

Task 1 : Weather Forecasting

TABLE OF CONTENT

1. Introduction.....	2
2. Dataset Description.....	2
3. Data Preprocessing.....	2
3.1 Data Inspection.....	2
3.2 Handle Missing Data.....	2
3.3 Non-stationary Columns.....	3
3.3 Filtering out the unwanted features.....	3
3.4 Outlier Detection.....	3
4. Exploratory Data Analysis (EDA).....	4
4.1 Consistency of the dataset.....	4
4.2 Stationarity of Time Series Data.....	4
4.3 Correlation matrix.....	4
4.3 Data Distribution Analysis.....	5
4.4 Decomposition.....	6
5. Model Selection.....	7
5.1 Defining Model for Time Series Data.....	7
5.1.1 Model Types.....	7
5.2 Defining Model to predict rain.....	9
5.2.1 Feature Importance.....	9
6. Model Evaluation.....	9
6.1 Time Series Data (Five feature columns).....	9
6.2 Rain_or_not prediction.....	11
7. Conclusion.....	12
8. References.....	12

1. Introduction

The objective of this project is to develop a machine learning model that predicts whether it will rain or not, enabling farmers to make better decisions regarding irrigation, planting, and harvesting. Traditional weather forecasts often fail to capture hyper-local variations, making accurate predictions crucial for smart agriculture.

The dataset consists of **300 days of daily weather observations**, manually recorded in an Excel sheet. It includes various weather attributes relevant to rainfall prediction. However, the data contains **missing values, incorrect entries, and formatting inconsistencies**, requiring thorough preprocessing before model training.

2. Dataset Description

Column	Description	Type
avg_temperature	Average temperature in °C	float64
humidity	Humidity in percentage	float64
avg_wind_speed	Average wind speed in km/h	float64
rain_or_not	Binary label (1 = rain, 0 = no rain)	binary
cloud_cover	-	float64
pressure	-	float64
date	Date of observation	float64

3. Data Preprocessing

3.1 Data Inspection

This dataset consists of 311 datarows.

5 key features include numerical values. And the target feature has a binary label.

3.2 Handle Missing Data

<pre>print(df.isnull().sum())</pre> <pre>date 0 avg_temperature 15 humidity 15 avg_wind_speed 15 rain_or_not 0 cloud_cover 15 pressure 0 dtype: int64</pre>	<p>There were 15 unique rows that had null values in the columns avg_temperature, humidity, avg_wind_speed and cloud_cover.</p> <p>These were imputed using KNN imputation that finds the most similar data points (based on available features like pressure and rain_or_not) and estimates missing values using their average.</p>
--	---

3.3 Non-stationary Columns

ADF Statistic: -2.430551
 p-value: 0.133299
 Critical Values:
 1%: -3.453
 5%: -2.871
 10%: -2.572

Out of all available time-series data columns, avg_temperature was not stationary.

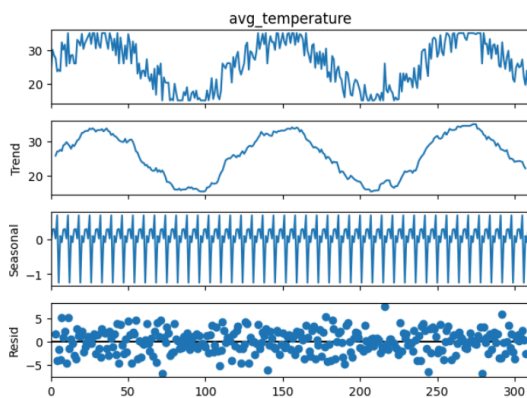
Carry out differencing before it is applied to a time-series model for prediction that requires data to be stationary.

3.3 Filtering out the unwanted features

All features were considered important so any columns were not dropped.

3.4 Outlier Detection

1. Method 1

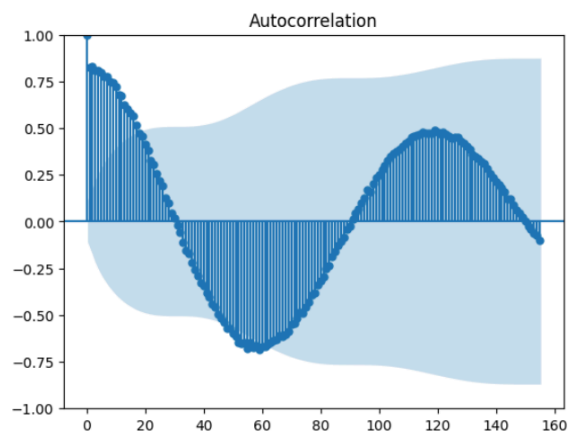


Outliers were then identified by decomposing time series data into trend, seasonality, and residual components, then identifying anomalies where residuals exceed **3 standard deviations** from the mean.

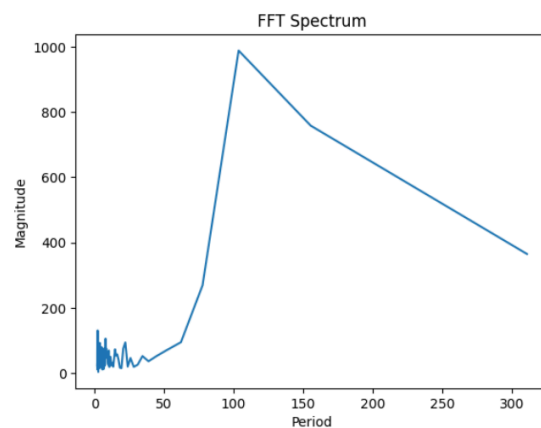
Prior to detecting outliers, we had to figure out the seasonality time period for all five columns.

Two main methods:

1. Autocorrelation plot



2. Fast Fourier transform

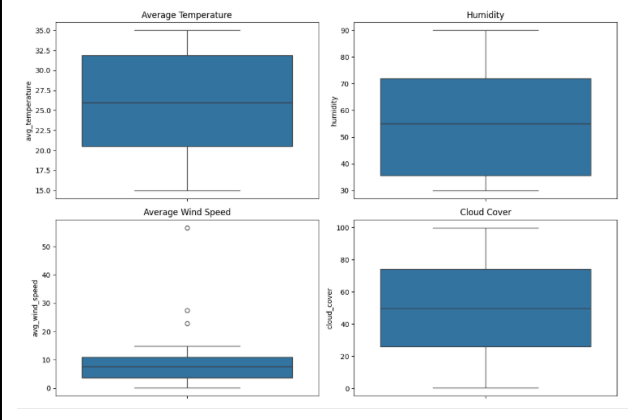


The highest peaks will tell us about Seasonality.

2. Method 2

Outliers were identified using the **Interquartile Range (IQR) method**, which identifies values that fall **outside 1.5 times the IQR** from the first (Q1) and third quartiles (Q3).

Outliers are values lower than $Q1 - 1.5 * IQR$ or higher than $Q3 + 1.5 * IQR$.



Outlier Removal: The identified outliers were removed and the values were replaced by new values using **linear interpolation**.

4. Exploratory Data Analysis (EDA)

4.1 Consistency of the dataset

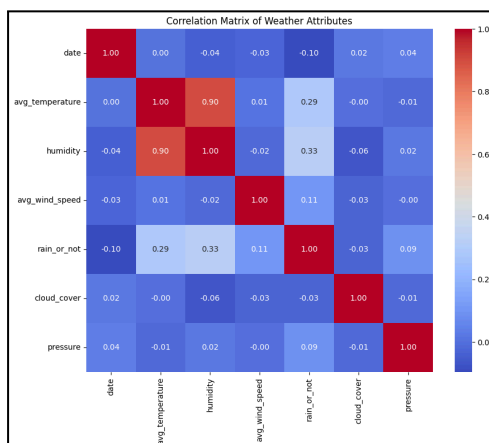
Dates are in chronological order.

4.2 Stationarity of Time Series Data

	Column	ADF Statistic	p-value	Critical Value (1%)	Stationary
0	avg_temperature	-2.43055	0.133299	-3.45256	No
1	humidity	-3.53646	0.00709872	-3.45279	Yes
2	avg_wind_speed	-18.6347	2.05626e-30	-3.45162	Yes
3	cloud_cover	-17.0402	8.2061e-30	-3.45162	Yes
4	pressure	-18.5718	2.08222e-30	-3.45162	Yes

4.3 Correlation matrix

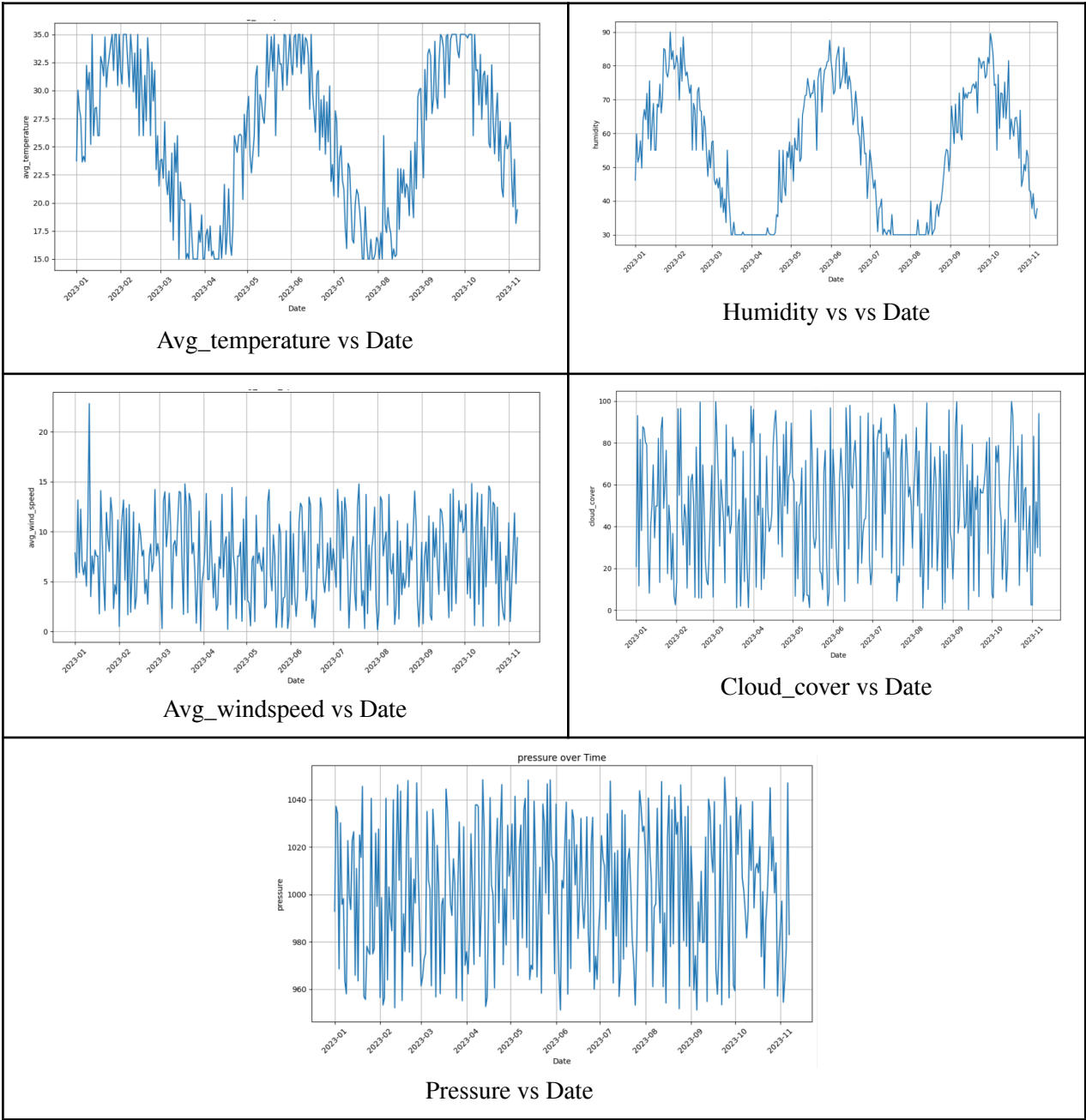
A heatmap of the correlation matrix is generated to understand the relationships between features.



Key takeaways:

- Temperature and humidity are strongly related.
- Rain probability increases with higher humidity and temperature.

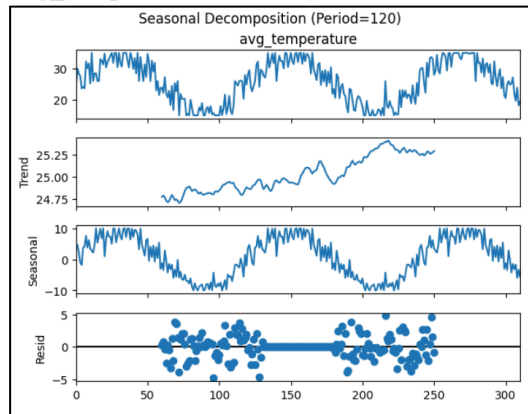
4.3 Data Distribution Analysis



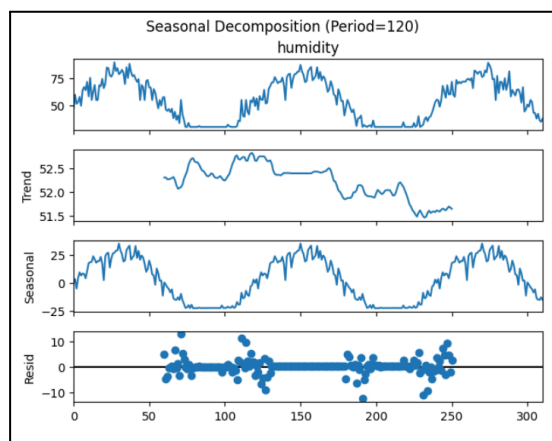
4.4 Decomposition

For the five attributes, we decomposed under varying seasonalities to pick the best time frame that captures seasonality and trend best.

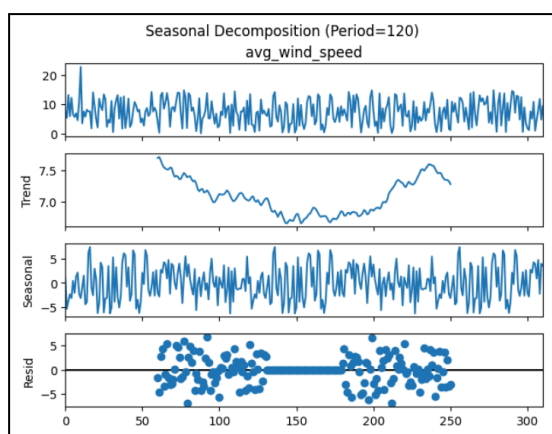
1. Avg_temperature: 120 days



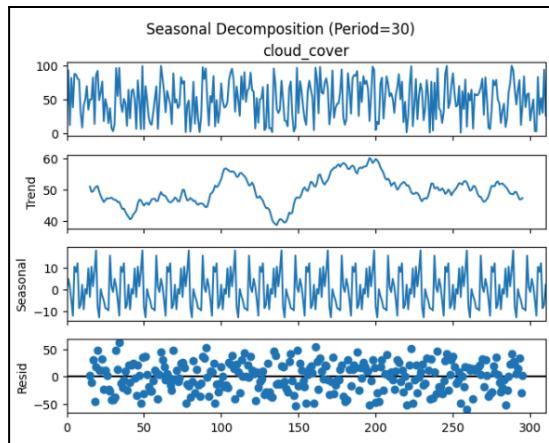
2. Humidity: 120 days



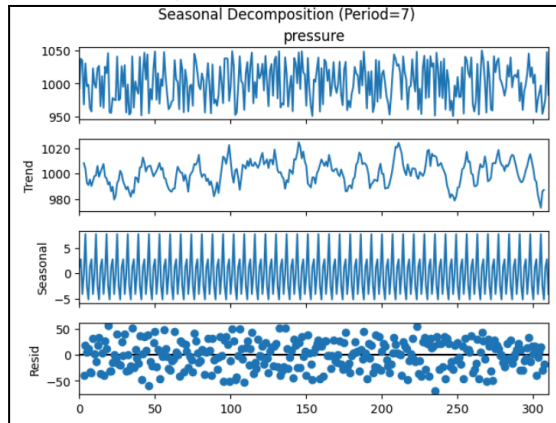
3. Avg_wind_speed: 120 days



4. Cloud_cover: 30 days



5. Pressure: 7 days



5. Model Selection

Model Selection happens in two parts

5.1 Defining Model for Time Series Data

Multiple models were tested across all time series data columns, and the best-performing model for each weather attribute was selected for further analysis.

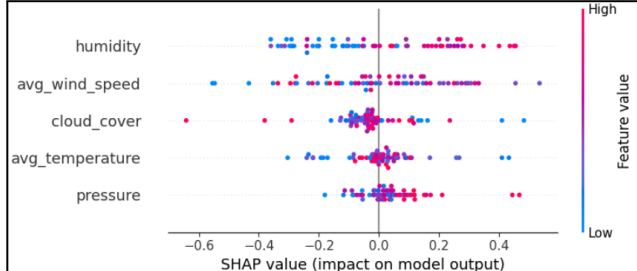
5.1.1 Model Types

Model	Weather attribute	Metrics
1. SARIMA Model (for the best seasonality)	avg_temperature	MAE: 3.396 RMSE: 4.123 MAPE: 13.28%
	humidity	MAE: 6.722 RMSE: 9.186 MAPE: 11.81%

	avg_wind_speed	MAE: 4.692 RMSE: 5.441 MAPE: 228.55%
	cloud_cover	MAE: 22.664 RMSE: 27.053 MAPE: 165.99%
	pressure	MAE: 19.698 RMSE: 24.770 MAPE: 1.98%
2. PROPHET	avg_temperature	MAE: 4.9445 RMSE: 5.9465 MAPE: 21.0034
	humidity	MAE: 11.3043 RMSE: 13.4617 MAPE: 22.6945
	avg_wind_speed	MAE: 3.9823 RMSE: 4.6337 MAPE: 194.2689
	cloud_cover	MAE: 22.6550 RMSE: 27.6244 MAPE: 193.0371
	pressure	MAE: 19.5683 RMSE: 24.5401 MAPE: 1.9679
3. LSTM	avg_temperature	MAE: 2.6810 RMSE: 3.2447 MAPE: 11.6024
	humidity	MAE: 7.3959 RMSE: 8.6781 MAPE: 15.8318
	avg_wind_speed	MAE: 3.9589 RMSE: 4.5221 MAPE: 171.2100
	cloud_cover	MAE: 23.1177 RMSE: 27.8132 MAPE: 214.9760
	pressure	MAE: 21.9026 RMSE: 27.3500 MAPE: 2.2069
4. Stacking Model	Although certain values were promising the model showed unexpected fluctuations. Hence, an unstable model that was not taken to consideration further	
*Selected models for given attributes have been highlighted in green		

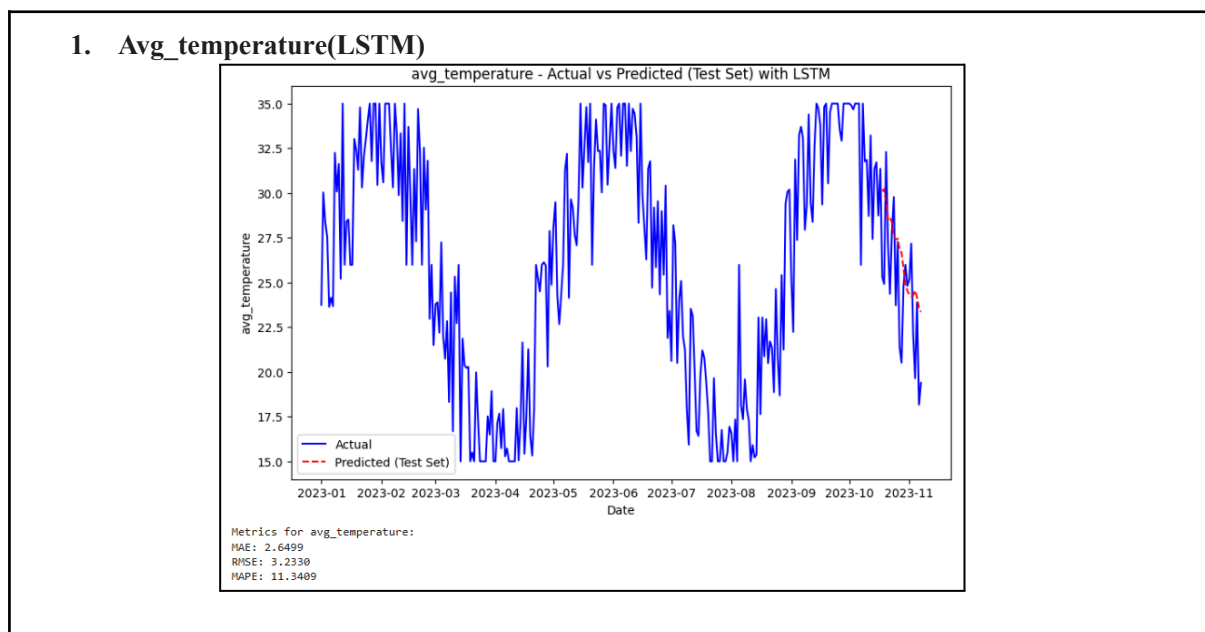
5.2 Defining Model to predict rain

5.2.1 Feature Importance

Test	Analysis																		
<div>1. Decision Tree model</div> <table><tr><th></th><th>Feature</th><th>Importance</th></tr><tr><td>2</td><td>avg_wind_speed</td><td>0.264115</td></tr><tr><td>1</td><td>humidity</td><td>0.210432</td></tr><tr><td>4</td><td>pressure</td><td>0.20019</td></tr><tr><td>0</td><td>avg_temperature</td><td>0.183424</td></tr><tr><td>3</td><td>cloud_cover</td><td>0.141839</td></tr></table>		Feature	Importance	2	avg_wind_speed	0.264115	1	humidity	0.210432	4	pressure	0.20019	0	avg_temperature	0.183424	3	cloud_cover	0.141839	<div><ul style="list-style-type: none">- avg_wind_speed (0.264) is the most important feature in predicting rain.- humidity (0.210) and pressure (0.200) also play significant roles.- avg_temperature (0.183) and cloud_cover (0.142) have lower importance but still contribute.</div> <div>All columns should be taken into account.</div>
	Feature	Importance																	
2	avg_wind_speed	0.264115																	
1	humidity	0.210432																	
4	pressure	0.20019																	
0	avg_temperature	0.183424																	
3	cloud_cover	0.141839																	
<div>2. Shap Analysis</div> <table><tr><th></th><th>Feature</th><th>SHAP Importance</th></tr><tr><td>humidity</td><td>humidity</td><td>0.200108</td></tr><tr><td>avg_wind_speed</td><td>avg_wind_speed</td><td>0.178172</td></tr><tr><td>cloud_cover</td><td>cloud_cover</td><td>0.0955053</td></tr><tr><td>avg_temperature</td><td>avg_temperature</td><td>0.0829048</td></tr><tr><td>pressure</td><td>pressure</td><td>0.0725688</td></tr></table>		Feature	SHAP Importance	humidity	humidity	0.200108	avg_wind_speed	avg_wind_speed	0.178172	cloud_cover	cloud_cover	0.0955053	avg_temperature	avg_temperature	0.0829048	pressure	pressure	0.0725688	<div><ul style="list-style-type: none">- All features were recognized to be important where pressure has the lowest importance.</div> <div></div>
	Feature	SHAP Importance																	
humidity	humidity	0.200108																	
avg_wind_speed	avg_wind_speed	0.178172																	
cloud_cover	cloud_cover	0.0955053																	
avg_temperature	avg_temperature	0.0829048																	
pressure	pressure	0.0725688																	

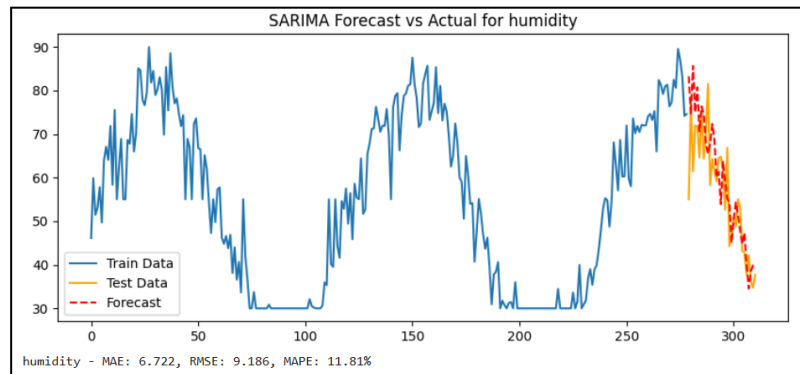
6. Model Evaluation

6.1 Time Series Data (Five feature columns)

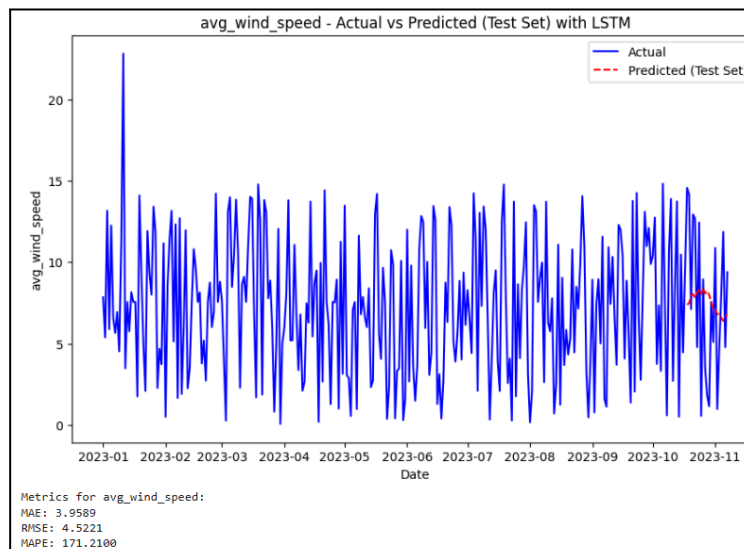


2. Humidity (SARIMA)

s:120 days

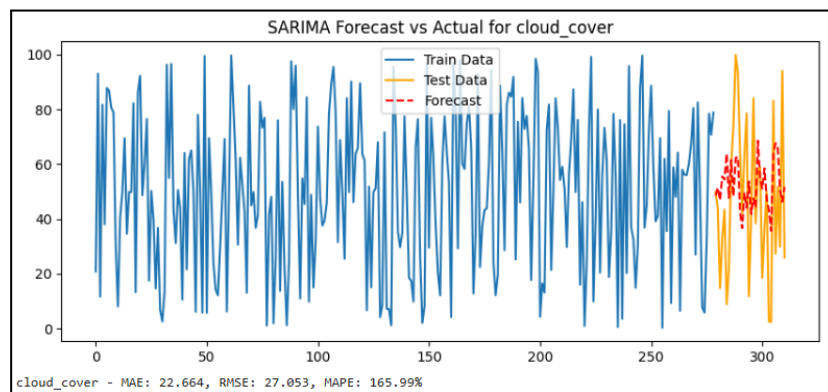


3. Avg_wind_speed (LSTM)

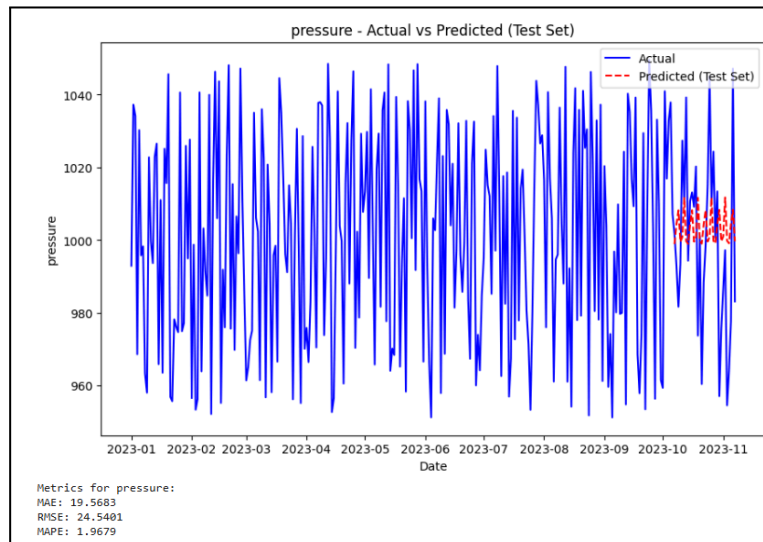


4. Cloud_cover (SARIMA)

s: 30 days



5. Pressure (SARIMA) s: 7 days



6.2 Rain_or_not prediction

The goal of this section is to find a classification model that predicts whether it will rain (1) or not (0) based on a time series dataset. Various machine learning models were tested to determine their effectiveness in capturing weather patterns and making accurate predictions.

	Model(I)	Accuracy	Model(II)	Accuracy
0	Ridge Classifier CV	0.6825	XGBoost	0.5714
1	Logistic Regression CV	0.6667	Gaussian NB	0.5714
2	Logistic Regression	0.6667	Random Forest	0.5714
3	SVM (RBF Kernel)	0.6508	Extra Trees	0.5714
4	Gaussian Process Classifier	0.6508	Passive Aggressive Classifier	0.5556
5	SVM (Linear Kernel)	0.6349	Perceptron	0.5238
6	CatBoost	0.6349	Decision Tree	0.5079
7	Bernoulli NB	0.6349	AdaBoost	0.5079
8	K-Nearest Neighbors	0.6032	Neural Network (MLP)	0.5079
9	LightGBM	0.6032	Gradient Boosting	0.4921

Ridge Classifier CV was selected.

7. Conclusion

The machine learning model successfully analyzes historical weather data to predict rainfall, enabling more accurate decision-making for farmers. The dataset underwent rigorous preprocessing, including handling missing values, outlier detection, and time series decomposition, ensuring high-quality inputs for model training.

Among the various classification models tested, Ridge Classifier CV emerged as the best-performing model for predicting rainfall (YES or NO), achieving the highest accuracy. Feature importance analysis highlighted average wind speed, humidity, and pressure as key factors influencing rainfall predictions.

8. References

- [1] [data-imputation-demystified-time-series-data](#)
- [2] [time-series-forecasting-real-world-challenges](#)
- [3] [top-10-binary-classification-algorithms](#)
- [4] [RidgeClassifier](#)