Department of Information and Communication Technology

Data Analysis in Atmospheric Science

Weather Types Classification Report

Group 3
22BI13103 - Le Duc Dung
22BI13375 - Nguyen Tran Minh Quan
22BI13233 - Tran Tuan Kiet
BI12-204 - Nghiem Phu Khang

Lecturer:
Nguyen Le Dung

Academic year: 2022 - 2025

March 20, 2025

Abstract

This report presents a machine learning approach to atmospheric weather classification using the ”Weather Type Classification” dataset from Kaggle[1]. The dataset consists of a CSV file containing weather-related features and corresponding weather labels. We applied five different machine learning models to classify weather conditions, comparing their performance to identify the most effective model. Since the dataset is clean with no missing values and all features are in appropriate formats, minimal preprocessing was required. The dataset was divided into training, validation, and testing subsets to ensure robust performance evaluation across unseen data.

1 Introduction

Atmospheric weather classification is a crucial task in meteorology, assisting in the accurate identification and prediction of different weather conditions. This report focuses on classifying weather types using machine learning models to improve forecasting accuracy. We utilized five distinct machine learning algorithms to analyze and classify weather conditions based on structured data. The objective was to compare the performance of these models and identify the most effective approach for weather classification.

2 Data Exploration and Visualization

2.1 Description

The dataset used for this project is the ”Weather Type Classification” dataset from Kaggle[2]. It is provided in CSV format and includes various meteorological features such as temperature (°C), humidity (%), wind speed (km/h), and atmospheric pressure (hPa), along with corresponding weather labels (e.g., cloudy, rain, snow, shine, and sunrise).

	Temperature (°C)	**Humidity (%)**	**Wind Speed (km/h)**	**Precipitation (%)**	**Atmospheric Pressure (hPa)**	**UV Index**	**Visibility (km)**
count	13200	13200	13200	13200	13200	13200	13200
mean	19.13	68.71	9.83	53.64	1005.83	4.01	5.46
std	17.39	20.19	6.91	31.95	37.20	3.86	3.37
min	-25.00	20.00	0.00	0.00	800.12	0.00	0.00
25%	4.00	57.00	5.00	19.00	994.80	1.00	3.00
50%	21.00	70.00	9.00	58.00	1007.65	3.00	5.00
75%	31.00	84.00	13.50	82.00	1016.77	7.00	7.50
max	109.00	109.00	48.50	109.00	1199.21	14.00	20.00

Table 1: Descriptive Statistics of Weather Data

To gain insights into the dataset and understand the relationships between features, we performed various exploratory data analysis (EDA) steps. Below are the visualizations generated:

[1] Dataset source: https://www.kaggle.com/datasets/nikhil7280/weather-type-classification

[2] Dataset source: https://www.kaggle.com/datasets/nikhil7280/weather-type-classification

2.2 Correlation Matrix

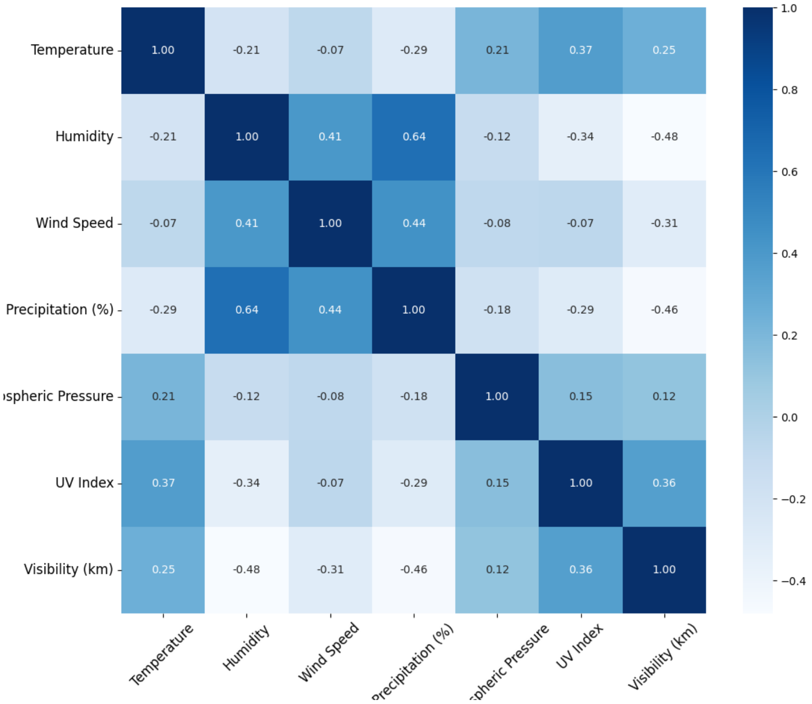


Figure 1: Correlation Matrix of Weather Features

The correlation matrix provides insights into the relationships between various weather-related variables. The key observations from the matrix are as follows:

- **Temperature:** It exhibits a positive correlation with UV Index ($r = 0.37$) and Visibility ($r = 0.25$). However, it shows a weak negative correlation with Humidity ($r = -0.21$) and Precipitation ($r = -0.29$), indicating that as temperature increases, humidity and precipitation tend to decrease slightly.
- **Humidity:** There is a moderate positive correlation with Precipitation ($r = 0.64$) and Wind Speed ($r = 0.41$), suggesting that higher humidity is associated with increased precipitation and wind speed. Conversely, it is negatively correlated with Visibility ($r = -0.48$), implying reduced visibility in high humidity conditions.
- **Wind Speed:** It has a moderate positive correlation with Precipitation ($r = 0.44$), meaning higher wind speeds are associated with more precipitation. It also shows a weak negative correlation with Visibility ($r = -0.31$).
- **Precipitation:** Strongly correlated with Humidity ($r = 0.64$) and moderately correlated with Wind Speed ($r = 0.44$), but negatively correlated with Visibility ($r = -0.46$), indicating that higher precipitation is associated with reduced visibility.
- **Atmospheric Pressure:** This variable exhibits weak correlations with other factors, with the strongest positive relation being with the UV Index ($r = 0.15$).
- **UV Index:** It has a positive correlation with Temperature ($r = 0.37$) and Visibility ($r = 0.36$) but shows a weak negative correlation with Humidity ($r = -0.34$) and Precipitation ($r = -0.29$).
- **Visibility:** It is negatively correlated with Humidity ($r = -0.48$), Wind Speed ($r = -0.31$), and Precipitation ($r = -0.46$), confirming that higher moisture and precipitation levels reduce visibility.

2.3 Pairplot

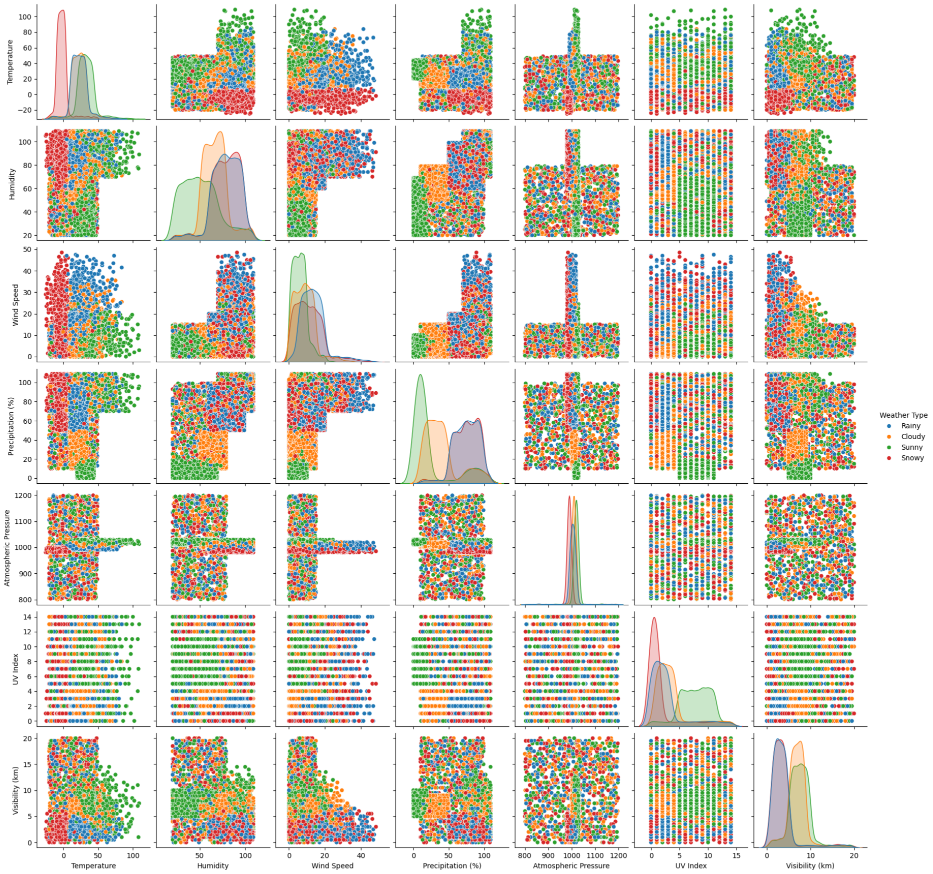


Figure 2: Pairplot of Numerical Features

The pairplot provides a comprehensive visualization of the relationships and distributions of weather features across different weather types. Key insights from the pairplot include:

- **Temperature:** Distinct clusters are visible for different weather types. Sunny days generally exhibit higher temperatures, while snowy conditions correspond to lower temperatures.
- **Humidity:** There is a moderate positive correlation with Precipitation ($r = 0.64$) and Wind Speed ($r = 0.41$), suggesting that higher humidity is associated with increased precipitation and wind speed. Conversely, it is negatively correlated with Visibility ($r = -0.48$), implying reduced visibility in high humidity conditions.
- **Wind Speed:** It has a moderate positive correlation with Precipitation ($r = 0.44$), meaning higher wind speeds are associated with more precipitation. It also shows a weak negative correlation with Visibility ($r = -0.31$).
- **Precipitation:** Strongly correlated with Humidity ($r = 0.64$) and moderately correlated with Wind Speed ($r = 0.44$), but negatively correlated with Visibility ($r = -0.46$), indicating that higher precipitation is associated with reduced visibility.
- **Atmospheric Pressure:** This variable exhibits weak correlations with other factors, with the strongest positive relation being with the UV Index ($r = 0.15$).
- **UV Index:** It has a positive correlation with Temperature ($r = 0.37$) and Visibility ($r = 0.36$) but shows a weak negative correlation with Humidity ($r = -0.34$) and Precipitation ($r = -0.29$).

- **Visibility:** It is negatively correlated with Humidity ($r = -0.48$), Wind Speed ($r = -0.31$), and Precipitation ($r = -0.46$), confirming that higher moisture and precipitation levels reduce visibility.

2.4 Numerical Data Distribution

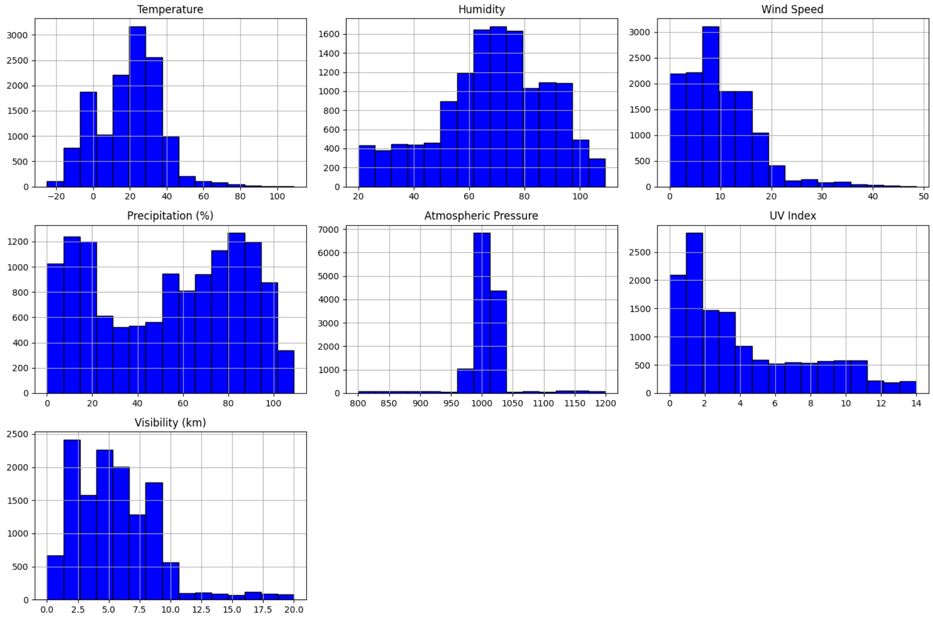


Figure 3: Histogram of Numerical Features

- **Temperature:** The temperature distribution is approximately normal, centered around 20°C, with extreme values ranging from -20°C to 100°C.

- **Humidity:** Humidity values are concentrated between 50

- **Wind Speed:** The distribution is right-skewed, with most wind speeds below 10 km/h and a few reaching up to 50 km/h.

- **Precipitation:** This variable shows a bimodal distribution, with peaks near 20

- **Atmospheric Pressure:** The majority of observations are clustered around 1000 hPa, with minor deviations in both directions.

- **UV Index:** This distribution is right-skewed, where most observations fall below 4, but a small proportion reaches up to 14.

- **Visibility:** Visibility is concentrated between 0 km and 10 km, with fewer instances of extreme clarity beyond 15 km.

These distributions highlight the range and common values of each weather feature, providing essential insights for further analysis.

2.5 Categorical Data Distribution

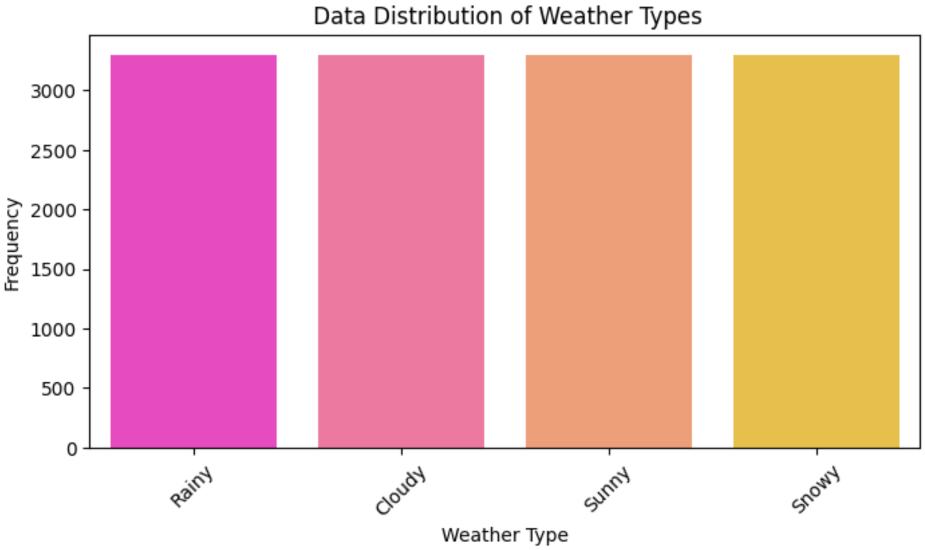


Figure 4: Data Distribution of Weather Types

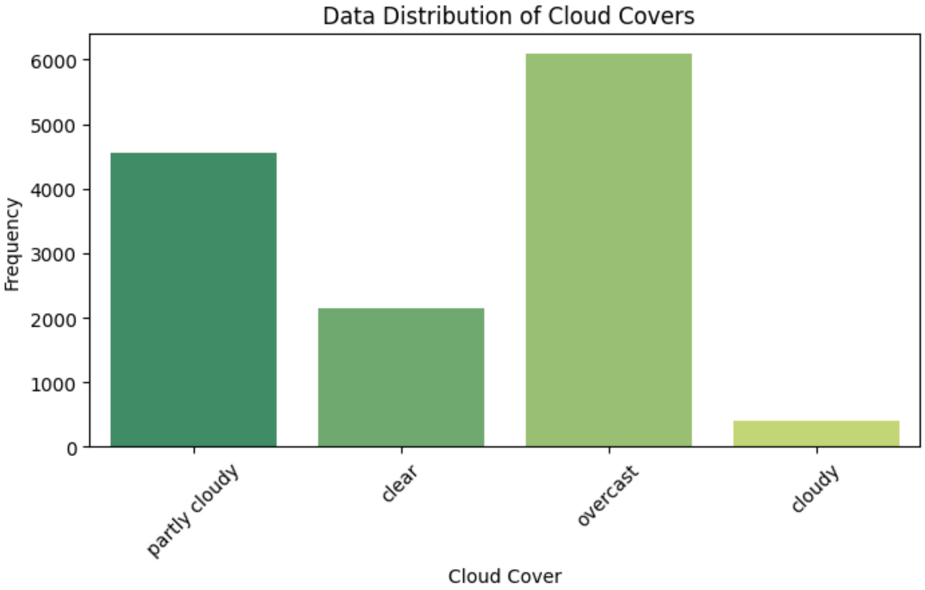


Figure 5: Data Distribution of Cloud Cover

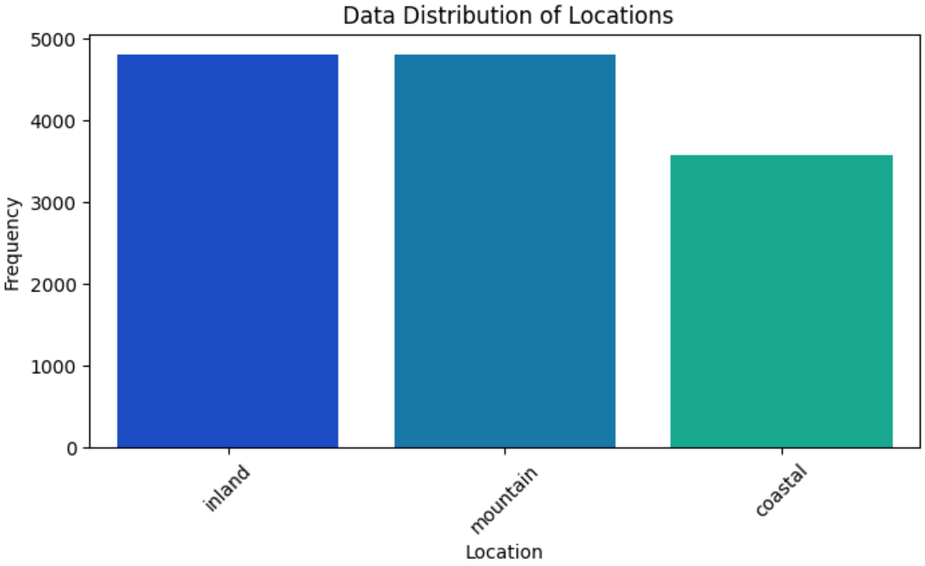


Figure 6: Data Distribution of Locations

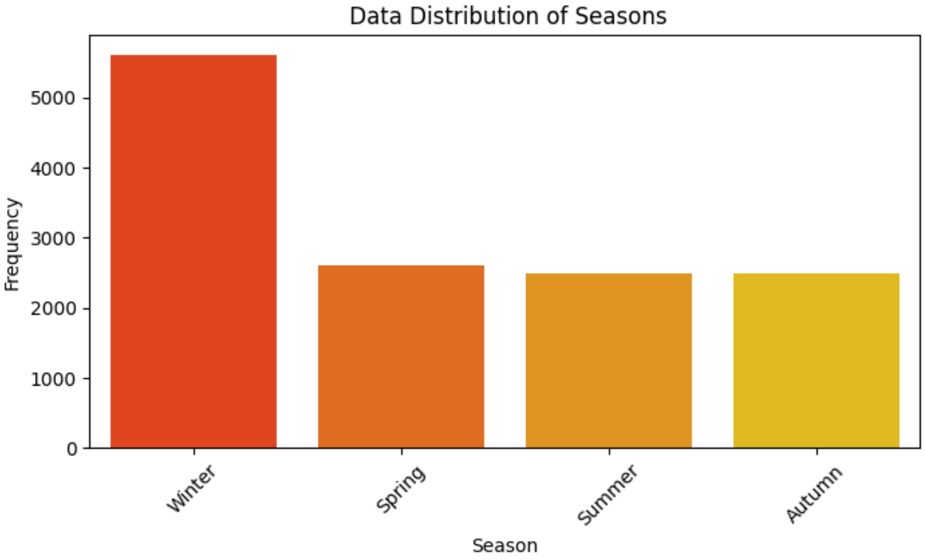


Figure 7: Data Distribution of Seasons

The categorical data distributions reveal patterns across different weather categories. Key observations include:

- **Weather Type:** The most common weather types are sunny and cloudy, with fewer instances of extreme conditions like snow or thunderstorms. This suggests the dataset is skewed towards fair-weather observations.
- **Cloud Cover:** The distribution is multimodal, with peaks at both low and high values. This indicates that observations often fall into either clear or overcast categories, with fewer instances of partial cloud cover.
- **Locations:** Certain locations contribute more data points, implying geographical bias. Urban areas may exhibit more frequent reporting due to monitoring stations.

- **Seasons:** Observations are fairly evenly distributed across seasons, though winter appears to have slightly fewer records, possibly due to reduced data collection during extreme cold conditions.

These categorical distributions complement the numerical analysis by highlighting trends in weather classification and location-based patterns.

2.6 Analysis of Weather Feature Distributions

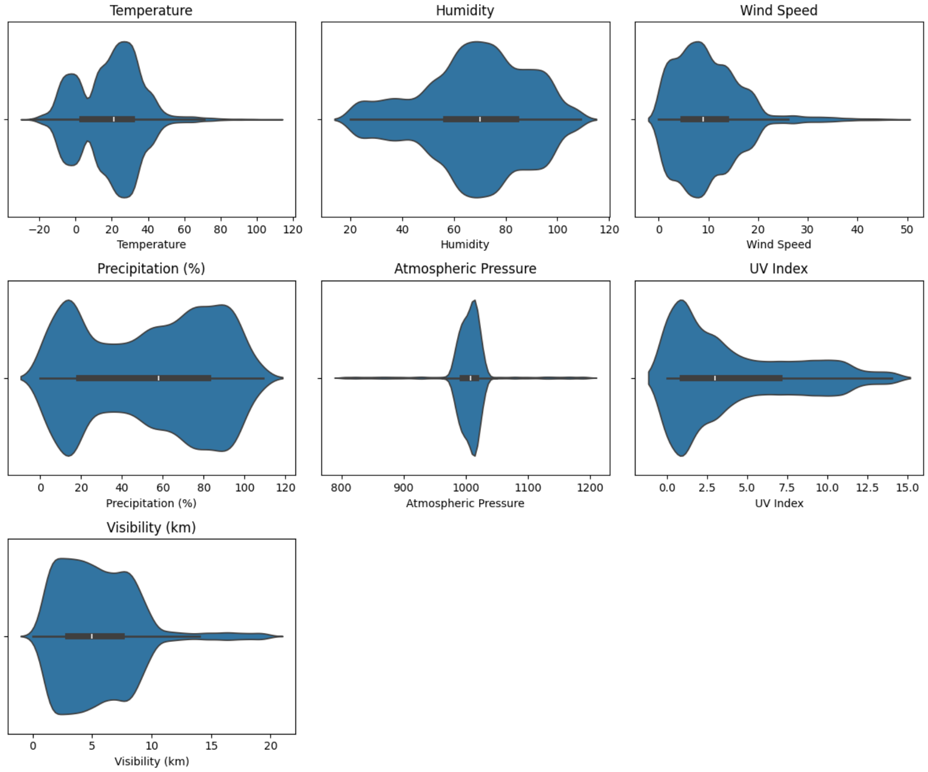


Figure 8: Weather Feature Distribtion (Violin plot)

The violin plots above illustrate the distribution of key weather-related variables. Each plot combines a box plot with a kernel density estimate to visualize both the central tendency and the spread of the data:

- **Temperature:** The distribution of temperature shows multiple peaks, suggesting the presence of distinct weather conditions. Most data points are concentrated between 0°C and 40°C, with extreme values extending beyond 100°C.
- **Humidity:** This feature exhibits a bimodal distribution, with data clustering around 40% and 80%, indicating varying humidity levels across different weather conditions.
- **Wind Speed:** The majority of wind speeds are concentrated below 20 km/h, while extreme wind speeds beyond 40 km/h are rare.
- **Precipitation (%):** The precipitation data is widely dispersed, with two main peaks near 0% and 100%, reflecting both dry and rainy conditions.
- **Atmospheric Pressure:** The distribution of atmospheric pressure is tightly centered around 1000 hPa, suggesting minimal variability in this parameter.
- **UV Index:** Most UV index values are concentrated between 0 and 10, with a peak near zero indicating many low-UV observations.

- **Visibility (km):** Visibility shows a bimodal distribution, with most observations concentrated between 0 and 10 km, reflecting clear and low-visibility conditions.

Overall, these violin plots highlight the distribution patterns, density, and presence of outliers for each weather-related feature, providing a comprehensive understanding of the dataset.

2.7 Feature Analysis

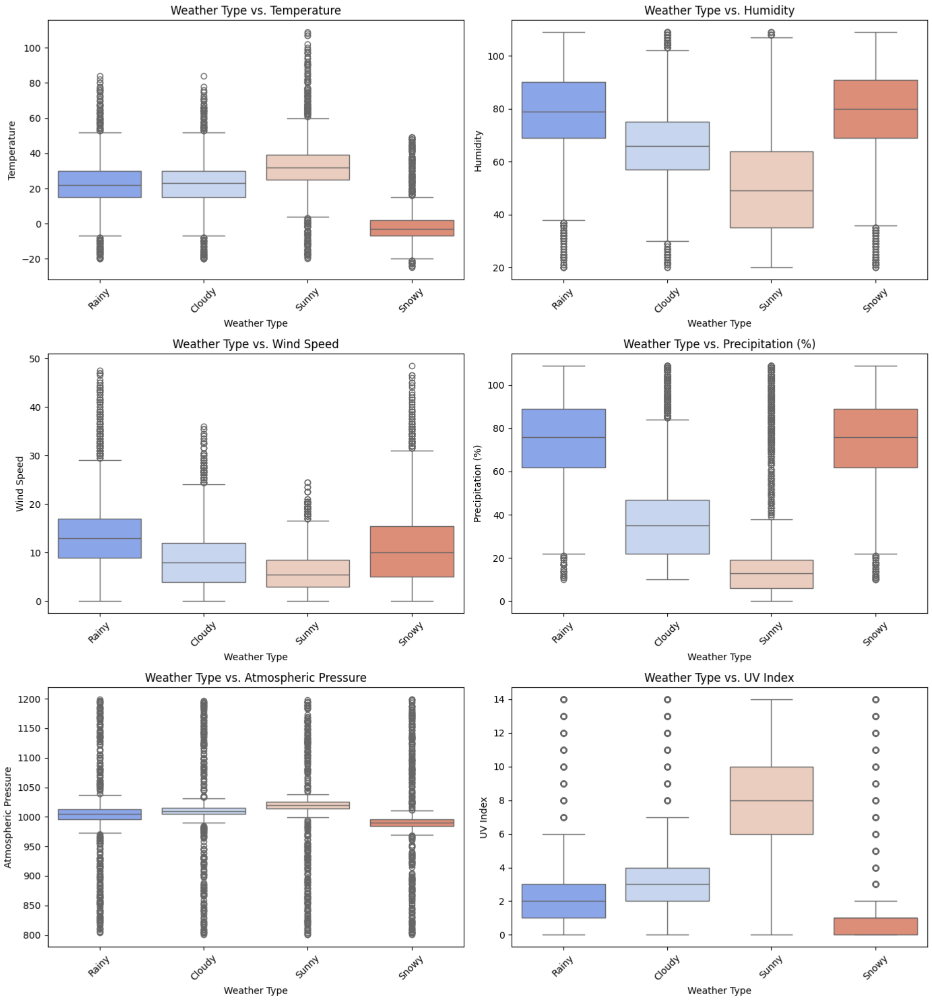


Figure 9: Weather Type vs Numerical Features (Boxplot)

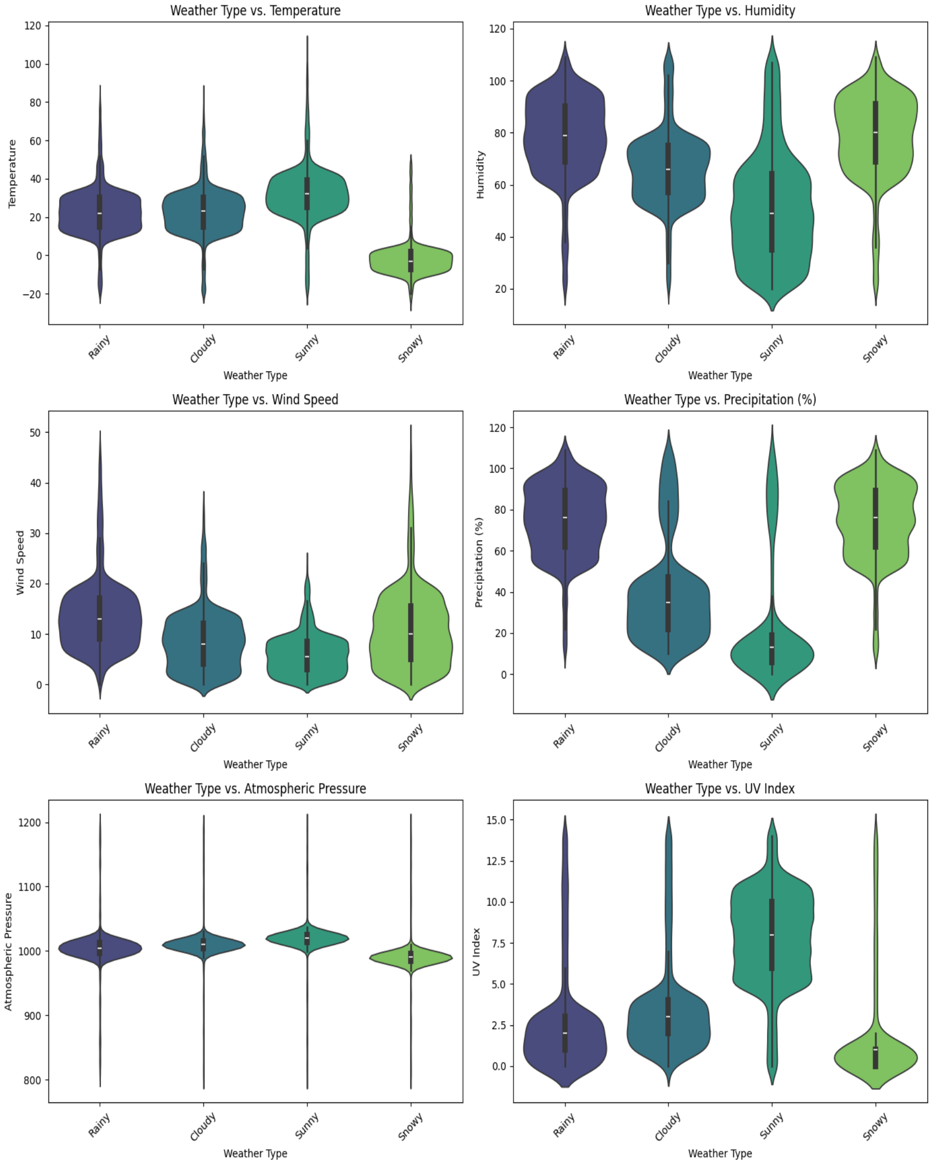


Figure 10: Weather Type vs Numerical Features (Violinplot)

From Figure 9, several observations can be made regarding the relationship between weather types and various numerical features:

- **Temperature:** Sunny weather exhibits the highest median temperature with a wider spread, while snowy conditions correspond to the lowest temperature range.
- **Humidity:** Rainy and snowy weather types are associated with consistently higher humidity levels, whereas sunny conditions display lower humidity.
- **Wind Speed:** Both rainy and snowy weather show higher wind speeds compared to sunny and cloudy weather, though all categories exhibit a large number of outliers.
- **Precipitation (%):** As expected, precipitation is significantly higher during rainy and snowy conditions, while sunny weather exhibits minimal precipitation.
- **Atmospheric Pressure:** Atmospheric pressure remains relatively stable across weather types, but rainy and cloudy conditions tend to show slightly lower values.

- **UV Index:** Sunny weather results in the highest UV index, while snowy conditions show the lowest values due to reduced sunlight exposure.
- **Visibility (km):** Visibility is notably reduced during rainy and snowy weather, while sunny and cloudy conditions provide better visibility.

These insights suggest distinct patterns across weather types, which may help in building predictive models for weather classification based on numerical features.

Figure **??** displays the distribution of key numerical weather features across different weather types (Rainy, Cloudy, Sunny, Snowy) using violin plots. The following observations can be drawn:

- **Temperature:**
  - Sunny weather shows the highest temperature range, typically between 20°C and 40°C.
  - Snowy conditions are associated with the lowest temperatures, with values mostly below 10°C.
  - Rainy and cloudy weather types exhibit moderate temperatures, centered around 20°C.
- **Humidity:**
  - Rainy and snowy conditions display consistently high humidity, with values concentrated between 80% and 100%.
  - Sunny weather shows a broader range but tends toward lower humidity levels.
  - Cloudy weather has intermediate humidity values, overlapping with both sunny and rainy distributions.
- **Wind Speed:**
  - Rainy and snowy weather types tend to have higher wind speeds, with values frequently exceeding 20 km/h.
  - Sunny and cloudy conditions exhibit lower wind speeds, typically below 20 km/h.
- **Precipitation (%):**
  - As expected, rainy and snowy weather conditions exhibit the highest precipitation values.
  - Sunny conditions show minimal precipitation, while cloudy weather has moderate levels but with a large variance.
- **Atmospheric Pressure:**
  - Atmospheric pressure remains relatively stable across weather types, with minor deviations.
  - Cloudy and rainy weather show slightly lower pressure on average compared to sunny weather.
- **UV Index:**
  - Sunny weather demonstrates the highest UV index, with values ranging from 5 to 12.
  - Snowy weather exhibits the lowest UV index, often remaining below 2.
  - Cloudy and rainy weather types show moderate UV index values with a similar distribution pattern.
- **Visibility (km):**
  - Visibility is reduced in rainy and snowy conditions, often falling below 5 km.
  - Cloudy weather displays intermediate visibility, while sunny conditions provide the best visibility, frequently exceeding 10 km.

These violin plots provide a comprehensive overview of how weather types influence key meteorological variables, highlighting clear patterns in temperature, humidity, wind speed, and other factors.

3 Data Pre-processing

For effective model training, the following pre-processing steps were performed:

- **Feature Scaling:** Numerical features were standardized to ensure all features contribute equally to model learning.
- **Categorical Encoding:** Categorical features were encoded to be used in machine learning models.
- **Data Splitting:** The dataset was split into training and testing sets using an 80-20 ratio to evaluate model performance on unseen data.

To handle outliers, we used the **Interquartile Range (IQR) method**, which focuses on the middle 50% of the data. The process is as follows:

1. **Calculate the first quartile (Q1, 25th percentile)** and **third quartile (Q3, 75th percentile)** for each numerical feature.
2. Compute the **Interquartile Range (IQR)**: IQR $= Q3 - Q1$.
3. Define outliers as any data point falling below $Q1 - 1.5 \times \text{IQR}$ or above $Q3 + 1.5 \times \text{IQR}$.
4. Remove these extreme values to ensure that our model is trained on reliable data without being influenced by extreme anomalies.

4 Model

- **Random Forest:** An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of individual trees. It is known for its robustness to overfitting and high accuracy on structured data.
- **Support Vector Machine (SVM):** A powerful supervised learning algorithm used for both classification and regression tasks. SVM works by finding the optimal hyperplane that best separates data points of different classes in a high-dimensional space. It is effective in high-dimensional spaces and suitable for complex decision boundaries.
- **Naive Bayes:** A probabilistic classifier based on Bayes' theorem with the assumption of conditional independence between features. It is computationally efficient and performs well with large datasets and high-dimensional data, especially in text classification.
- **Stochastic Gradient Descent (SGD) Classifier:** An iterative optimization algorithm for large-scale linear classification tasks. It is particularly useful for datasets with many samples due to its efficiency in handling large data and its adaptability to various loss functions.
- **Multi-layer Perceptron (MLP) Classifier:** A type of artificial neural network with multiple layers of nodes. MLP uses backpropagation for training and is capable of capturing complex relationships in data. It is effective for both classification and regression tasks, especially with non-linear data.

Model	**Hyperparameters**
Random Forest	{bootstrap: True, max_depth: 20, max_features: 'sqrt', min_samples_leaf: 2, min_samples_split: 5, n_estimators: 200}
SVM	{C: 10, degree: 2, gamma: 'auto', kernel: 'rbf'}
Naive Bayes	{var_smoothing: 0.001}
SGD Classifier	{alpha: 0.01, learning_rate: 'optimal', loss: 'modified_huber', max_iter: 1000, penalty: 'l1', tol: 0.001}
MLP Classifier	{activation: 'relu', alpha: 0.1, hidden_layer_sizes: (50, 50), learning_rate: 'constant', max_iter: 200, solver: 'adam'}

Table 2: Model Hyperparameters

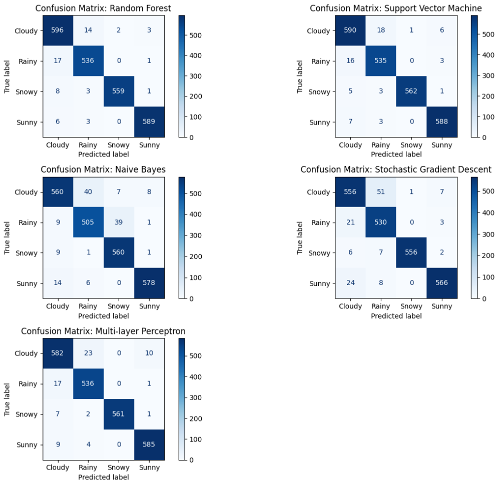
5 Result

5.1 Result

Model	Accuracy	Precision	F1-Score	Recall
Random Forest	0.9752	0.9754	0.9753	0.9752
SVM	0.9731	0.9732	0.9731	0.9731
Naive Bayes	0.9423	0.9427	0.9422	0.9423
SGD Classifier	0.9444	0.9459	0.9447	0.9444
MLP Classifier	0.9683	0.9686	0.9684	0.9683

Table 3: Model Performance Comparison (Accuracy, Precision, F1-Score, Recall)

The following confusion matrices display the performance of five different machine learning models: Random Forest, Support Vector Machine (SVM), Naive Bayes, Stochastic Gradient Descent (SGD), and Multi-layer Perceptron (MLP). Each matrix visualizes the predictions against the true labels, providing insight into the accuracy and misclassification patterns of each model.



Confusion Matrix: Random Forest

Confusion Matrix: Support Vector Machine

Confusion Matrix: Naive Bayes

Confusion Matrix: Stochastic Gradient Descent

Confusion Matrix: Multi-layer Perceptron

Figure 11: Enter Caption

5.2 Conclusion

Based on the performance metrics—**Accuracy**, **Precision**, **F1-Score**, and **Recall**—the **Random Forest** model demonstrates the best overall performance across all evaluated criteria. It achieves the highest accuracy (0.9752), along with balanced and superior precision (0.9754), F1-score (0.9753), and recall (0.9752), indicating that it consistently classifies the data with minimal errors.

5.3 Model Selection Consideration

Although the Random Forest model performs the best, the performance gap between all models is relatively small. Therefore, to optimize computational efficiency, I will select the model that requires the least resources.

Final Decision: Given its simplicity and lower computational cost, the **Naive Bayes** model is the most practical choice despite its slightly lower performance metrics. It is computationally efficient and suitable for large datasets, making it a better option when resource usage is a concern.