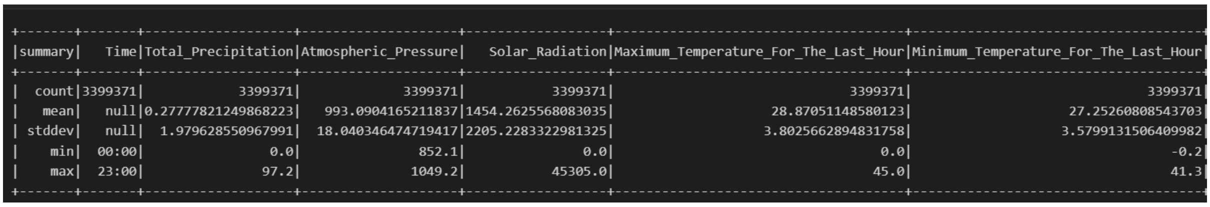
2.8 Feature Engineering:

- Generate new features that may capture additional information relevant to forecasting like lag columns to history indicators.
- Select or create relevant features based on domain knowledge and exploratory data analysis insights.

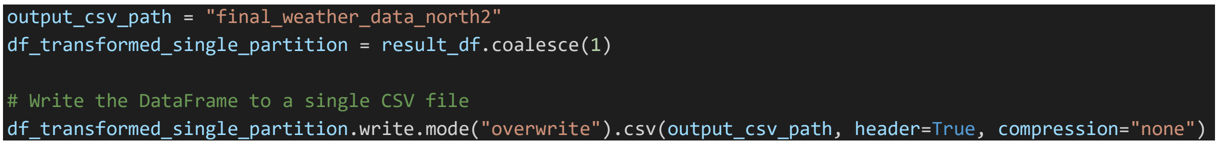


```
df4 = df3.filter((df3.Total_Precipitation>=0) & (df3.Atmospheric_Pressure>=100)
                & (df3.Atmospheric_Pressure<=1050) & (df3.Solar_Radiation>=0)
                & (df3.Maximum_Temperature_For_The_Last_Hour>-273)
                & (df3.Maximum_Temperature_For_The_Last_Hour<135) & (df3.Minimum_Temperature_For_The_Last_Hour>-273)
                &(df3.Minimum_Temperature_For_The_Last_Hour<135) & (df3.Maximum_Relative_Humid_Temperature_For_The_Last_Hour>=0)
                & (df3.Maximum_Relative_Humid_Temperature_For_The_Last_Hour<=100) & (df3.Minimum_Relative_Humid_Temperature_For_The_Last_Hour>=0)
                & (df3.Minimum_Relative_Humid_Temperature_For_The_Last_Hour<=100) & (df3.Wind_speed_In_Metres_Per_Second<=150)
                & (df3.Wind_speed_In_Metres_Per_Second>=0) & (df3.Latitude>=-90) & (df3.Latitude<=90)
                & (df3.Longitude>=-180) & (df3.Longitude<=180))

df4.count()
```

Here are some rows after the above transformation.



```
+-------+-------+-------------------+--------------------+------------------+-------------------------------------+-------------------------------------+
|summary|   Time|Total_Precipitation|Atmospheric_Pressure|   Solar_Radiation|Maximum_Temperature_For_The_Last_Hour|Minimum_Temperature_For_The_Last_Hour|
+-------+-------+-------------------+--------------------+------------------+-------------------------------------+-------------------------------------+
|  count|3399371|            3399371|             3399371|           3399371|                              3399371|                              3399371|
|   mean|   null|0.27777821249868223|   993.0904165211837|1454.2625568083035|                    28.87051148580123|                    27.25260808543703|
| stddev|   null| 1.979628550967991|  18.040346474719417|2205.2283322981325|                   3.8025662894831758|                   3.5799131506409982|
|    min|  00:00|                0.0|               852.1|               0.0|                                  0.0|                                 -0.2|
|    max|  23:00|               97.2|              1049.2|           45305.0|                                 45.0|                                 41.3|
+-------+-------+-------------------+--------------------+------------------+-------------------------------------+-------------------------------------+
```

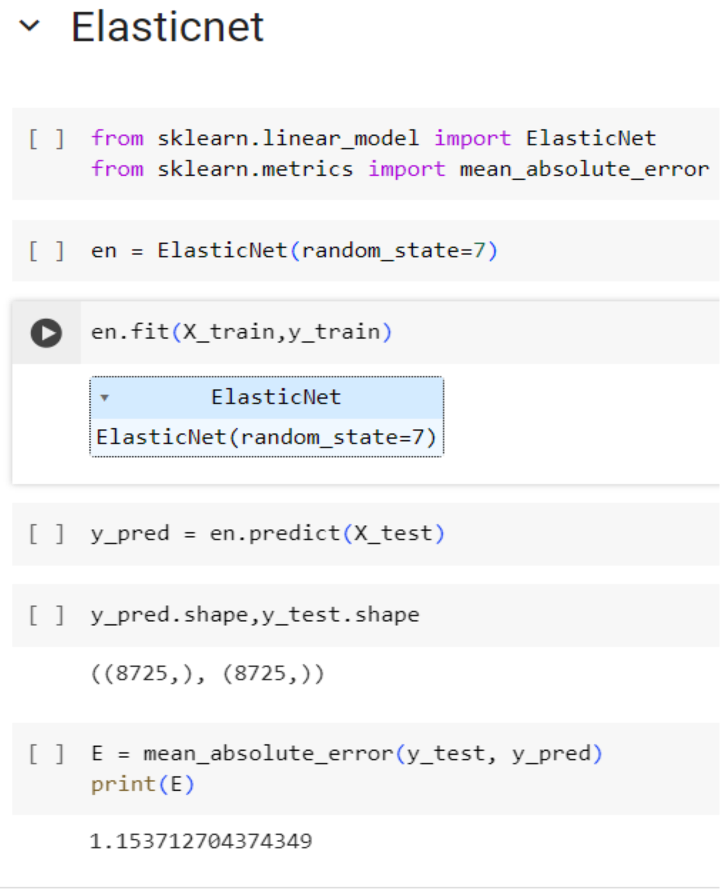


```
output_csv_path = "final_weather_data_north2"
df_transformed_single_partition = result_df.coalesce(1)

# Write the DataFrame to a single CSV file
df_transformed_single_partition.write.mode("overwrite").csv(output_csv_path, header=True, compression="none")
```

with the help of above code we downloaded our transformed file so that we can use for visualization in tableau.

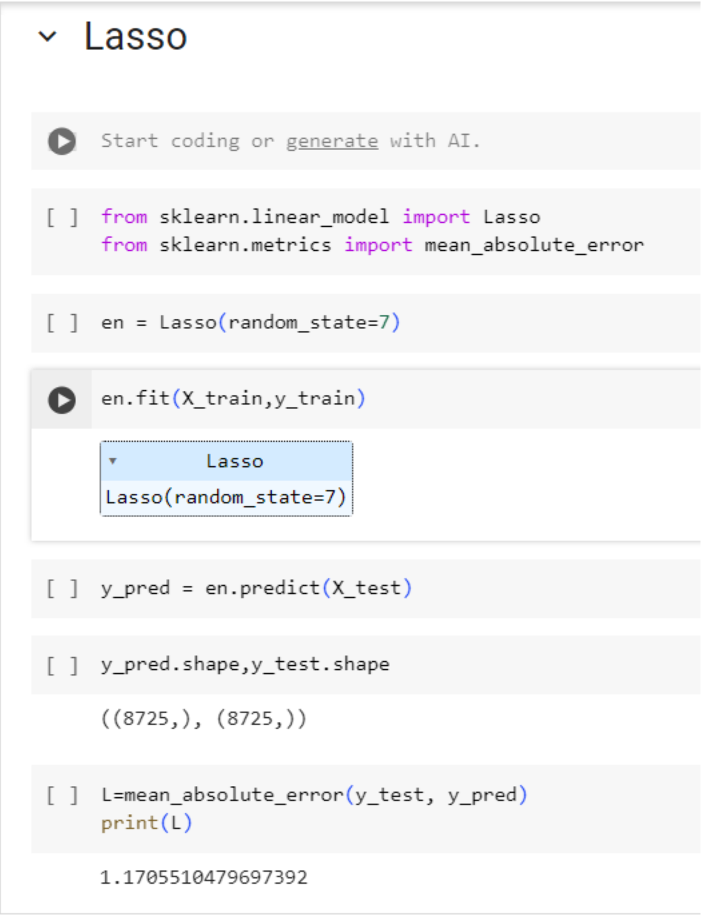
2.9 Model Development:

- Split the dataset into training, validation, and test sets to evaluate model performance effectively.
- Select appropriate machine learning algorithms for the predictive task, such as elasticnet , lasso , ridge.



Elasticnet

```
[ ] from sklearn.linear_model import ElasticNet
    from sklearn.metrics import mean_absolute_error

[ ] en = ElasticNet(random_state=7)

    en.fit(X_train,y_train)

          ElasticNet
    ElasticNet(random_state=7)

[ ] y_pred = en.predict(X_test)

[ ] y_pred.shape,y_test.shape

    ((8725,), (8725,))

[ ] E = mean_absolute_error(y_test, y_pred)
    print(E)

    1.153712704374349
```



Lasso

```
[ ] from sklearn.linear_model import Lasso
    from sklearn.metrics import mean_absolute_error
```

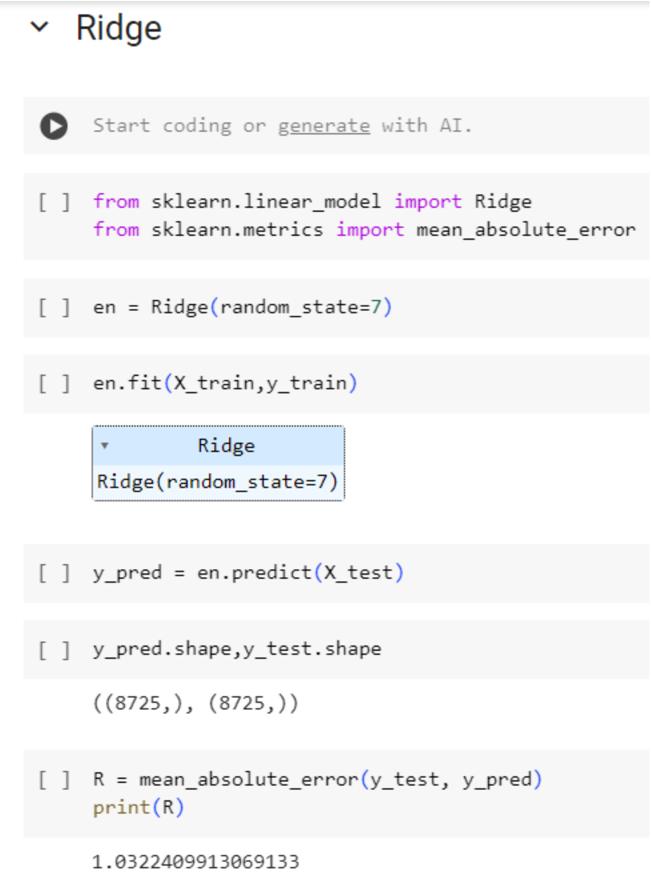
```
[ ] en = Lasso(random_state=7)
```

```
en.fit(X_train,y_train)

    Lasso
Lasso(random_state=7)
```

```
[ ] y_pred = en.predict(X_test)
```

```
[ ] y_pred.shape,y_test.shape

((8725,), (8725,))
```

```
[ ] L=mean_absolute_error(y_test, y_pred)
    print(L)

1.1705510479697392
```



Ridge

```
[ ] from sklearn.linear_model import Ridge
    from sklearn.metrics import mean_absolute_error
```

```
[ ] en = Ridge(random_state=7)
```

```
[ ] en.fit(X_train,y_train)

    Ridge
Ridge(random_state=7)
```

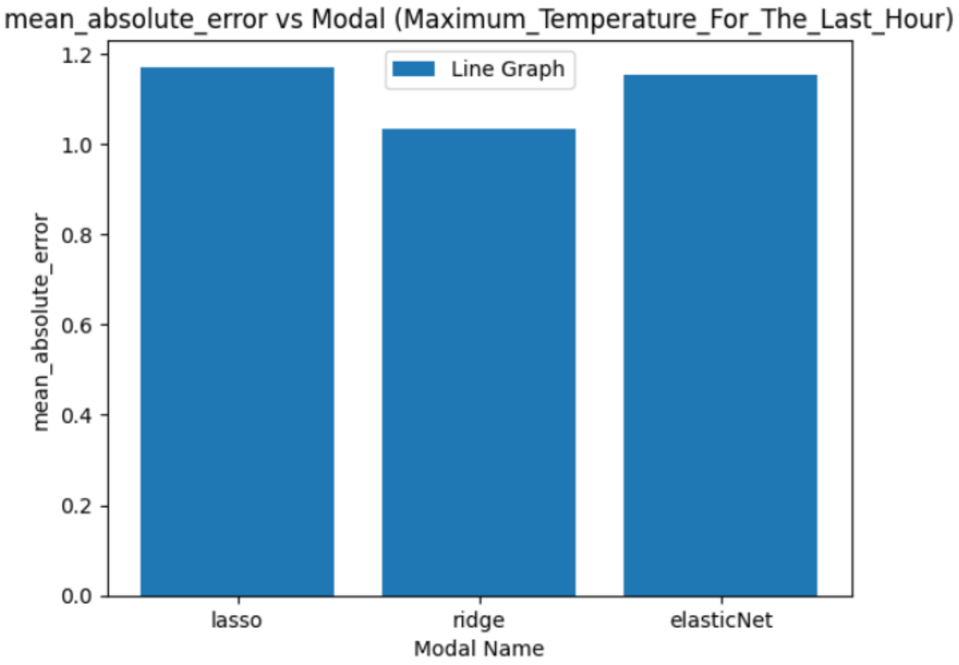
```
[ ] y_pred = en.predict(X_test)
```

```
[ ] y_pred.shape,y_test.shape

((8725,), (8725,))
```

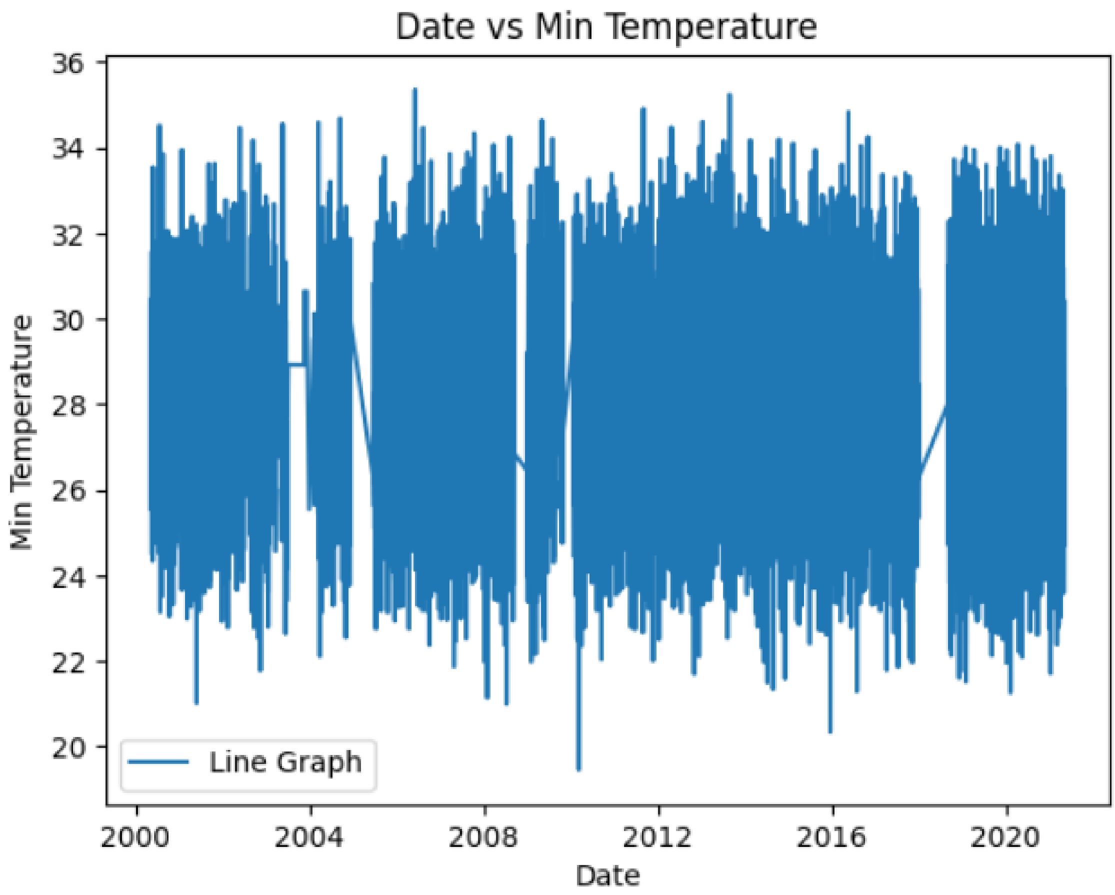
```
[ ] R = mean_absolute_error(y_test, y_pred)
    print(R)

1.0322409913069133
```

2.10 Model Evaluation:

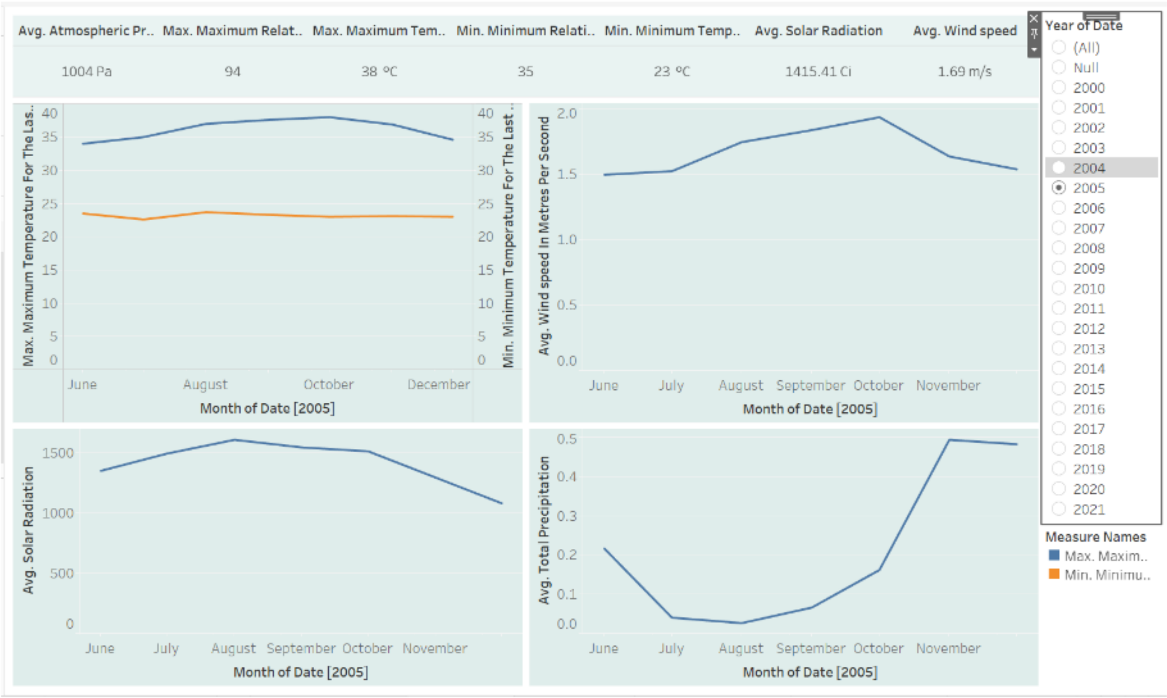
- Compare the performance of different models and select the best-performing one based on predefined criteria.
-



Below is the final forecasting of temperature. .



Here is one dashboard that we have created.



Avg. Atmospheric Pr..	Max. Maximum Relat..	Max. Maximum Tem..	Min. Minimum Relati..	Min. Minimum Temp..	Avg. Solar Radiation	Avg. Wind speed
1004 Pa	94	38 ºC	35	23 ºC	1415.41 Ci	1.69 m/s

Year of Date

○ (All)
○ Null
○ 2000
○ 2001
○ 2002
○ 2003
○ 2004
● 2005
○ 2006
○ 2007
○ 2008
○ 2009
○ 2010
○ 2011
○ 2012
○ 2013
○ 2014
○ 2015
○ 2016
○ 2017
○ 2018
○ 2019
○ 2020
○ 2021

Measure Names

🟦 Max. Maxim..
🟧 Min. Minimu..

2.11 Model Interpretation:

- Interpret model predictions and feature importance to understand the factors driving accurate forecasting .
- Communicate findings to user in a clear and interpretable manner, highlighting actionable insights and recommendations.

2.12 Model Deployment:

- Deploy the selected model into production environments, integrating it into the forecasting process to assist in decision-making.
- Implement monitoring mechanisms to track model performance and recalibrate the model as needed over time.

2.13 Documentation and Reporting:

- Document the entire project workflow, including data pre-processing steps, model development process, evaluation results, and deployment procedures.
- Prepare a comprehensive report summarizing the project findings, insights.

3. FUTURE SCOPE

The future scope of "Weather Data Analysis and Forecasting" using ML algorithms and Tableau for visualization is promising, with opportunities for advancements in technology, research, and user-centric applications. Here are some potential directions for future development:

1. **Improved Machine Learning Models:**
   - Continued research and development of advanced machine learning models, including ensemble methods, deep learning architectures, and hybrid models. These models can better capture intricate relationships within weather data and improve forecasting accuracy.
2. **Hyperparameter Optimization and AutoML:**
   - Emphasis on automated machine learning (AutoML) techniques for hyperparameter optimization, model selection, and feature engineering. This can streamline the model development process and lead to more effective forecasting models.
3. **Probabilistic Forecasting with ML:**
   - Advancements in probabilistic forecasting using machine learning algorithms. Enhancing models to provide not only deterministic predictions but also quantify uncertainties, enabling more informed decision-making.
4. **Integration of Big Data and Cloud Computing:**
   - Utilization of big data technologies and cloud computing platforms for handling large volumes of weather data. This can improve scalability, storage, and processing capabilities for enhanced model training and prediction.
5. **Real-Time Data Streaming:**
   - Integration of real-time data streaming technologies to enable the ingestion of live weather data into ML models. This allows for continuous updates and adaptability to changing weather conditions.
6. **Explainable AI in Weather Forecasting:**
   - Incorporation of explainable AI techniques to enhance the interpretability of ML-based forecasting models. Ensuring transparency in model predictions is crucial for building trust among users and stakeholders.
7. **Customizable Visualization Dashboards:**
   - Development of customizable and interactive visualization dashboards using tools like Tableau. These dashboards can

empower users to explore weather data, trends, and forecasts based on their specific needs.

8. **Geospatial Analytics:**
    - Integration of geospatial analytics into visualization tools for a more detailed and location-specific representation of weather patterns. This is particularly important for industries with geographically distributed operations.
9. **User-Centric Applications with Tableau:**
    - Creation of user-centric weather forecasting applications with intuitive interfaces using Tableau. Tailoring visualizations to specific industries (e.g., agriculture, energy, aviation) allows users to make informed decisions based on forecasted weather conditions.
10. **Predictive Analytics for Extreme Events:**
    - Implementation of predictive analytics to identify and forecast extreme weather events, such as hurricanes, floods, or heatwaves. This can contribute to improved disaster preparedness and response.
11. **Collaboration with Meteorological Agencies:**
    - Strengthening collaborations with meteorological agencies for data sharing, model validation, and benchmarking. Such partnerships can lead to the development of more accurate and robust forecasting systems.
12. **Machine Learning for Climate Change Impact Assessment:**
    - Extending ML applications to assess the impact of climate change on weather patterns. This involves analyzing long-term trends, identifying climate change indicators, and predicting future climate scenarios.
13. **Enhanced Accessibility and Mobile Applications:**
    - Improving accessibility to weather forecasts through the development of user-friendly mobile applications. Providing timely and localized weather information for individuals on the go.

As technology and methodologies advance, the convergence of ML algorithms and visualization tools like Tableau holds tremendous potential for revolutionizing how we analyze, understand, and forecast weather patterns. The future is likely to see a more seamless and user-friendly integration of these technologies for enhanced decision-making in various sectors.

REQUIREMENTS SPECIFICATION

4.1 Hardware Requirement:

- 500 GB hard drive (Minimum requirement)
- 8 GB RAM (Minimum requirement)
- PC x64-bit CPU

4.2 Software Requirement:

- Windows/Mac/Linux
- Apache Spark 2.4.x and above
- VS Code/Anaconda/Spyder
- Python Extension for VS Code

Libraries:

- Numpy 1.18.2
- Pandas 1.2.1
- Matplotlib 3.3.3
- Scikit-learn 0.24.1

4. CONCLUSION

In the culmination of our project on "Weather Data Analysis and Forecasting" using PySpark, machine learning algorithms, and Tableau for visualization, we have successfully traversed the realms of big data processing, advanced analytics, and impactful visualization to glean valuable insights from extensive weather datasets. This multifaceted approach has empowered us to not only understand historical weather patterns but also to make informed predictions for future conditions.

Leveraging the robust capabilities of PySpark, we efficiently processed large volumes of weather data, ensuring data quality through meticulous cleaning and filtering.

The application of machine learning algorithms enabled us to uncover intricate relationships within the weather data.

Tableau emerged as a powerful ally in transforming complex data into intuitive visualizations. Our Tableau dashboards offer dynamic and interactive insights, enabling users to explore weather patterns, trends, and forecasts effortlessly.

Looking forward, the project opens avenues for continuous improvement and expansion, **Advanced Machine Learning Models, Real-Time Data Streaming, Collaboration and Integration, Climate Change Impact Assessment.**

5. REFERENCES

- https://www.kaggle.com/datasets/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region?select=columns_description.csv
- https://gemini.google.com
- https://www.kaggle.com/code/nibukdk93/weather-data-analysis