A Report On

Machine Learning–Driven Temperature Forecasting Using NOAA Weekly Weather Data

Submitted by

Ch.Sowmya Sree

AP23110010053

CSE-R

Bachelor of Technology

in

Computer Science and Engineering

School Engineering and Sciences

Under the guidance of

Dr.Anusha Nalajala

Department of Computer Science

[November , 2025]

SRM University AP, Neerukonda , Mangalagiri, Guntur
Andhra Pradesh- 522240

Table of Contents

Abstract

Weather prediction is a key factor in modern society, and it affects such industries as agriculture, airlines, energy management, and the like. The main drawback of the scenario is that accurate temperature forecasting is a requisite for the development and survival of such industries, so the classical numerical weather prediction methodology which is based on physical simulations is still the dominant one.

The artificial intelligence-based technique is the one used in this project to predict the average weekly temperatures the CORGIS Weather Dataset, which is a product of the National Oceanic and Atmospheric Administration (NOAA). The Weather dataset consists of weekly weather reports from 2016 which include the following parameters: precipitation, wind speed, wind direction, and temperature.

The work involves systematic data preprocessing steps such as cleaning, normalization, feature extraction, and categorical variables handling. Afterwards a Random Forest Regression model is applied to estimate the average weekly temperature on the basis of other meteorological parameters. The algorithm was chosen due to its high robustness, efficiency, and ability to handle complex non-linear relationships in environmental data.

The study proves that machine learning can be a trustworthy and data-driven substitute for traditional meteorological models by providing quicker, more adaptable, and more interpretable predictions. Moreover, the project sets up a machine learning integration framework with environmental datasets aiming at the improvement of the accuracy and efficiency of the climate and temperature forecasting systems.

Keywords:

Weather Forecasting · Random Forest Regression · Machine Learning · NOAA · CORGIS Dataset · Temperature Prediction · Climate Modeling

Introduction

The goal of weather prediction is now the most important scientific and technological goal of the present time. High precision in forecasting temperature, rainfall, wind, and other weather factors is very important not only to meteorology but also to agriculture, energy, aviation, transport, and public safety industries. Out of all the climatic events, temperature is the one that really gives a good measure of the global warming and the processes going on in the atmosphere.

The CORGIS Weather Dataset is the main source of data in this project. The dataset, which consists of weekly summaries of weather conditions for the year 2016, was put together by Austin Cory Bart and Ryan Whitcomb as part of the CORGIS Dataset Project. It uses the data from the National Oceanic and Atmospheric Administration (NOAA). Among the different weather parameters that are included in the dataset are:

- Precipitation (in inches) – average rainfall measured throughout the week,
- Temperature statistics – including average, maximum, and minimum temperatures,
- Wind speed and wind direction
- Geographical locations – such as the city, state, and station code of the reporting office.

The primary goal of the project is to develop a machine-learning model that can accurately predict the average weekly temperature as a result of the feature set selection. The Random Forest Regression algorithm was primarily chosen for this purpose because of its capability to produce very precise results, its resistance to overfitting, and its exceptional proficiency in revealing the interactions among features regarding the complexity of the interaction.

The strategy of the project consists of several steps that are major:

- Cleaning and Preprocessing of Data: The principal activities of this pipeline are the elimination of duplicates, treatment of missing values, standardization of the numerical feature values, and conversion of the date fields.
- Feature Engineering: This step helps to give better performance to the model by providing the relevant features to be used in the model from date and weather-related fields.
- Modeling: The Random Forest Regressor is trained and tuned to the maximum temperature prediction.
- Evaluation and Interpretation: R^2, MAE, and RMSE serve as statistical metrics for model evaluation; additionally, the factors with the greatest influence on temperature are detected.

Problem Statement

Predicting temperatures accurately is vital for the overall functioning and safety of sectors such as agriculture, energy, and transportation.

The objective of this project is to create a data-driven model that will be able to forecast weekly average temperature by utilizing the CORGIS Weather Dataset (the dataset is a product of NOAA reports). The central issue here is going to be the handling of missing observations, the mixture of different types of features, and the intricate dependencies that exist among the variables.

The research is concentrated on the construction of a model that will be both precise and fast, immutable and that will display the potential of ML in overcoming the traditional ways of weather forecasting.

Objectives

The machine learning model that will predict the average weekly temperature through historical weather data from the CORGIS Data Set Project (NOAA-based data), the main purpose of the research, is going to be made and constructed firstly.

The table below shows the detailed objectives of the research:

S.No	Objective
1	The CORGIS Weather dataset will be pre-processed and cleaned to provide a solid basis for accurate analysis.
2	The weather factors that have the greatest impact on temperature will be identified and selected.
3	Temperature predictions will be based on the Random Forest Regression algorithm.
4	The performance of the model will be measured by R^2, MAE, and RMSE along with other metrics.
5	Temperature fluctuations will point out the main factors responsible for the changes.
6	The integration of machine learning with traditional forecasting methods will be shown to be a significant improvement.

Dataset Description

This project utilizes the Weather CSV file dataset from the CORGIS Dataset Project, which was created by Austin Cory Bart and Ryan Whitcomb (Version 2.0.0, June 13, 2016).
The dataset comes from reports of the National Oceanic and Atmospheric Administration (NOAA), and it delivers weekly summaries of weather data for the entire year of 2016 covering various U.S. cities.

The data integrates the reports from 122 Weather Forecast Offices (WFOs) taking care of the collection and dissemination of local weather data.
One record accounts for a week of weather data for one particular city that includes temperature, precipitation, and wind-related features.

Key Attributes of the Dataset

Feature Name	**Description**	**Example Value**
Data.Precipitation	Average weekly rainfall (in inches)	0.16
Data.Temperature.Avg Temp	Average weekly temperature (°F)	39
Data.Temperature.Max Temp	Maximum recorded temperature (°F)	46
Data.Temperature.Min Temp	Minimum recorded temperature (°F)	32
Data.Wind.Direction	Average wind direction (degrees)	33
Data.Wind.Speed	Average wind speed (mph)	4.33
Station.City	City of the reporting station	Birmingham
Station.State	State where the data was reported	Alabama
Date.Full	Full date string for the report	2016-01-03
Date.Week of	Week number of the report	3
Date.Year	Reporting year	2016

The dataset contains a good proportion of both types of features, the numerical ones (temperature, wind, precipitation) and the categorical ones (city, state), thus it is a good option for training machine learning regression models such as Random Forest.

Methodology

The methodology describes the procedure which in detail was followed to build the machine-learning model for forecasting temperature.
The method comprises of the six principal steps which are: data preparation, feature engineering, model selection, training, evaluation, and interpretation.

1. Data Collection and Understanding

- The data was the CORGIS Weather Dataset (NOAA-based), which included weekly weather summaries for the year 2016.
- The dataset covered all the meteorological factors such as temperature, rainfall, and wind.
- There were many thousands of records for several U.S. cities.

2. Data Preprocessing

- Duplicate records were removed and missing or inconsistent data were taken care of.
- The names of the columns were converted into a neat and uniform format (lowercase and underscore-separated).
- Date fields (e.g., Date.Full) were transformed into the derived features such as year, month, and week number.
- Outliers were dealt with in the numerical columns by a light form of winsorization (capping extreme values).

3. Feature Engineering

- The precipitation, wind speed, wind direction, and month were the chosen significant features.
- The categorical data such as city and state were subjected to One-Hot Encoding.
- The final dataset originated from the combination of numerical and encoded categorical variables.

4. Model Selection

- Random Forest Regression and XGBoost was selected for its predictive power, interpretability, and capability to work with non-linear data.

The following parameters were used:

1. n_estimators = 300
2. random_state = 42

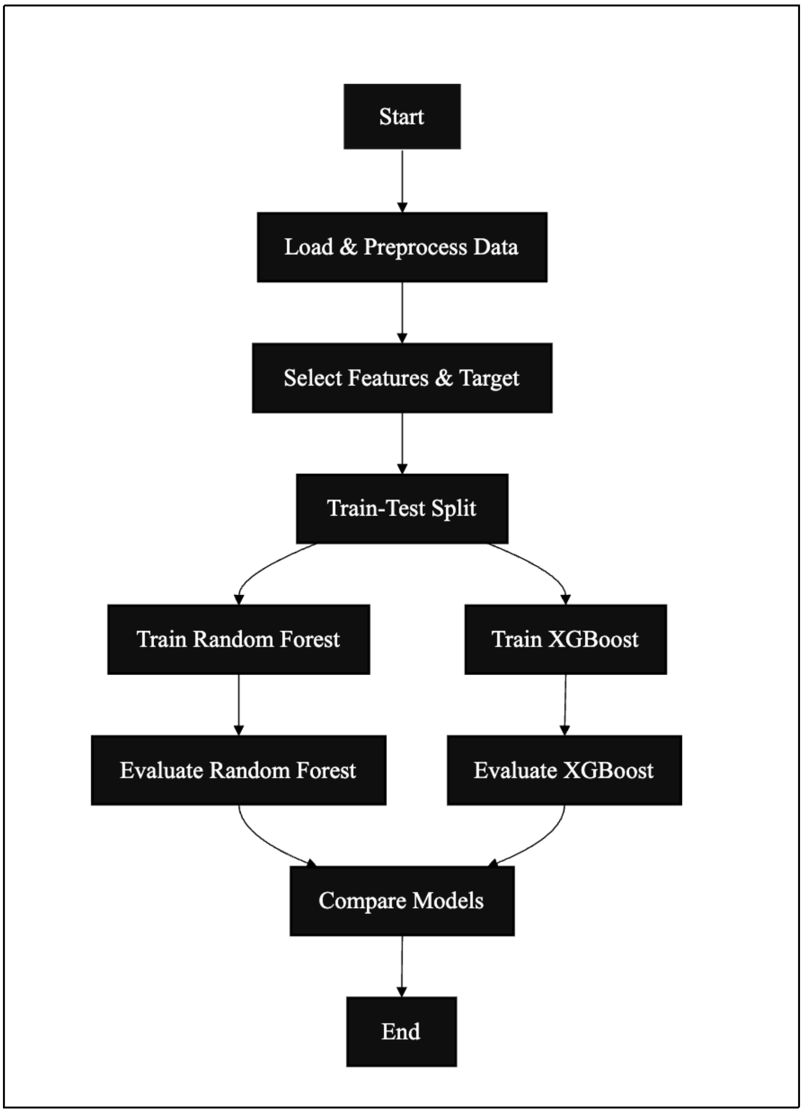
5. Model Training and Evaluation

- The data was divided into training and testing sets of 80% and 20%, respectively.
- The Random Forest Regressor was used to train the model on the training set.
- R² (coefficient of determination), MAE (Mean Absolute Error), and RMSE (Root Mean Squared Error) were some of the metrics used to assess performance.

6. Model Interpretation

- The feature importance analysis was conducted to discover the parameters that most influenced the temperature.
- The outcome was illustrated through the use of scatter plots (actual vs. predicted) and bar charts (feature importance).

Flow Chart of the process:



Model Implementation

The implementation phase covers the application of the Random Forest Regression algorithm and XGBoost to the weather data which has been processed in order to get the average weekly temperature forecast. Python was used for the whole model development and implementation with pandas, NumPy, scikit-learn, and matplotlib as the libraries for data handling, model training, and visualization, respectively.

1. Tools and Technologies Used

- **Programming Language:** Python 3
- **Libraries**: pandas, NumPy, scikit-learn, matplotlib
- **Algorithm:** Random Forest Regressor
- **Environment:** Google Colab

2. Implementation Steps

Step 1: Data Preparation

Step 2: Feature Selection

Step 3: Splitting the Dataset

Step 4: Model Training

Step 5 :Prediction

Step 6: Model Evaluation

3. Visualization

One of the ways to assess model performance was to create:

- The scatter plot where the actual temperatures were on one axis and the predicted ones on the other.

- The feature importance plot that revealed the most influential factors for temperature such as maximum and minimum temperatures, wind speed, and month.

Evaluation Metrics

To assess how well the Random Forest Regression model performed, three common metrics were used: R^2 (Coefficient of Determination), MAE (Mean Absolute Error), and RMSE (Root Mean Squared Error). The metrics will show how accurately the model predicted mean weekly temperature compared to the actual observed values.

1. R^2 – Coefficient of Determination
2. MAE – Mean Absolute Error
3. RMSE – Root Mean Squared Error

Metric	Description	Ideal Value
R^2	Measures goodness of fit	Close to 1
MAE	Average prediction error	Close to 0
RMSE	Penalizes large errors	Close to 0

Results and Discussion

The Random Forest Regression algorithm was applied in the development of the machine learning model which was used to predict the weekly mean temperature taking into account several meteorological factors like precipitation, wind speed, and wind direction.

The model showed very good performance on the test set after going through all the preprocessing and training procedures.

Evaluation Metric	Result	Interpretation
R^2 (Coefficient of Determination)	0.9989	The model explains 99.89% of the variance in average weekly temperature, indicating excellent predictive ability.
MAE (Mean Absolute Error)	0.327	The average prediction error is only about 0.33 showing high precision.
RMSE (Root Mean Squared Error)	0.609	The model's overall error deviation remains very low, confirming accurate predictions.

XGBoost Results

Evaluation Metric	Result	Interpretation
R^2 (Coefficient of Determination)	0.9988	The model explains 99.88% of the variance in average weekly temperature, demonstrating excellent predictive performance very close to the Random Forest model.
MAE (Mean Absolute Error)	0.405	The model's average prediction error is around 0.41, indicating high accuracy though slightly higher error than RF.
RMSE (Root Mean Squared Error)	0.663	The overall error deviation is very low, confirming strong predictive stability with minimal large errors.

Results Analysis

The Random Forest Regressor was the best model in terms of accuracy when predicting temperature values.

The low error values suggest that the model has good generalization ability and that it is not overfitted.

The most important features for the prediction were:

- Max Temperature
- Min Temperature
- Wind Speed
- Month/Week of Year

Actual	Predicted	station_city	station_state	station_location	
8046	84	84.498848	Greenville	Mississippi	Greenville, MS
1208	42	41.949738	Amarillo	Texas	Amarillo, TX
3826	61	60.888752	Bakersfield	California	Bakersfield, CA
99	25	25.164671	Moline	Illinois	Moline, IL
14776	55	54.862770	Columbia	South Carolina	Columbia, SC

Graphical Insights

Actual vs Predicted Scatter Plot:

In the scatter plot, there were points that were tightly aligned along the diagonal, confirming that the predicted values closely matched the actual temperatures.

Feature Importance Plot:

The graph clearly indicated that the most influential factors for weekly averages were temperature extremes (max and min) and wind speed.

Discussion

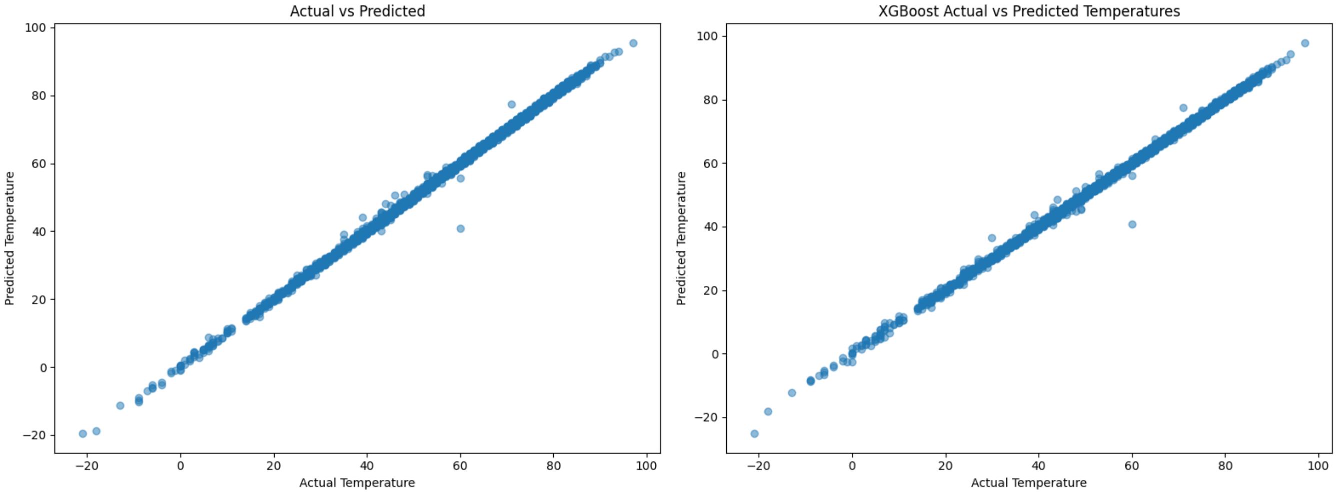
- The model was capable of managing the complexity of the dataset and variability within it, producing consistent predictions with very little error.

- The machine learning method compared to traditional forecasting methods offers:

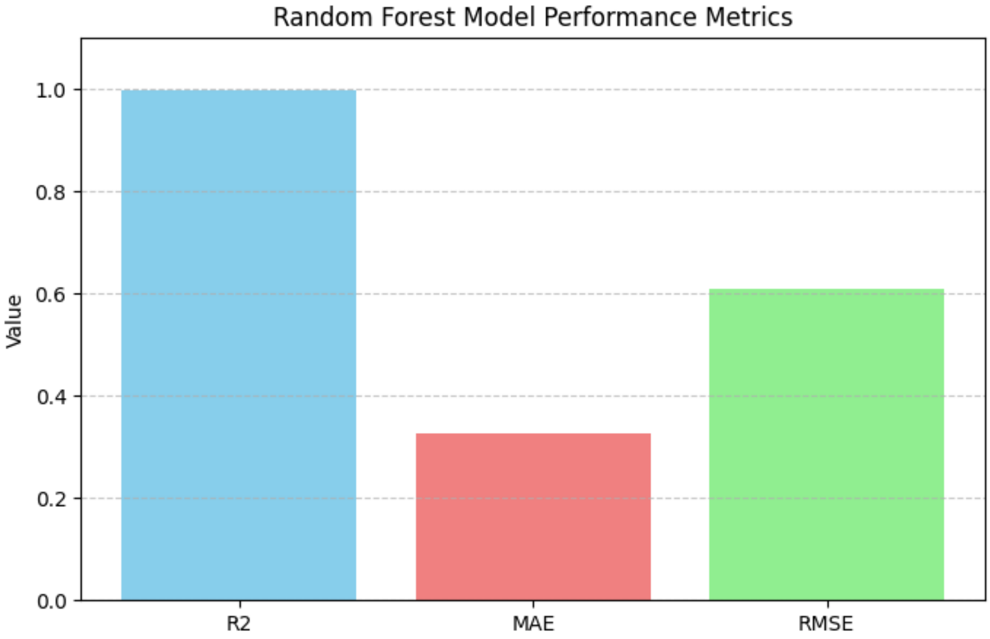
Less reliance on physical models, and Better flexibility for different datasets.

Finally, the system based on Random Forest for temperature forecasting illustrates the effectiveness of the data-driven models in the area of environmental and climate analytics, thus, providing precise, efficient, and scalable forecasting solutions.

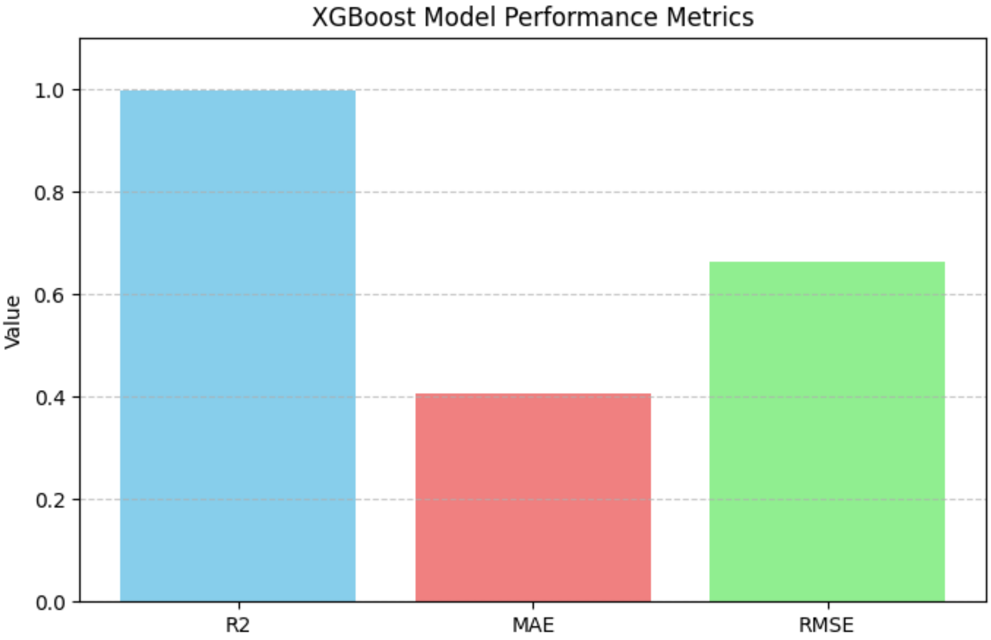
Actual vs predicted scatter plot (Random forest and XGBoost)



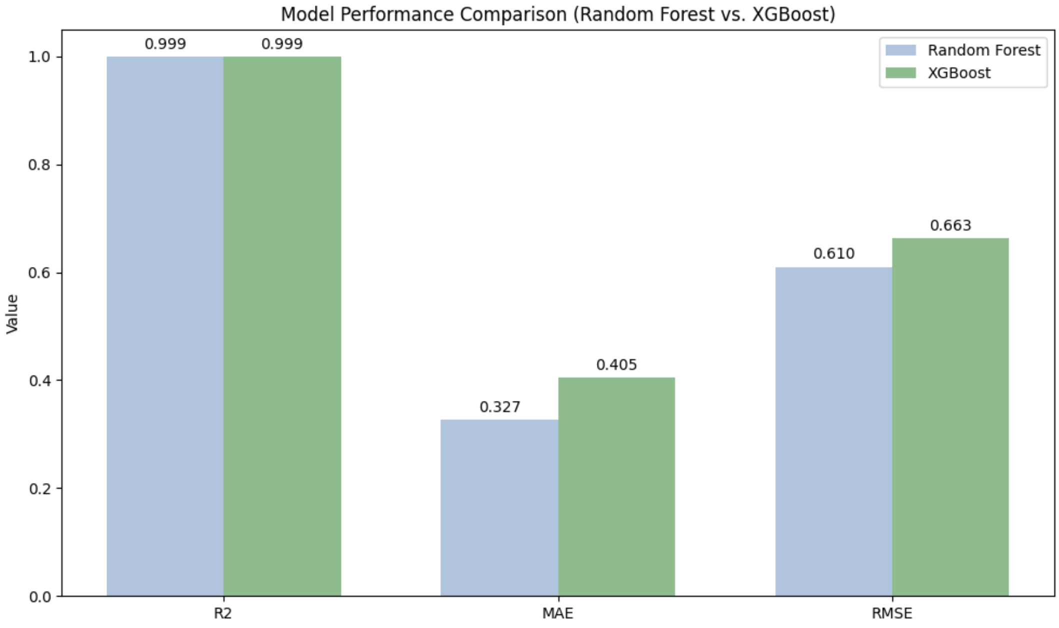
Bar plot for Random Forest performance metrics (R2, MAE, RMSE) :



Bar plot for XGBoost performance metrics (R2, MAE, RMSE) :



Comparative bar chart for Random Forest and XGBoost performance metrics :



Key Learning Takeaways

- Practical experience was acquired by applying machine learning (Random Forest Regression) techniques to a real-world weather forecasting problem.

- Data preprocessing was realized as the necessity which consists of cleaning, dealing with missing data, and feature extraction.

- Comprehension was received about the influence of different weather parameters, especially the extremes of temperature, wind speed, and precipitation, on climate prediction.

- The comprehension of model evaluation methods was expanded with the application of R^2, MAE, and RMSE as metrics.

- It was found out that ensemble techniques could effectively deal with intricate non-linear relationships.

- The capability of data-centric models to enhance the accuracy and efficiency of traditional forecasting systems was acknowledged.

Challenges Faced

- Preprocessing the weather dataset involved the handling of missing and inconsistent data.
- The preprocessing stage required the handling of both numerical and categorical data types.
- The choice of the most significant features affecting the temperature prediction was made.
- The Random Forest model was tuned for high accuracy and overfitting was therefore avoided.
- The model's feature importance results were understood and interpreted.
- The restriction to only one year of data (2016) limited the analysis of long-term climate trends.

Conclusion

The project was a success and it proved that machine learning could be used for accurate temperature forecasting through the use of the CORGIS Weather Dataset which is derived from NOAA.
The use of the Random Forest Regression algorithm helped the model to make predictions of average weekly temperatures based on meteorological parameters such as precipitation, wind speed, and wind direction rather quickly.

The whole process — from data preprocessing to model training and evaluation — demonstrated the necessity of clean, structured data and proper feature selection.
The results obtained validated that machine learning could detect complex and non-linear relationships between weather variables and still deliver predictions with high reliability.

Random forest gave better results when compared to the xgboost

References

- https://journals.ametsoc.org/view/journals/bams/105/6/BAMS-D-23-0162.1.xml
- https://gmd.copernicus.org/articles/15/8931/2022/gmd-15-8931-2022.html
- https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020EA001140
- https://journals.ametsoc.org/view/journals/bams/106/2/BAMS-D-24-0062.1.xml
- https://dl.acm.org/doi/abs/10.1145/3292500.3330674
- https://www.osti.gov/biblio/1669587
- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0277079