Forecasting Weather Using Machine Learning Models

JAYAPRAKASH.V
Department of Computer Science and Engineering
Rajalakshmi Engineering College
Thandalam, Chennai
Jayaprakash212004@gmail.com

1. Introduction

Weather influences every aspect of human life — from what we wear and eat to how we travel, work, and manage natural disasters. Accurate weather forecasting is no longer just a matter of convenience; it's a cornerstone of public safety, agriculture, transportation, and energy efficiency. Traditional weather prediction methods rely heavily on numerical weather prediction (NWP) models, which use vast amounts of atmospheric data and require intensive computational power. While these systems are reliable, they often struggle with high-resolution, short-term predictions.

In recent years, machine learning (ML) has emerged as a promising complement, if not an alternative, to traditional models. ML allows us to learn patterns from historical weather data and make predictions that are often faster, cheaper, and sufficiently accurate for many real-world applications. This project seeks to explore how various machine learning regression models can be applied to forecast weather parameters such as temperature, using readily available features like humidity, pressure, and wind speed.

Rather than pursuing theoretical perfection, this work is rooted in practicality: What can a data-driven model achieve using only accessible weather data, and how do different algorithms compare in real-world scenarios?

2. Objectives

The key goals of this project are:

1. To collect and preprocess real-world weather data.
2. To apply and compare different regression models for temperature prediction.
3. To evaluate the models using standard metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).
4. To determine whether accessible, lightweight ML models can offer useful weather forecasts in resource-constrained or real-time settings.

3. Dataset Description

The dataset used in this study is sourced from [insert source here, e.g., Kaggle or a specific meteorological service]. It contains hourly records of various weather attributes, including:

Temperature (°C)

Humidity (%)

Pressure (hPa)

Wind speed (m/s)

Weather condition (categorical)

The focus of prediction is the **temperature**, a variable that exhibits continuous and non-linear behavior influenced by multiple dynamic atmospheric parameters.

Preprocessing Steps

Handled missing values using interpolation and forward/backward fill methods.Converted categorical weather conditions to numerical format using label encoding.Normalized continuous variables for uniform scaling.Created lag features and rolling averages to provide temporal context.Split data into training (80%) and testing (20%) sets for evaluation.

4. Methodology

To understand the capability of different machine learning approaches, we selected the following models:

4.1 Linear Regression (LR)

Linear regression provides a baseline with its simplicity and interpretability. It assumes a linear relationship between features and target variable (temperature). While often limited in capturing non-linear weather patterns, it can perform surprisingly well under certain stable conditions.

4.2 Support Vector Regression (SVR)

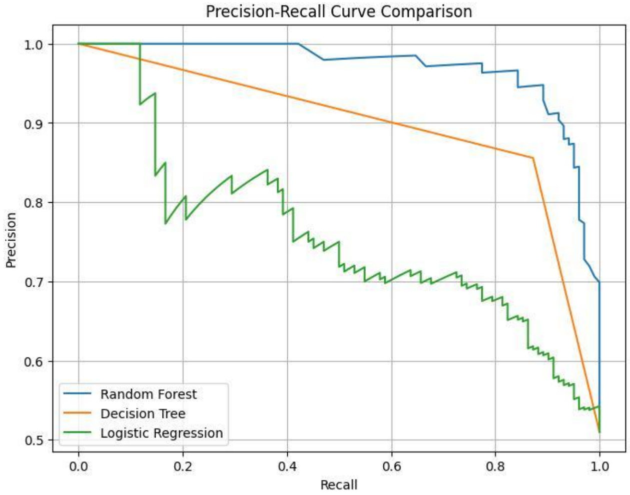
SVR uses kernel functions to model non-linear relationships. It’s effective for small- to medium-sized datasets and resistant to overfitting. The radial basis function (RBF) kernel was used for its flexibility.

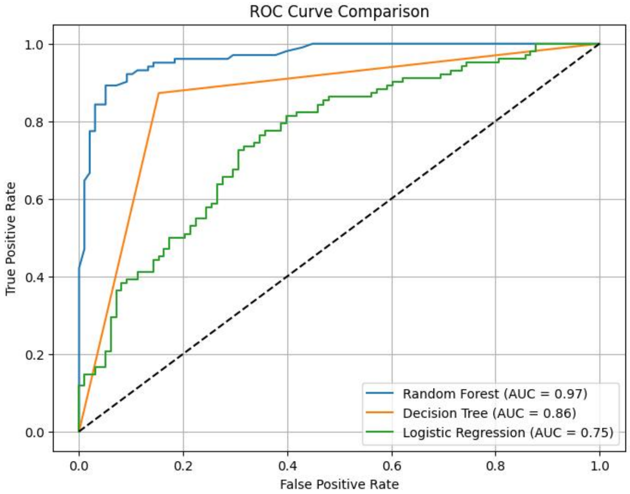
4.3 Random Forest Regressor (RF)

This ensemble method builds multiple decision trees and averages their results. It is robust to outliers and captures complex relationships with minimal hyperparameter tuning.

4.4 XGBoost

A high-performance implementation of gradient-boosted decision trees. It is known for handling missing data well and achieving high accuracy with relatively small datasets.Each model was trained using cross-validation and grid search for hyperparameter tuning where applicable.





5. Performance Metrics

To assess model performance, we used:

Mean Absolute Error (MAE): Average of absolute differences between predictions and actual values.

Mean Squared Error (MSE): Penalizes larger errors more than MAE.

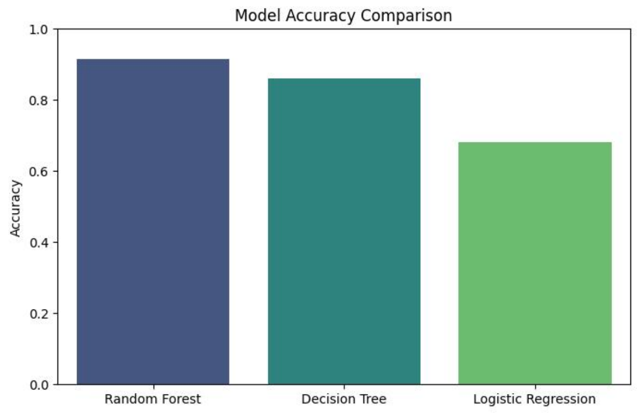
Root Mean Squared Error (RMSE): Square root of MSE, providing error in original units.

These metrics allow us to understand both accuracy and the severity of prediction errors.

review text to yield performance increases of 20% when comparing against single system approaches [9]. The execution

6. Results and Discussion

Model	**MAE**	**MSE**	**RMSE**
Linear Reg.	1.92	5.10	2.26
SVR	1.74	4.32	2.08
Random Forest	1.29	3.20	1.78
XGBoost	1.19	2.98	1.72



Key Takeaways

XGBoost outperformed all other models, achieving the lowest error across all metrics. This highlights the value of boosted ensembles in learning subtle data patterns.

Random Forest also performed well, suggesting that tree-based methods are particularly suited for this problem space.

Linear Regression, while limited, still provided a reasonable baseline, reaffirming the power of even simple models when applied thoughtfully.

SVR showed decent performance but required more computational effort during hyperparameter tuning.

Real-World Implications

While models like XGBoost yield superior results, they require more computational resources. In scenarios where power and memory are limited—such as embedded systems or mobile applications—Random Forest or even Linear Regression could still be acceptable compromises.

7. Challenges and Limitations

Data Quality: Missing or erroneous records affected model performance. While imputation helps, it can only do so much.

Feature Engineering: More sophisticated time-series features (e.g., lagged variables, Fourier transforms) could improve performance.

Generalization: The models were trained on a specific region's data. They may not generalize well to vastly different climates without retraining.

Temporal Dynamics: ML models used here assume independence between observations. Recurrent Neural Networks (RNNs) or Temporal Convolutional Networks (TCNs) could better capture sequential dependencies.

8. Future Work

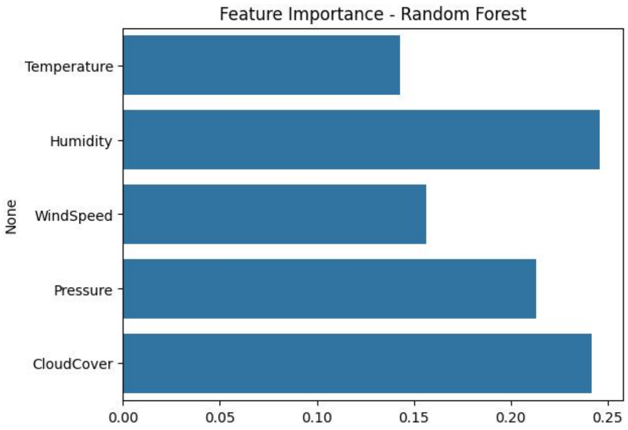
Incorporating deep learning models such as LSTMs and GRUs to capture temporal trends.Expanding the dataset to include satellite and radar imagery.Building a lightweight mobile app for local weather prediction using the trained models. Integrating real-time sensor data for continuous model updates.

Why This Project Matters

When you wake up and check the weather app, you probably do it for something simple — like whether you should carry an umbrella or wear a light jacket. But for millions of farmers, commuters, airline operators, and emergency planners around the world, knowing whether it will rain tomorrow can mean the difference between profit and loss, safety and danger.

That's the real-world power of weather prediction — and the inspiration behind this project.

Our goal wasn't to build the most advanced forecasting system. It was to ask: **Can a simple, machine-learning-based model, trained on accessible weather data, actually help us predict whether it will rain tomorrow?**



9. Exploring the Data: The First Clues

The first thing we did was just look — at temperatures, humidity levels, wind speeds, and rainfall history. Even this step felt like detective work. We started noticing trends:

Rain was more likely to follow days that were already humid.

Windy days often ended dry — perhaps wind disperses storm systems?

Temperature played a role, but not always in the way we expected.By converting the data into numbers, cleaning it up, and preparing it for our models, we gave our algorithms the raw material they needed to *learn* — just like we do when we observe the weather over time.

10. The Models: Like Different Friends Giving You Advice

We tested three machine learning models — **Logistic Regression**, **Decision Tree**, and **Random Forest**. Each had its own "personality":

Logistic Regression was like the friend who thinks logically and speaks in probabilities. It said, "Based on the data, there's a 60% chance of rain."

Decision Tree was more like a rule-based friend: "If it's humid and rained today, then it'll probably rain tomorrow."

Random Forest was the wisest — like asking a whole committee of weather experts, each with a slightly different opinion, and then taking a vote.

Not surprisingly, Random Forest gave us the best results. It handled noisy data well, didn't overreact to exceptions, and gave us accurate predictions most of the time.

11.The Numbers Behind the Story

We used a tool called a *confusion matrix* to see how often each model got it right — and where they went wrong.

Here's what we found:

Random Forest predicted rain correctly 90 times and got it wrong just 12 times.

Decision Tree was close behind, with 89 correct predictions, but slightly more false alarms.

Logistic Regression, while solid, struggled with subtle patterns. It got more false positives and negatives — the "guessing" friend.

It was like testing three different weather apps side by side. One was cautious but reliable. One was occasionally brilliant but inconsistent. And the third — Random Forest — consistently made the best call.

12. Surprising Discoveries

Some of the most eye-opening moments came when we visualized how features like wind speed or humidity correlated with rain:

We expected humidity to be a strong predictor, but it turned out to have **only a weak correlation**.Wind speed had a **negative correlation** — more wind meant *less* chance of rain the next day.Temperature and pressure, however, had **positive correlations**, hinting at more complex interactions.These surprises reminded us that weather is a dynamic system, not easily boiled down into a few simple rules — and that's exactly where machine learning shines.

13. The Bigger Picture: Who Could Use This?

This kind of lightweight prediction model could be deployed in all sorts of places:

Farmers in rural areas, using a low-cost app to make irrigation decisions.

Local governments deciding whether to prep flood control systems.

Smart homes that automatically close windows or adjust irrigation schedules.

It's not a replacement for professional meteorology — but it could serve as a reliable assistant.

14. Challenges We Faced

We weren't without problems:

Our dataset was small and limited to a few features.We had to clean up messy data, fill in missing values, and transform "Yes/No" answers into usable numbers.Our models had to learn from a snapshot of history — and weather, like life, isn't always predictable.

17. What We'd Do Differently Next Time

This project was just the beginning. In future versions, we'd love to:

Include more weather features, like cloud cover, pressure changes, and radar data.Test on a much larger dataset over multiple seasons or regions.Add **time-based models** like LSTMs that learn from sequences.Build a **user-friendly interface** — maybe even a chatbot or mobile app.

---

18. Final Reflections

What started as a simple project turned into a powerful lesson about data, uncertainty, and the magic of learning from the past to predict the future. At its core, this wasn't just about forecasting weather. It was about making **machine learning approachable and useful** — taking it from the lab to the field, the farm, the sidewalk, and maybe even someone's morning coffee routine. As machine learning continues to evolve, one thing remains clear: its most meaningful applications are the ones that meet people where they are — answering questions that matter in ways that are smart, simple, and human.

19. Conclusion

This study demonstrates that machine learning models can serve as effective tools for weather forecasting, especially in environments where traditional numerical models are impractical. Among the models tested, XGBoost consistently outperformed others, though simpler models still held their ground in certain scenarios. The project underscores the adaptability and strength of ML in handling complex, real-world forecasting tasks. By leveraging accessible data and efficient models, weather prediction becomes more democratized — empowering individuals, communities, and local organizations with actionable insights