Weather Prediction

Project Report

Introduction

The world's weather is constantly and quickly changing. Today's society relies heavily on accurate forecasts. We significantly rely on weather forecasts for everything from agriculture to business, transport, and daily commute. In order to maintain simple and seamless movement as well as safe day-to-day operations, it is crucial to predict the weather accurately because the entire world is experiencing the effects of ongoing climate change.

Problem Statement

With data about various atmospheric conditions like humidity, precipitation, temperature, wind speed, etc., we need to find out the correlation between these weather attributes and their influence on the possible likely weather conditions like rain, snow, fog, sunlit etc., Essentially, using past data about weather conditions, we need to predict the most likely weather scenario using statistical and algorithmic models.

Motivation

Since we often listen to weather forecast news for local and regional long- or short-term weather predictions, there is a widespread and growing interest in weather information. Leading weather research organizations and businesses have been creating weather prediction systems that can identify, predict, and forecast weather threats and occurrences using cutting-edge scientific methods. This project aims to create a reasonably accurate prediction model with reduced computing power.

Methodology

- Data collection through various relevant and publicly available datasets in Kaggle
- Pre-processing of collected data for checking data quality and data cleaning by fixing or removing incorrect and corrupted data within the data set
- Building various training models using common statistical and algorithmic models like regression, decision trees and artificial neural networks
- Training the various models built and then testing their performance
- Evaluation of trained models and comparison with each other to identify the best model for the given dataset

A brief description of the various models used in this project has been given below:

Gaussian Naive Bayes Classifier

It is employed in numerous applications involving categorization. The "naive" assumption is the notion that the model's input variables are unrelated to one another and have unrelated distributions. If we alter the value of one feature, the algorithm's other characteristics won't be affected. Each class is presumed to follow a Gaussian distribution using Gaussian Naive Bayes.

Decision Tree

In this data is continually divided according to a certain parameter and represented by a tree structure. It is one of the most widely used machine learning algorithms and is used to resolve classification and regression tasks.

Random forest

It consists of several tree-structured classifiers whose outputs are combined to produce a single result. It consists of a collection of classifiers called decision trees and can be used for both classification and regression problems. The Random Forest Classifier is renowned for its ability to make precise predictions, flexibility, and lowered overfitting risk.

Methodology

Gradient boosting Classifier

It can be used for regression and classification tasks in machine learning; they are effective at classifying complex datasets and prediction accuracy is improved through the development of multiple models in succession, each of which aims to correct the errors of the previous one. Gradient Boosting classifiers combine many weak learning models, especially decision trees, to create a strong predictive model.

Logistic Regression

Modelling the likelihood of a discrete outcome given an input variable is what this technique entails. The most popular types of logistic regression provides a binary result, such as true or false, yes or no, and so on. Using multinomial logistic regression, events with more than two distinct possible outcomes can be modelled.

When attempting to establish which category a new sample most closely resembles, classification problems are a good place to employ logistic regression as an analysis technique. Logistic regression is a helpful analytical method since cyber security involves classification difficulties, such as attack detection.

K Nearest Neighbors Classifier

It is a non-parametric, supervised learning classifier that employs proximity to classify or anticipate how a particular data point will be grouped.

Extreme Gradient Boosting Classifier

A type of ensemble machine learning techniques known as "gradient boosting" can be applied to classification or regression-based predictive modelling issues. Decision tree models are the building blocks for ensembles.

Methodology

Stochastic Gradient Descent (SGD):
It is an effective method for fitting (linear) Support Vector Machines and logistic regression under convex loss functions.

SVM Classifier
It is a group of supervised learning techniques for classifying data, doing regression analysis, and identifying outliers.

ANN Classifier
As a function of the inputs, this classifier simply assigns an observation to a discrete class, by modelling the entire problem via a neural network, that is fitted to the dataset, by modifying weights in the network.

Results

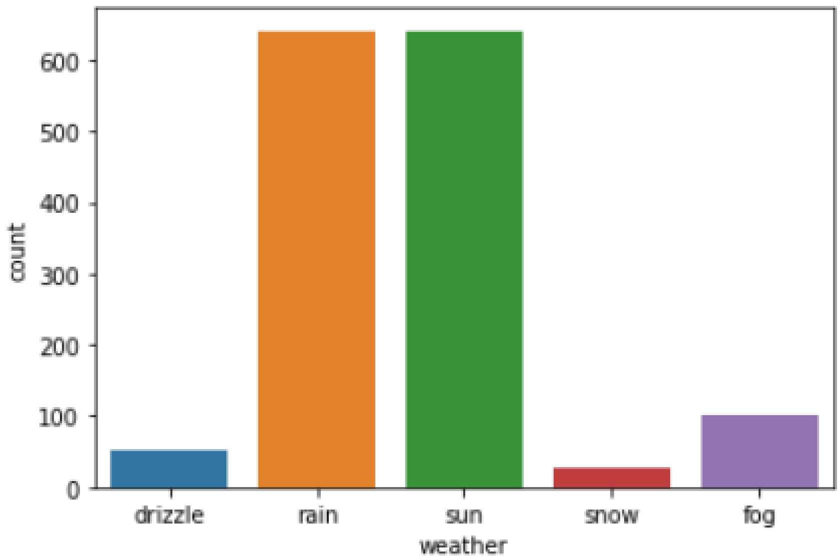
Data Exploration

	date	precipitation	temp_max	temp_min	wind	weather
0	2012-01-01	0.0	12.8	5.0	4.7	drizzle
1	2012-01-02	10.9	10.6	2.8	4.5	rain
2	2012-01-03	0.8	11.7	7.2	2.3	rain
3	2012-01-04	20.3	12.2	5.6	4.7	rain
4	2012-01-05	1.3	8.9	2.8	6.1	rain

Dataset Excerpt

	precipitation	temp_max	temp_min	wind
count	1461.000000	1461.000000	1461.000000	1461.000000
mean	3.029432	16.439083	8.234771	3.241136
std	6.680194	7.349758	5.023004	1.437825
min	0.000000	-1.600000	-7.100000	0.400000
25%	0.000000	10.600000	4.400000	2.200000
50%	0.000000	15.600000	8.300000	3.000000
75%	2.800000	22.200000	12.200000	4.000000
max	55.900000	35.600000	18.300000	9.500000

Numerical Data Properties




```
rain       641
sun        640
fog        101
drizzle     53
snow        26
Name: weather, dtype: int64
```

Target Variable Count in Dataset

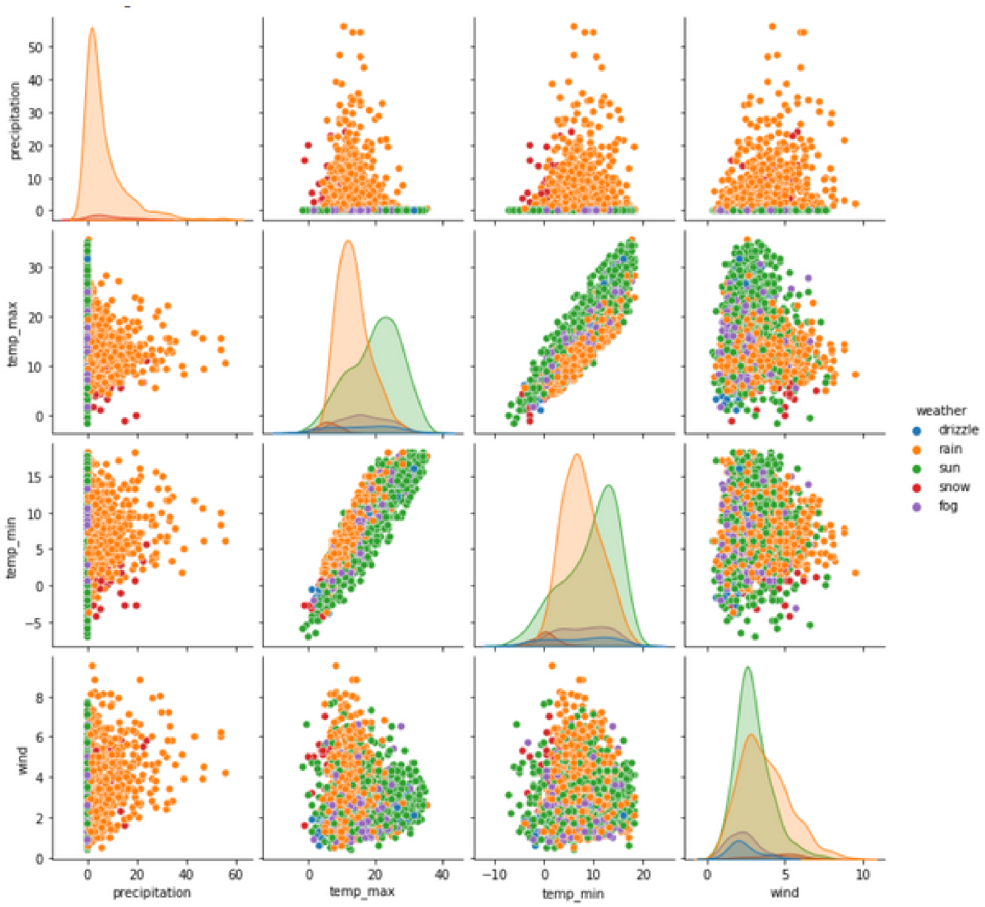
Results

Data Exploration

Check for missing values:- Data records with missing values were checked and removed to have consistent data.

```
date             0
precipitation    0
temp_max         0
temp_min         0
wind             0
weather          0
dtype: int64
```

Missing Value Check Results



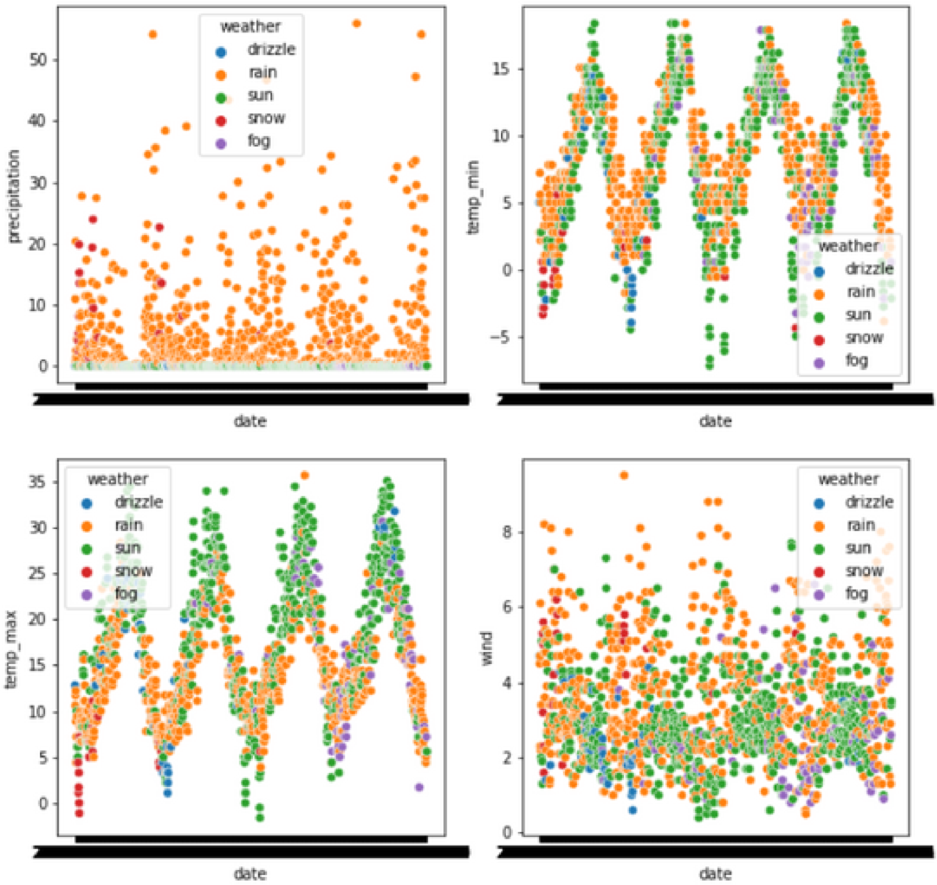
Pairplot between data attributes

Pairplot based on weather type has been generated to know the statistics of each weather type, which gives an idea about the pattern trends as well as shows the outliers in the data.

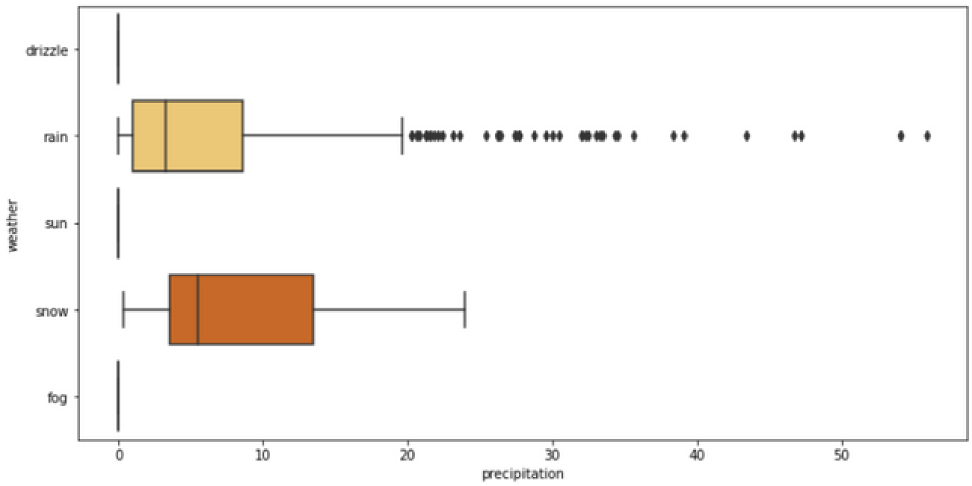
Results

Data Exploration

Similarly, a Scatterplot of date vs weather type is plotted to understand the weather type and its trend across different days.



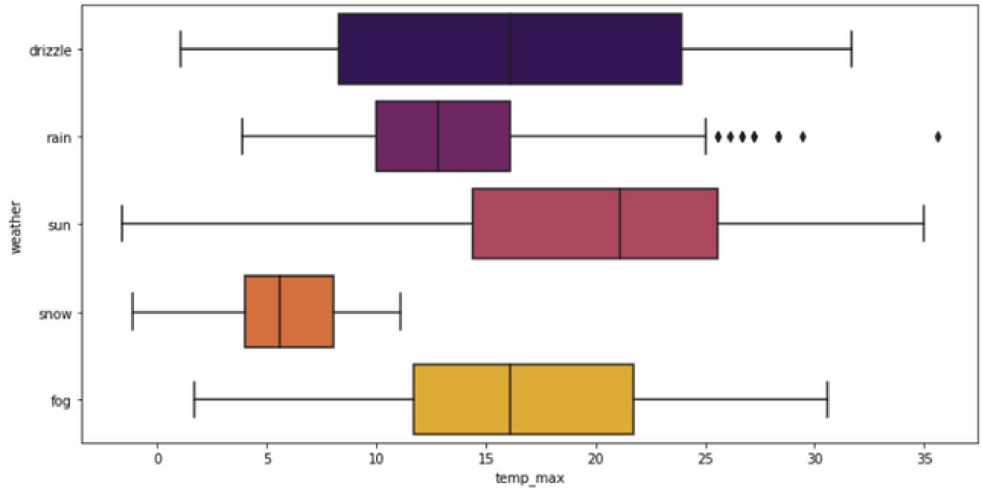
Scatterplot between Data Atributes



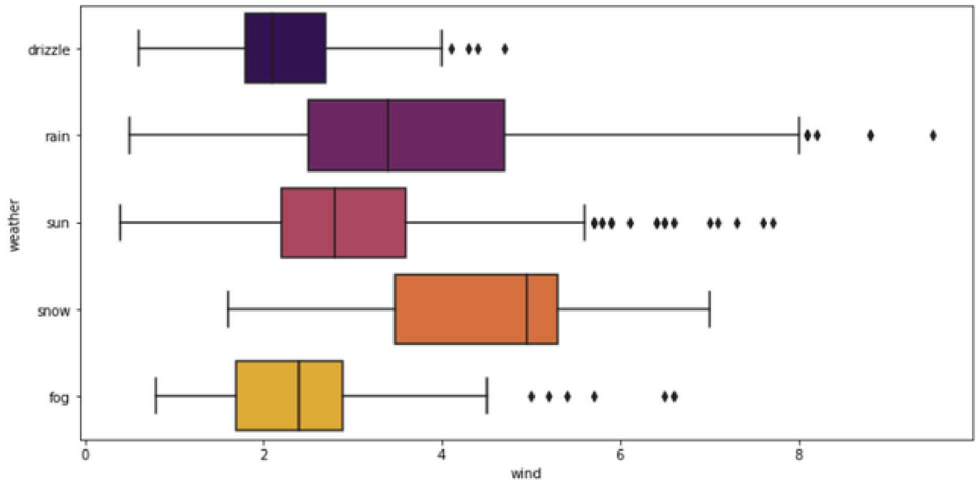
Boxplot between Weather and Precipitation

Results

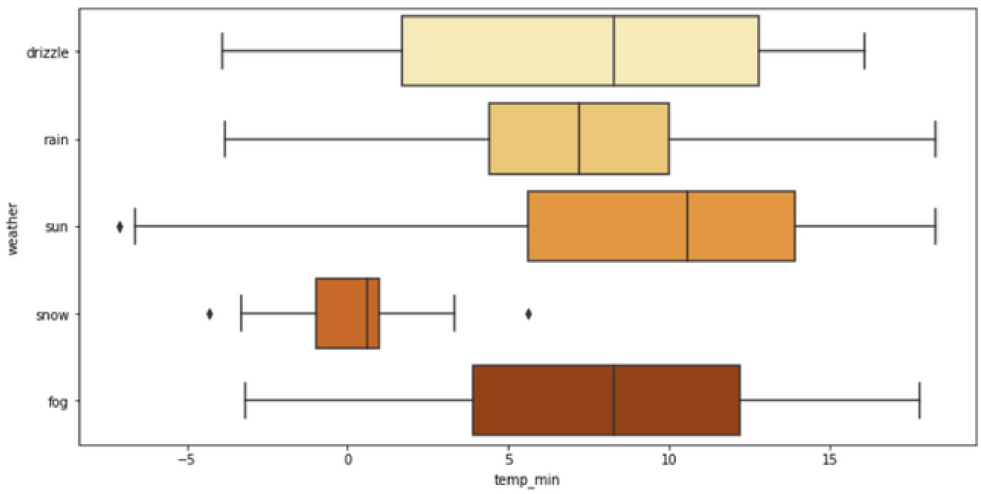
Data Exploration



Boxplot between Weather and Max Temperature



Boxplot between Weather and Wind

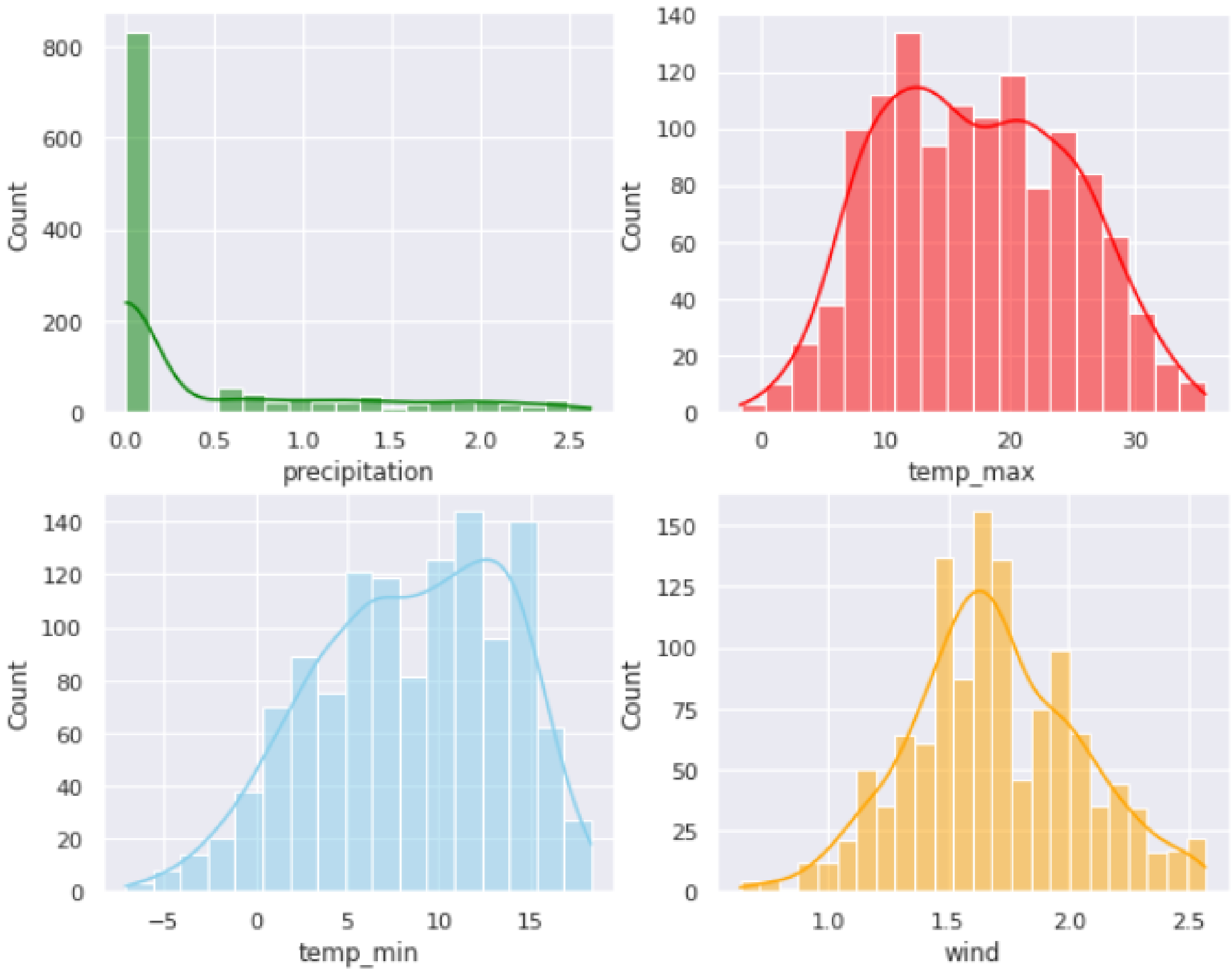


Boxplot between Weather and Min Temperature

Results

Data PreProcessing

Shorter boxplots show the high level of wind and precipitation compared to larger boxplots. The variability in the sizes of boxplots indicates the impact of temperatures, wind and precipitation on different weather conditions.

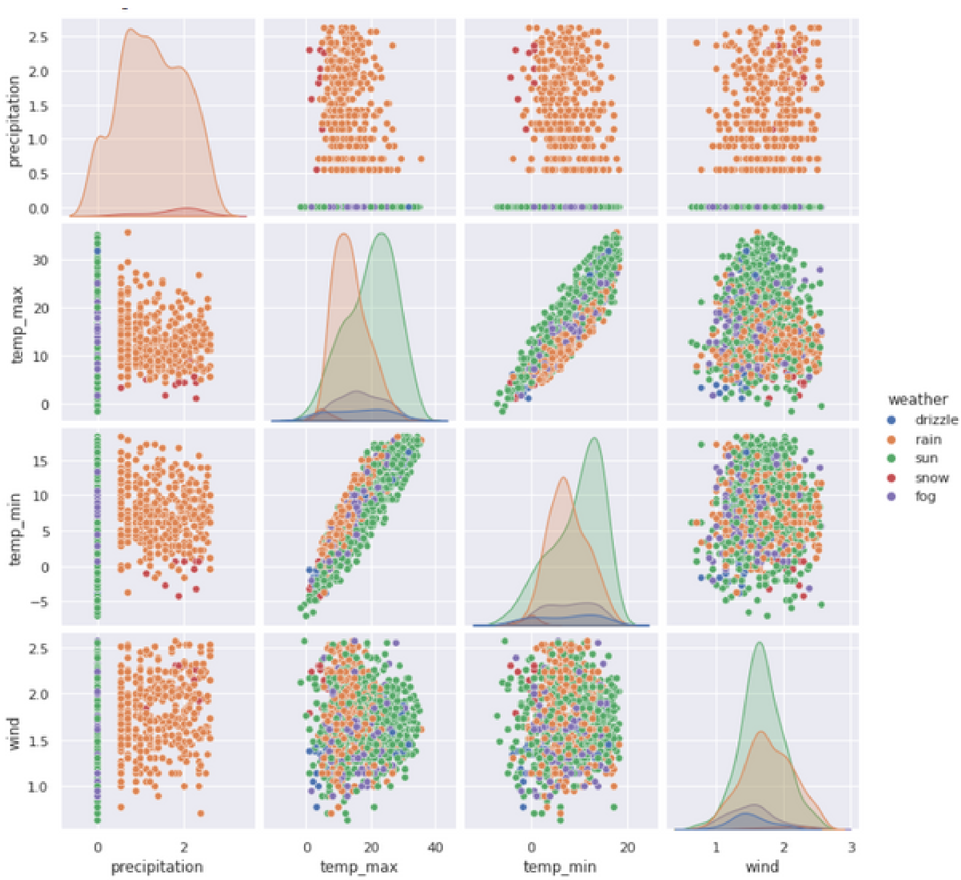


Histogram Plot of Data Attributes

The histogram plot shows the data distribution, after treating the data values of wind and precipitation for skewness and removal of outliers from the entire dataset, in order to lessen the effect of bias towards these data attributes while training the models.

Results

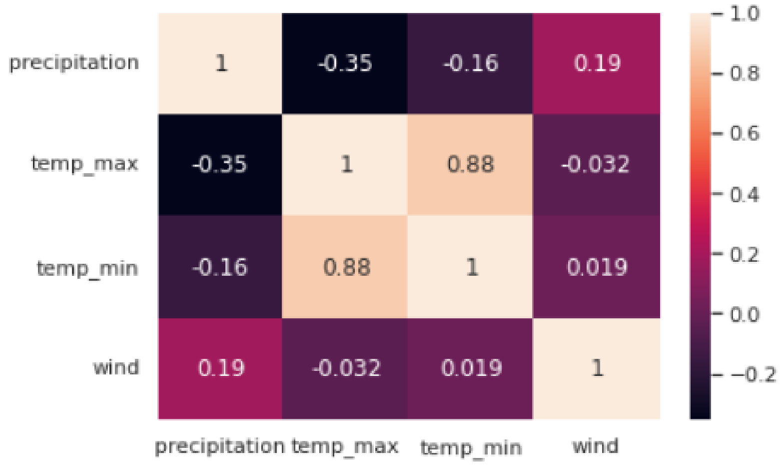
Data PreProcessing



Pairplot between data attributes

Pairplot grid after removing outliers, shows a better distributed spread of the data across parameters like precipitation, temperature and wind.

The hues or intensities and the corresponding data values in this heat map shows the corresponding correlation between any two dimensions of precipitation, temperature and wind.



Heatmap for Correlation between Data Attributes

The data has been then encoded using categorical encoding to make the dataset more suitable for training the models used for further classification.

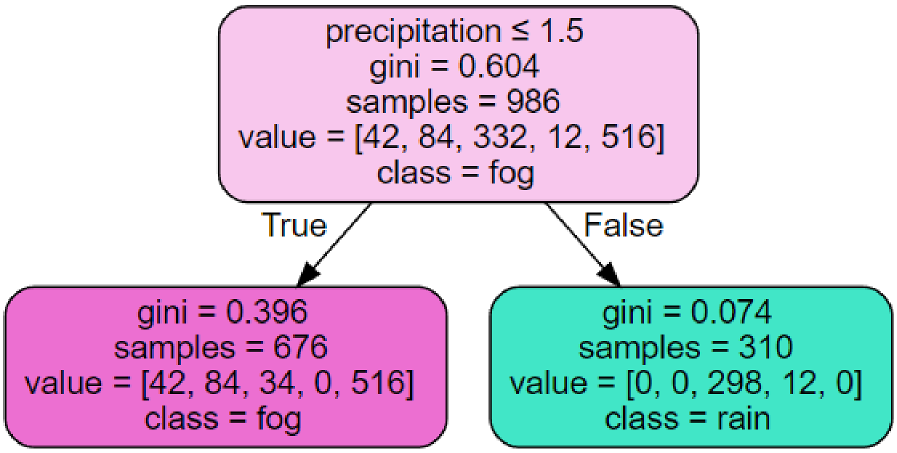
Results

Model Building and Testing

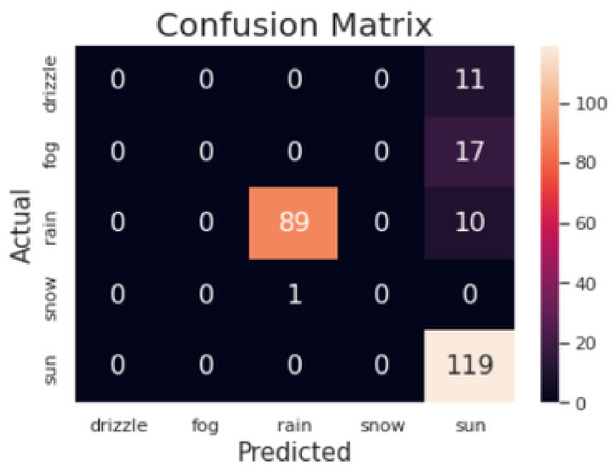
After the data has been cleaned and suitably preprocessed, it has been split into training and testing sets with 80% for the former and 20% in the latter. The same training and testing set has been used to train and test all the 11 models used in this project in order to reduce any random bias arising from using different training and testing sets across different models.

The figures below, show the confusion matrix, reports and various evaluation metrics of each of the models used to train and test the preprocessed dataset.

Decision Tree Classifier



Decision Tree Visualization



	precision	recall	f1-score	support
drizzle	0.00	0.00	0.00	11
fog	0.00	0.00	0.00	17
rain	0.99	0.90	0.94	99
snow	0.00	0.00	0.00	1
sun	0.76	1.00	0.86	119
accuracy			0.84	247
macro avg	0.35	0.38	0.36	247
weighted avg	0.76	0.84	0.79	247

Results

Model Building and Testing

Accuracy:

	0
drizzle	0.955466
fog	0.931174
rain	0.955466
snow	0.995951
sun	0.846154

Sensitivity:

	0
drizzle	0.00000
fog	0.00000
rain	0.89899
snow	0.00000
sun	1.00000

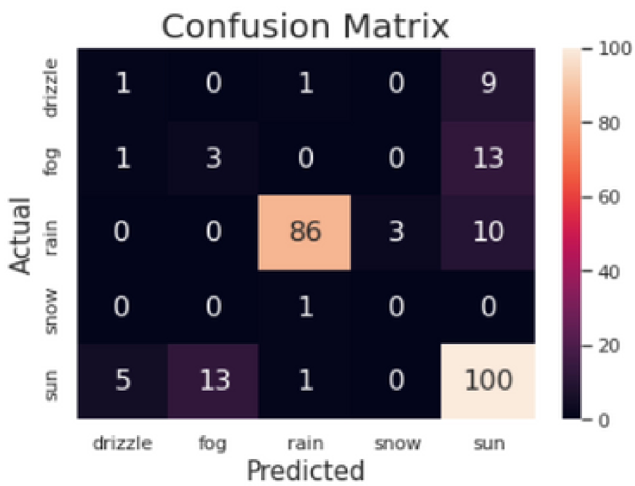
Error:

	0
drizzle	0.044534
fog	0.068826
rain	0.044534
snow	0.004049
sun	0.153846

Specificity:

	0
drizzle	1.000000
fog	1.000000
rain	0.993243
snow	1.000000
sun	0.703125

Bagging Classifier



	precision	recall	f1-score	support
drizzle	0.14	0.09	0.11	11
fog	0.19	0.18	0.18	17
rain	0.97	0.87	0.91	99
snow	0.00	0.00	0.00	1
sun	0.76	0.84	0.80	119
accuracy			0.77	247
macro avg	0.41	0.40	0.40	247
ighted avg	0.77	0.77	0.77	247

Accuracy:

	0
drizzle	0.935223
fog	0.890688
rain	0.935223
snow	0.983806
sun	0.793522

Sensitivity:

	0
drizzle	0.090909
fog	0.176471
rain	0.868687
snow	0.000000
sun	0.840336

Error:

	0
drizzle	0.064777
fog	0.109312
rain	0.064777
snow	0.016194
sun	0.206478

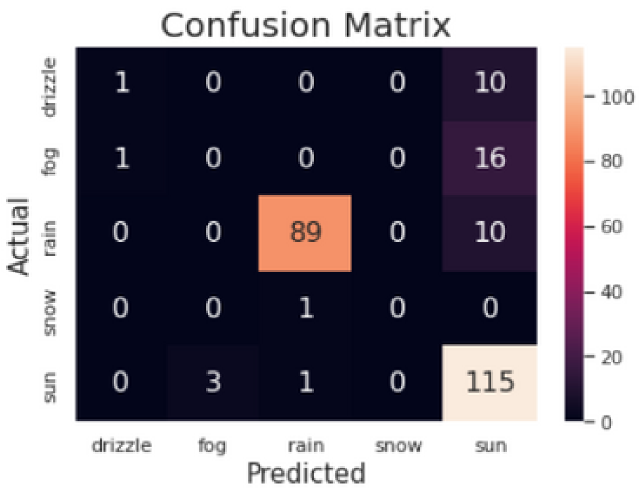
Specificity:

	0
drizzle	0.974576
fog	0.943478
rain	0.979730
snow	0.987805
sun	0.750000

Results

Model Building and Testing

Gradient Boosting Classifier



	precision	recall	f1-score	support
drizzle	0.50	0.09	0.15	11
fog	0.00	0.00	0.00	17
rain	0.98	0.90	0.94	99
snow	0.00	0.00	0.00	1
sun	0.76	0.97	0.85	119
accuracy			0.83	247
macro avg	0.45	0.39	0.39	247
weighted avg	0.78	0.83	0.79	247

Accuracy:

	0
drizzle	0.955466
fog	0.919028
rain	0.951417
snow	0.995951
sun	0.838057

Sensitivity:

	0
drizzle	0.090909
fog	0.000000
rain	0.898990
snow	0.000000
sun	0.966387

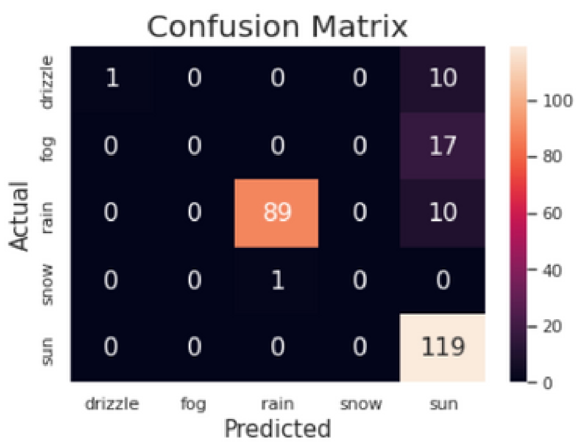
Error:

	0
drizzle	0.044534
fog	0.080972
rain	0.048583
snow	0.004049
sun	0.161943

Specificity:

	0
drizzle	0.995763
fog	0.986957
rain	0.986486
snow	1.000000
sun	0.718750

Extreme Gradient Boosting Classifier



	precision	recall	f1-score	support
drizzle	1.00	0.09	0.17	11
fog	0.00	0.00	0.00	17
rain	0.99	0.90	0.94	99
snow	0.00	0.00	0.00	1
sun	0.76	1.00	0.87	119
accuracy			0.85	247
macro avg	0.55	0.40	0.39	247
weighted avg	0.81	0.85	0.80	247

Results

Model Building and Testing

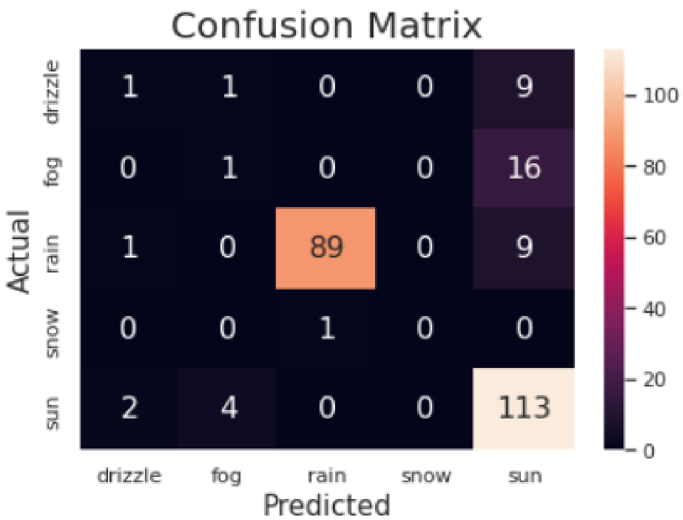
```
Accuracy:
                0
drizzle  0.959514
fog      0.931174
rain     0.955466
snow     0.995951
sun      0.850202
```

```
Sensitivity:
                0
drizzle  0.090909
fog      0.000000
rain     0.898990
snow     0.000000
sun      1.000000
```

```
Error:
                0
drizzle  0.040486
fog      0.068826
rain     0.044534
snow     0.004049
sun      0.149798
```

```
Specificity:
                0
drizzle  1.000000
fog      1.000000
rain     0.993243
snow     1.000000
sun      0.710938
```

Random Forest Classifier



	precision	recall	f1-score	support
drizzle	0.25	0.09	0.13	11
fog	0.17	0.06	0.09	17
rain	0.99	0.90	0.94	99
snow	0.00	0.00	0.00	1
sun	0.77	0.95	0.85	119
accuracy			0.83	247
macro avg	0.43	0.40	0.40	247
weighted avg	0.79	0.83	0.80	247

```
Accuracy:
                0
drizzle  0.947368
fog      0.914980
rain     0.955466
snow     0.995951
sun      0.838057
```

```
Sensitivity:
                0
drizzle  0.090909
fog      0.058824
rain     0.898990
snow     0.000000
sun      0.949580
```

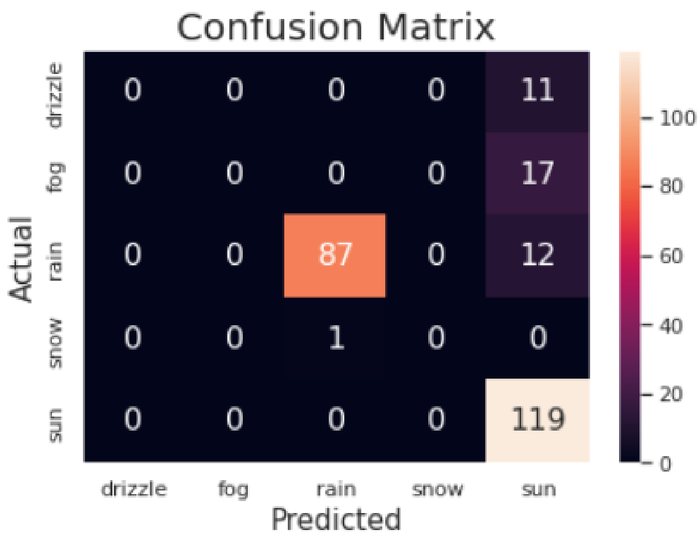
```
Error:
                0
drizzle  0.052632
fog      0.085020
rain     0.044534
snow     0.004049
sun      0.161943
```

```
Specificity:
                0
drizzle  0.987288
fog      0.978261
rain     0.993243
snow     1.000000
sun      0.734375
```

Results

Model Building and Testing

Logistic Regression Classifier



	precision	recall	f1-score	support
drizzle	0.00	0.00	0.00	11
fog	0.00	0.00	0.00	17
rain	0.99	0.88	0.93	99
snow	0.00	0.00	0.00	1
sun	0.75	1.00	0.86	119
accuracy			0.83	247
macro avg	0.35	0.38	0.36	247
weighted avg	0.76	0.83	0.79	247

Accuracy:

	0
drizzle	0.955466
fog	0.931174
rain	0.947368
snow	0.995951
sun	0.838057

Sensitivity:

	0
drizzle	0.000000
fog	0.000000
rain	0.878788
snow	0.000000
sun	1.000000

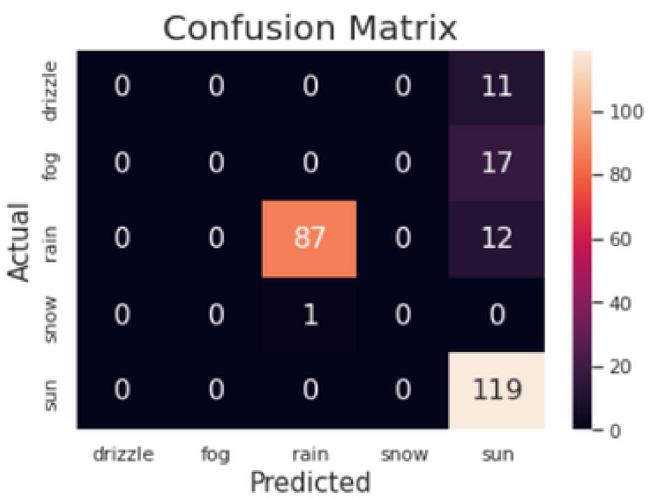
Error:

	0
drizzle	0.044534
fog	0.068826
rain	0.052632
snow	0.004049
sun	0.161943

Specificity:

	0
drizzle	1.000000
fog	1.000000
rain	0.993243
snow	1.000000
sun	0.687500

Stochastic Gradient Descent Classifier



	precision	recall	f1-score	support
drizzle	0.00	0.00	0.00	11
fog	0.00	0.00	0.00	17
rain	0.99	0.88	0.93	99
snow	0.00	0.00	0.00	1
sun	0.75	1.00	0.86	119
accuracy			0.83	247
macro avg	0.35	0.38	0.36	247
weighted avg	0.76	0.83	0.79	247

Results

Model Building and Testing

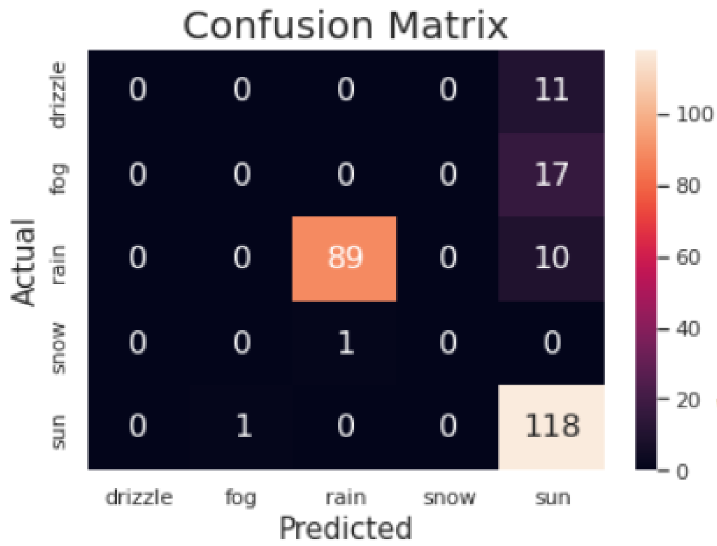
```
Accuracy:
                0
drizzle  0.955466
fog      0.927126
rain     0.947368
snow     0.995951
sun      0.834008
```

```
Sensitivity:
                0
drizzle  0.000000
fog      0.000000
rain     0.878788
snow     0.000000
sun      0.991597
```

```
Error:
                0
drizzle  0.044534
fog      0.072874
rain     0.052632
snow     0.004049
sun      0.165992
```

```
Specificity:
                0
drizzle  1.000000
fog      0.995652
rain     0.993243
snow     1.000000
sun      0.687500
```

Gaussian Naive Bayes Classifier



	precision	recall	f1-score	support
drizzle	0.00	0.00	0.00	11
fog	0.00	0.00	0.00	17
rain	0.99	0.90	0.94	99
snow	0.00	0.00	0.00	1
sun	0.76	0.99	0.86	119
accuracy			0.84	247
macro avg	0.35	0.38	0.36	247
weighted avg	0.76	0.84	0.79	247

```
Accuracy:
                0
drizzle  0.955466
fog      0.927126
rain     0.955466
snow     0.995951
sun      0.842105
```

```
Sensitivity:
                0
drizzle  0.000000
fog      0.000000
rain     0.898990
snow     0.000000
sun      0.991597
```

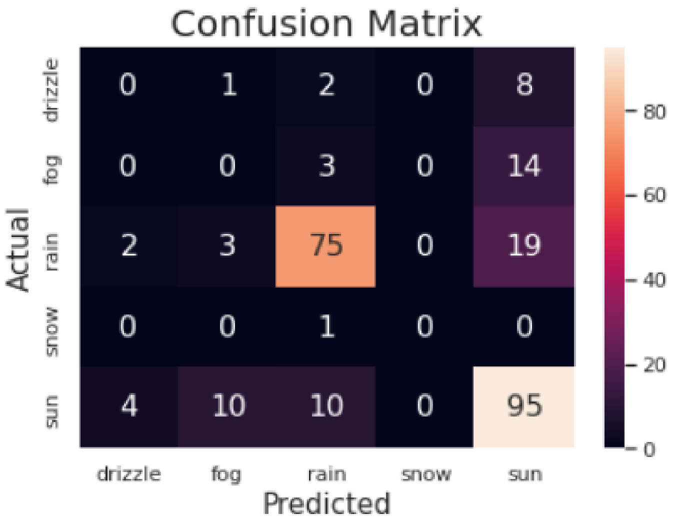
```
Error:
                0
drizzle  0.044534
fog      0.072874
rain     0.044534
snow     0.004049
sun      0.157895
```

```
Specificity:
                0
drizzle  1.000000
fog      0.995652
rain     0.993243
snow     1.000000
sun      0.703125
```

Results

Model Building and Testing

K Nearest Neighbors Classifier



	precision	recall	f1-score	support
drizzle	0.00	0.00	0.00	11
fog	0.00	0.00	0.00	17
rain	0.82	0.76	0.79	99
snow	0.00	0.00	0.00	1
sun	0.70	0.80	0.75	119
accuracy			0.69	247
macro avg	0.30	0.31	0.31	247
weighted avg	0.67	0.69	0.68	247

Accuracy:

	0
drizzle	0.931174
fog	0.874494
rain	0.838057
snow	0.995951
sun	0.736842

Sensitivity:

	0
drizzle	0.000000
fog	0.000000
rain	0.757576
snow	0.000000
sun	0.798319

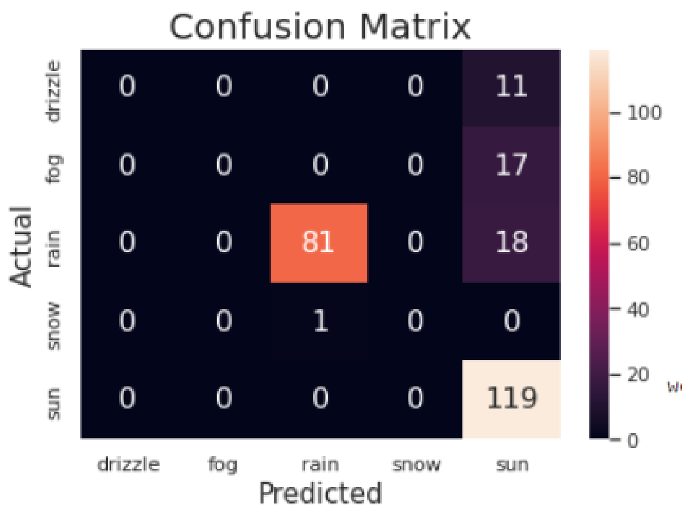
Error:

	0
drizzle	0.068826
fog	0.125506
rain	0.161943
snow	0.004049
sun	0.263158

Specificity:

	0
drizzle	0.974576
fog	0.939130
rain	0.891892
snow	1.000000
sun	0.679688

Support Vector Machine Classifier



	precision	recall	f1-score	support
drizzle	0.00	0.00	0.00	11
fog	0.00	0.00	0.00	17
rain	0.99	0.82	0.90	99
snow	0.00	0.00	0.00	1
sun	0.72	1.00	0.84	119
accuracy			0.81	247
macro avg	0.34	0.36	0.35	247
weighted avg	0.74	0.81	0.76	247

Results

Model Building and Testing

Accuracy:

	0
drizzle	0.955466
fog	0.931174
rain	0.923077
snow	0.995951
sun	0.813765

Sensitivity:

	0
drizzle	0.000000
fog	0.000000
rain	0.818182
snow	0.000000
sun	1.000000

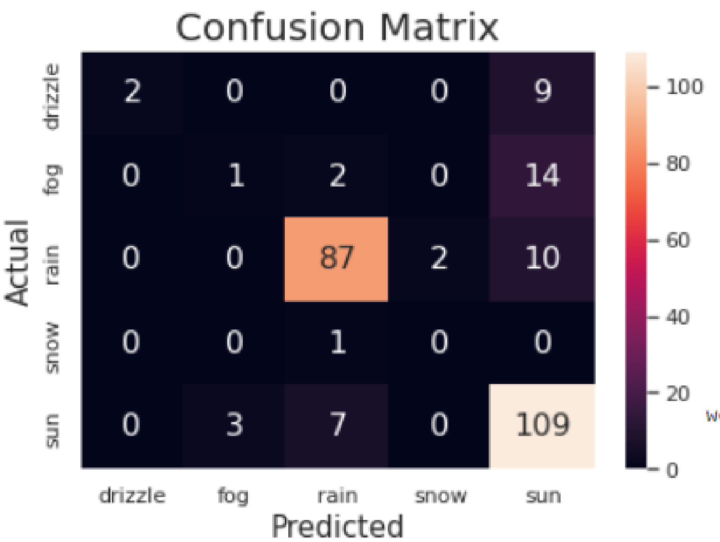
Error:

	0
drizzle	0.044534
fog	0.068826
rain	0.076923
snow	0.004049
sun	0.186235

Specificity:

	0
drizzle	1.000000
fog	1.000000
rain	0.993243
snow	1.000000
sun	0.640625

ANN Classifier



	precision	recall	f1-score	support
drizzle	1.00	0.18	0.31	11
fog	0.25	0.06	0.10	17
rain	0.90	0.88	0.89	99
snow	0.00	0.00	0.00	1
sun	0.77	0.92	0.84	119
accuracy			0.81	247
macro avg	0.58	0.41	0.43	247
weighted avg	0.79	0.81	0.78	247

Accuracy:

	0
drizzle	0.963563
fog	0.923077
rain	0.910931
snow	0.987854
sun	0.825911

Sensitivity:

	0
drizzle	0.181818
fog	0.058824
rain	0.878788
snow	0.000000
sun	0.915966

Error:

	0
drizzle	0.036437
fog	0.076923
rain	0.089069
snow	0.012146
sun	0.174089

Specificity:

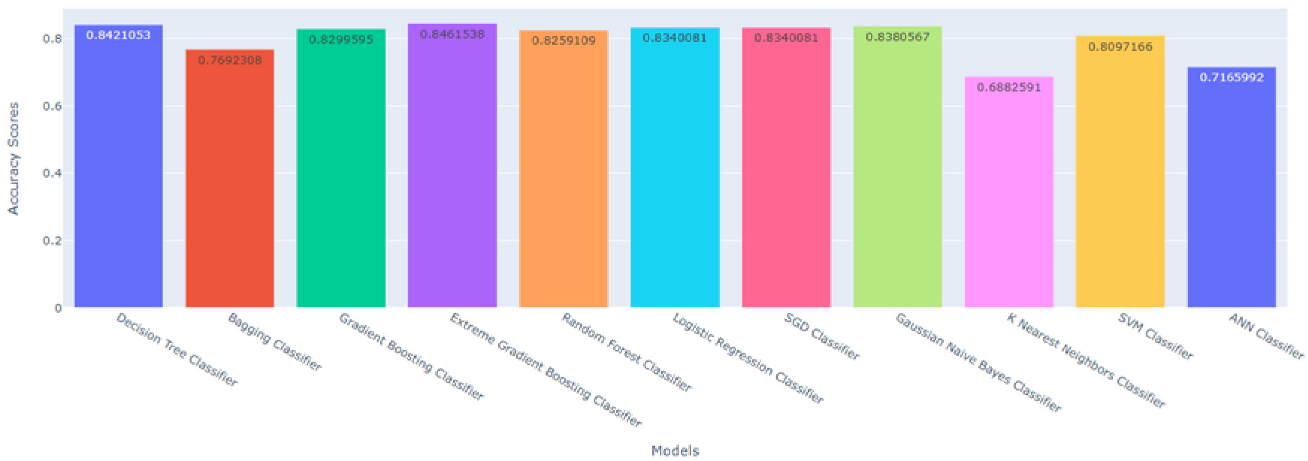
	0
drizzle	1.000000
fog	0.986957
rain	0.932432
snow	0.991870
sun	0.742188

Analysis

Before modeling the data, missing values and outliers need to be removed in order to get representative outcomes. In pairplot and scatter plot grids, outliers can be seen and hence, they have been removed as a result of pre-processing. The cleaned data is then classified into different classes in the model training process, using temperature, precipitation and wind as the significant criteria of classification. The results obtained from each of the classification models used are outlined below.

In the decision tree, out of 986 samples, 676 samples were identified as samples having precipitation values less than 1.5 with 0.396 as gini value. 310 samples were identified to have precipitation more than 1.5 with gini value 0.074. The confusion matrix shows that the sun has more number of true positives, i.e.119 followed by rain, which has 89 true positives, indicating that the actual and predicted number of sunny days by the decision tree model is 119 and rainy days turned out to be 89. However, the accuracy rate for the snowy weather is found to be the highest.

Similarly, for all the other classification models used, the predicted class of sun has the highest number of true positives, followed by the rain class. However, for all the models, the accuracy rate of snowy weather class is found to be the highest. This might be probably because the samples for the snow class are the least in number in the dataset, thereby reducing the complexity in prediction of the class for all the models.



Analysis

With respect to the 11 models used for training and testing on the given dataset, the **Extreme Gradient Boosting Classifier** has given the best performance, in terms of accuracy score, followed closely by the Decision Tree Classifier, whereas the least performance within the 11 models used is exhibited by the K Nearest Neighbors Classifier. Surprisingly, a robust algorithm like ANN has not been found to perform as well as significantly lesser complex models, like the Decision Tree Classifier, thereby indicating that more complex non-linear models may always not be providing the best performance across different samples of the same dataset.

Summary

With the advancement of science and technology, people have inculcated the practice of using sophisticated models to predict and forecast accurate weather events repeatedly with minimum deviation.

In this project, data has been pre-processed post which models have been built and after which the models have underwent training and testing using the training and testing data sets respectively. The different models such as decision tree, Bagging classifier, Gradient Boosting Classifier, Extreme Gradient Boosting classifier, Random Forest, Logistic Regression, Stochastic Gradient Descent, Gaussian Naive Bayes, K Nearest Neighbors, Support Vector Machine Classifier, ANN were built and their performance has been compared, of which the Extreme Gradient Boosting Classifier has been found to be the best fit, pertaining to our weather dataset.