RAIN PREDICTION IN AUSTRALIA: A MACHINE LEARNING APPROACH

CS – 513 A

Group No. 7

Sarthak Achyut Vaidya
20016184

Piyush Devendra Kataktalware
20022156

Sai Venkata Subbaraya Akhil Pulipaka
20012406

CONTENTS

PROBLEM STATEMENT

- Australia's vast and varied climatic conditions present a significant challenge in predicting rainfall accurately.
- The unpredictability of rain patterns affects agriculture, urban planning, and environmental sustainability.
- Our goal is to leverage machine learning models to improve the accuracy of rain predictions, aiding in effective planning and resource management across different Australian regions.

BRIEF IDEA OF “RAIN PREDICTION” CLASSIFICATION MODEL

This classification model, uses various features to provide accurate and timely forecasts to inform decision-making in agriculture, urban planning, and disaster management.

This information can be used to help in crop planning and irrigation scheduling to optimize agricultural output.

Rain prediction models aids in designing effective drainage systems, flood mitigation strategies, and water conservation measures. It enables early warnings for floods and storms, allowing for evacuation and strategic deployment of resources. It also supports the preservation of ecosystems by anticipating and managing changes in water availability.

ABOUT OUR DATA

1. Date
2. Location
3. Min Temp
4. Max Temp
5. Rainfall
6. Evaporation
7. Sunshine
8. Wind Gust Dir
9. Wind Gust Speed
10. WindDir9am
11. WindDir3pm
12. WindSpeed9am
13. WindSpeed3pm
14. Humidity9am
15. Humidity3pm
16. Pressure9am
17. Pressure3pm
18. Cloud9am
19. Cloud3pm
20. Temp9am
21. Temp3pm
22. Rain Today
23. Rain Tomorrow

Raw data was taken from:

https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package/data

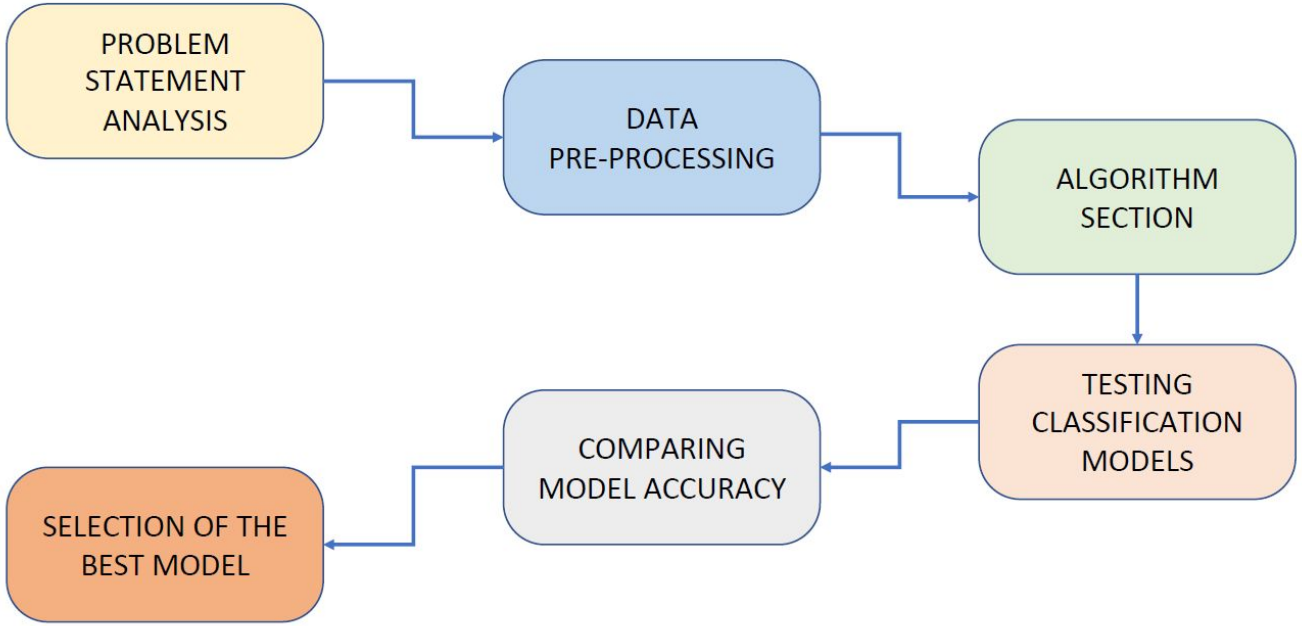
Sample of the used dataset

1	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporati	Sunshine	WindGust	WindGust	WindDir9a	WindDir3p	WindSpee	WindSpee	Humidity9	Humidity3	Pressure9	Pressure3	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
2	01-12-2008	Albury	13.4	22.9	0.6	NA	NA	W	44	W	WNW	20	24	71	22	1007.7	1007.1	8	NA	16.9	21.8	No	No
3	02-12-2008	Albury	7.4	25.1	0	NA	NA	WNW	44	NNW	WSW	4	22	44	25	1010.6	1007.8	NA	NA	17.2	24.3	No	No
4	03-12-2008	Albury	12.9	25.7	0	NA	NA	WSW	46	W	WSW	19	26	38	30	1007.6	1008.7	NA	2	21	23.2	No	No
5	04-12-2008	Albury	9.2	28	0	NA	NA	NE	24	SE	E	11	9	45	16	1017.6	1012.8	NA	NA	18.1	26.5	No	No
6	05-12-2008	Albury	17.5	32.3	1	NA	NA	W	41	ENE	NW	7	20	82	33	1010.8	1006	7	8	17.8	29.7	No	No
7	06-12-2008	Albury	14.6	29.7	0.2	NA	NA	WNW	56	W	W	19	24	55	23	1009.2	1005.4	NA	NA	20.6	28.9	No	No
8	07-12-2008	Albury	14.3	25	0	NA	NA	W	50	SW	W	20	24	49	19	1009.6	1008.2	1	NA	18.1	24.6	No	No
9	08-12-2008	Albury	7.7	26.7	0	NA	NA	W	35	SSE	W	6	17	48	19	1013.4	1010.1	NA	NA	16.3	25.5	No	No
10	09-12-2008	Albury	9.7	31.9	0	NA	NA	NNW	80	SE	NW	7	28	42	9	1008.9	1003.6	NA	NA	18.3	30.2	No	Yes
11	10-12-2008	Albury	13.1	30.1	1.4	NA	NA	W	28	S	SSE	15	11	58	27	1007	1005.7	NA	NA	20.1	28.2	Yes	No
12	11-12-2008	Albury	13.4	30.4	0	NA	NA	N	30	SSE	ESE	17	6	48	22	1011.8	1008.7	NA	NA	20.4	28.8	No	Yes
13	12-12-2008	Albury	15.9	21.7	2.2	NA	NA	NNE	31	NE	ENE	15	13	89	91	1010.5	1004.2	8	8	15.9	17	Yes	Yes
14	13-12-2008	Albury	15.9	18.6	15.6	NA	NA	W	61	NNW	NNW	28	28	76	93	994.3	993	8	8	17.4	15.8	Yes	Yes
15	14-12-2008	Albury	12.6	21	3.6	NA	NA	SW	44	W	SSW	24	20	65	43	1001.2	1001.8	NA	7	15.8	19.8	Yes	No
16	15-12-2008	Albury	8.4	24.6	0	NA	NA	NA	NA	S	WNW	4	30	57	32	1009.7	1008.7	NA	NA	15.9	23.5	No	NA
17	16-12-2008	Albury	9.8	27.7	NA	NA	NA	WNW	50	NA	WNW	NA	22	50	28	1013.4	1010.3	0	NA	17.3	26.2	NA	No
18	17-12-2008	Albury	14.1	20.9	0	NA	NA	ENE	22	SSW	E	11	9	69	82	1012.2	1010.4	8	1	17.2	18.1	No	Yes
19	18-12-2008	Albury	13.5	22.9	16.8	NA	NA	W	63	N	WNW	6	20	80	65	1005.8	1002.2	8	1	18	21.5	Yes	Yes
20	19-12-2008	Albury	11.2	22.5	10.6	NA	NA	SSE	43	WSW	SW	24	17	47	32	1009.4	1009.7	NA	2	15.5	21	Yes	No
21	20-12-2008	Albury	9.8	25.6	0	NA	NA	SSE	26	SE	NNW	17	6	45	26	1019.2	1017.1	NA	NA	15.8	23.2	No	No
22	21-12-2008	Albury	11.5	29.3	0	NA	NA	S	24	SE	SE	9	9	56	28	1019.3	1014.8	NA	NA	19.1	27.3	No	No

Summary of the dataset

index	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm
count	143975	144199	142199	82670	75625	135197	143693	142398	142806	140953	130395	130432	89572	86102	143693	141851
mean	12.1940343	23.22134827	2.3609181	5.4682315229	7.6111775	40.0352300716731	14.0434259149715	18.6626567788873	68.8808313376188	51.53911588	1017.64993979	1015.2558888	4.4474612602	4.509930082	16.9906314	21.68339031800974
std	6.39849497	7.119048845	8.4780597	4.1937040941	3.7854829	13.6070622673813	8.915375322679528	8.8098000212514	19.0291644518441	20.7959016560	7.10653028752	7.0374138081	2.8871588535	2.7203573	6.4887531	6.936650460035525
min	-8.5	-4.8	0	0	0	6	0	0	0	0	980.5	977.1	0	0	-7.2	-5.4
25%	7.6	17.9	0	2.6	4.8	31	7	13	57	37	1012.9	1010.4	1	2	12.3	16.6
50%	12	22.6	0	4.8	8.4	39	13	19	70	52	1017.6	1015.2	5	5	16.7	21.1
75%	16.9	28.2	0.8	7.4	10.6	48	19	24	83	66	1022.4	1020	7	7	21.6	26.4
max	33.9	48.1	371	145	14.5	135	130	87	100	100	1041	1039.6	9	9	40.2	46.7

PROJECT FLOW



EXPLORATORY DATA ANALYSIS

- We removed the date column from the dataset as it was not going to be a useful feature.
- We converted the Rain Today, Rain Tomorrow features into binary form, where we denoted rain being there as 1 and rain not being there as 0.
- We have removed all the unknown values(NA's) for ease of data analysis.
- We have added one hot encoding for the following columns

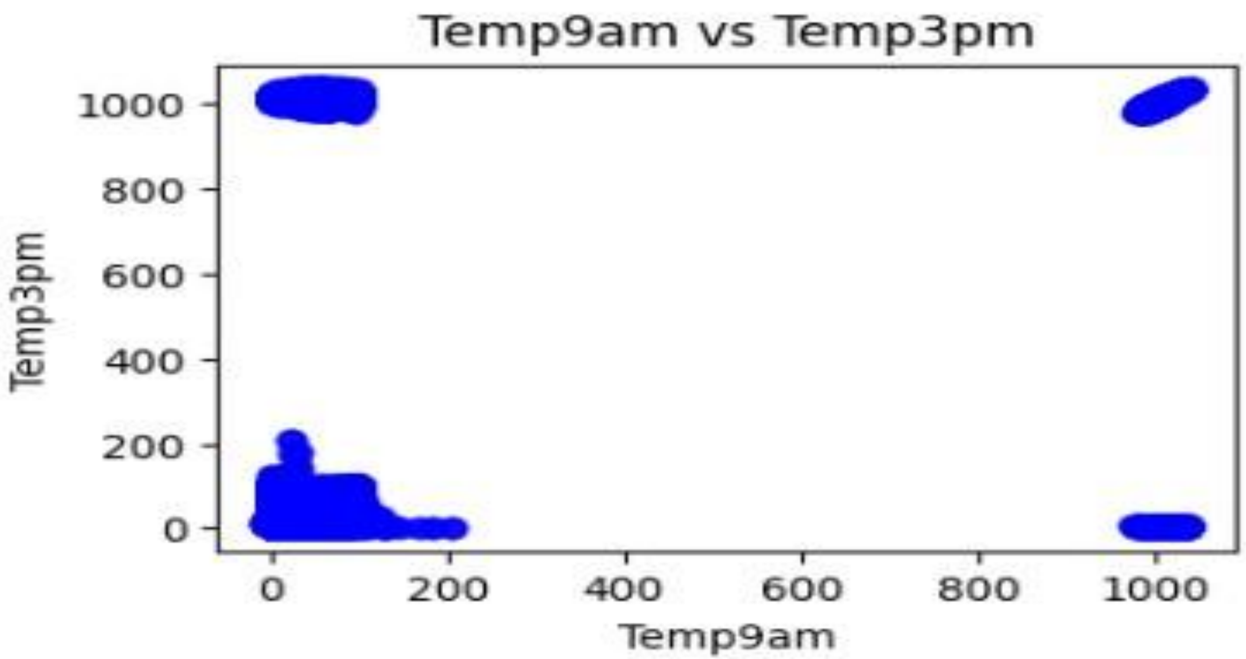
  Location, WindGustDir, WindDir9am, WindDir3pm

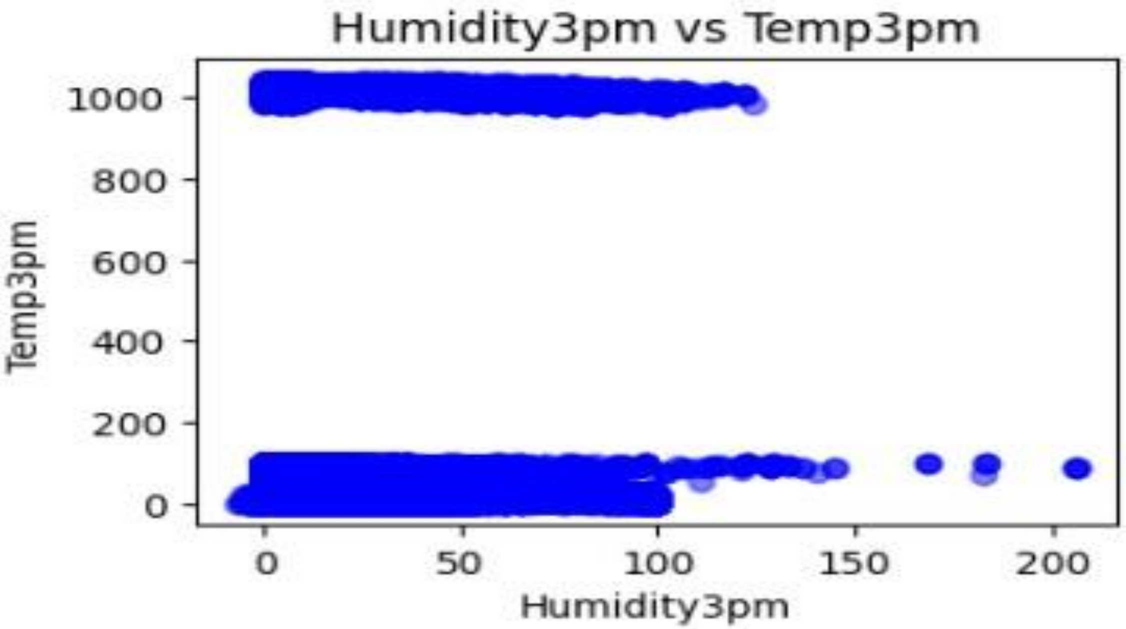
HANDLING THE IMBALANCED DATA

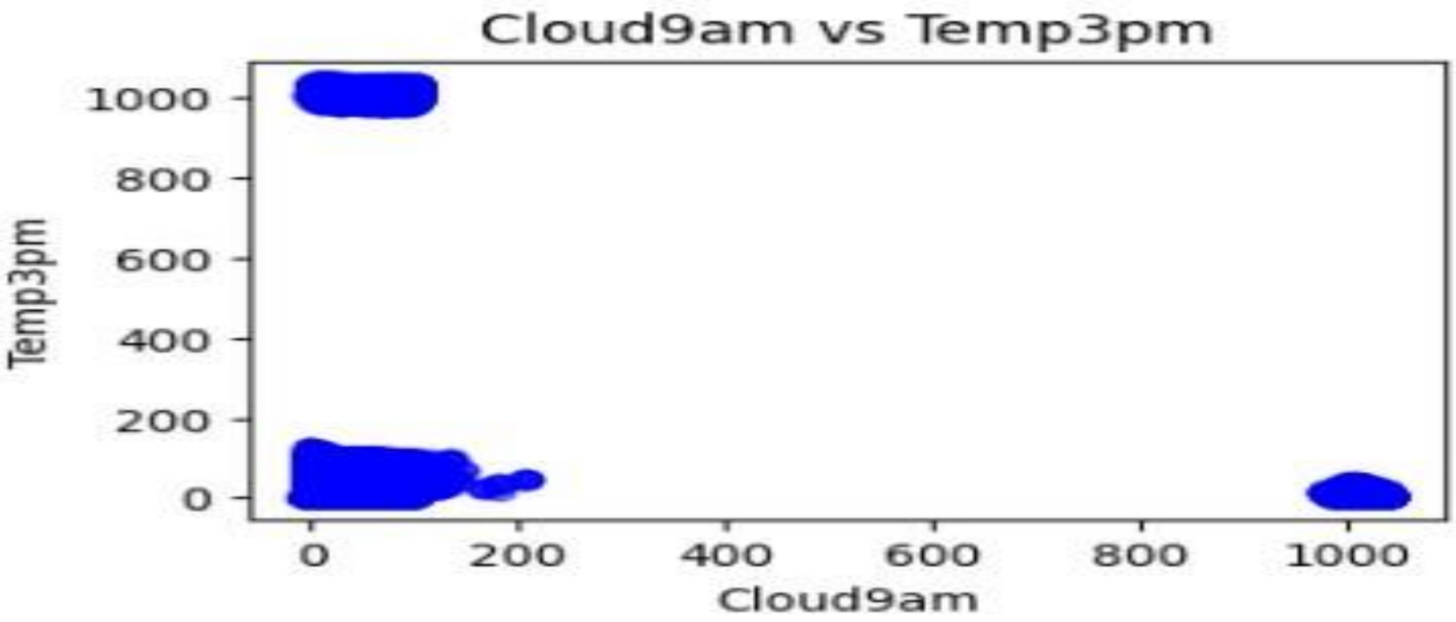
- The dataset had sparse occurrences of significant rainfall compared to periods of little to no rain, leading to an imbalanced dataset. To address this imbalance, we utilized the Python package Imbalanced-learn.

- To equalize the distribution of data representing different rainfall intensities, we created additional samples of the underrepresented class — in this case, significant rainfall events.

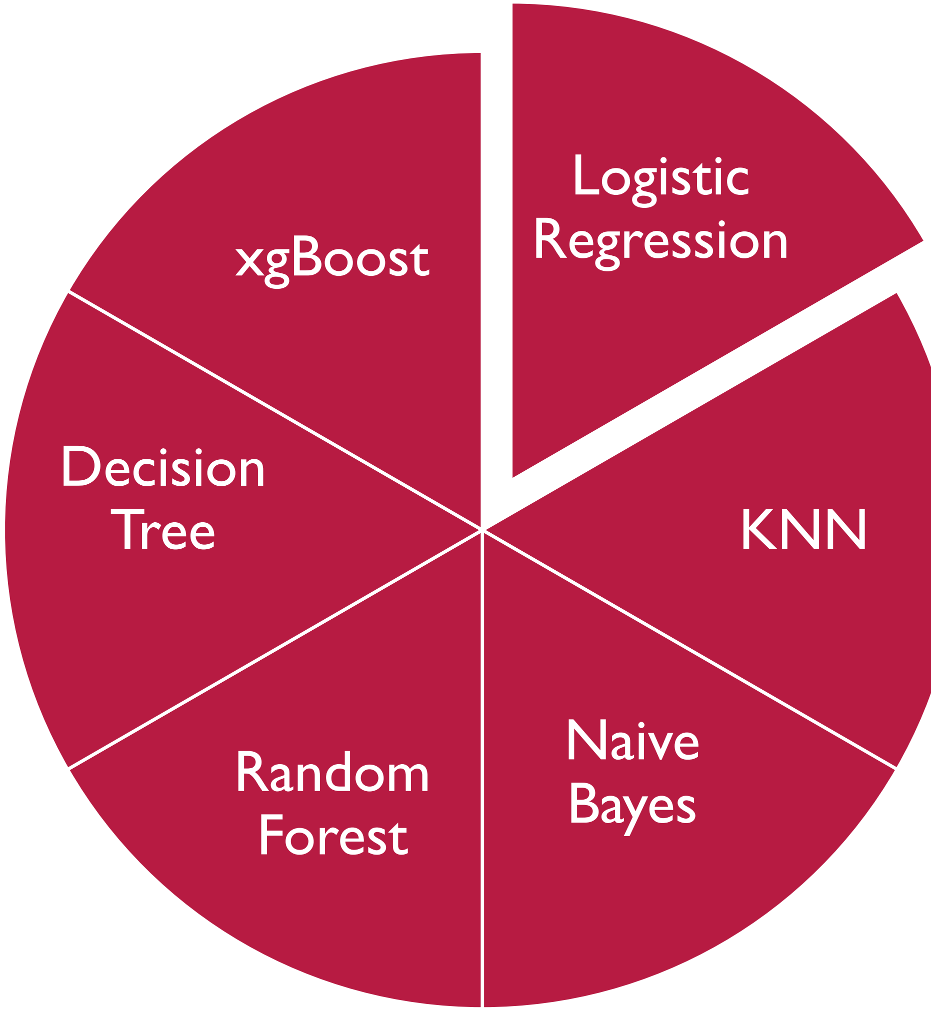
- This was achieved by employing the 'RandomOverSampler' function from the imbalanced-learn library, thus enhancing our model's ability to predict rain by learning from a more balanced dataset.


Temp9am vs Temp3pm
1000
800
600
400
200
0
Temp3pm
0
200
400
600
800
1000
Temp9am


Humidity3pm vs Temp3pm
1000
800
600
400
200
0
Temp3pm
0
50
100
150
200
Humidity3pm


Cloud9am vs Temp3pm
1000
800
600
400
200
0
Temp3pm
0
200
400
600
800
1000
Cloud9am

CLASSIFICATION MODELS



LOGISTIC REGRESSION

- Logistic Regression is a statistical method used for binary classification. In rain prediction, it helps in determining the probability of rain occurrence on a specific day, based on historical weather data. This model is valued for its simplicity and effectiveness in binary outcome predictions.
- LR Accuracy: 0.8089555631321741

LOGISTIC REGRESSION

```
Confusion Matrix:
 [[7176 1620]
 [1742 7060]]
True Positives(TP) = 7060
True Negatives(TN) = 7176
False Positives(FP) = 1620
False Negatives(FN) = 1742
```

classification report:

	precision	recall	f1-score	support
0	0.80	0.82	0.81	8796
1	0.81	0.80	0.81	8802
accuracy			0.81	17598
macro avg	0.81	0.81	0.81	17598
weighted avg	0.81	0.81	0.81	17598

K-NEAREST NEIGHBORS (KNN)

- KNN is a non-parametric method used for classification and regression. In rain prediction, it involves analyzing weather patterns by considering the 'k' nearest data points to predict rainfall. KNN is appreciated for its adaptability and ease of implementation.

- k = 3 Accuracy: 0.844641436526878I

- k = 5 Accuracy: 0.8063984543698147

- k = 10 Accuracy: 0.7824184566428003

K NEAREST NEIGHBOR

```
Confusion Matrix:
 [[6867 1929]
 [ 805 7997]]
True Positives(TP) = 7997
True Negatives(TN) = 6867
False Positives(FP) = 1929
False Negatives(FN) = 805
```

```
classification report:
              precision    recall  f1-score   support

           0       0.90      0.78      0.83      8796
           1       0.81      0.91      0.85      8802

    accuracy                           0.84     17598
   macro avg       0.85      0.84      0.84     17598
weighted avg       0.85      0.84      0.84     17598
```

NAÏVE BAYES

- Naive Bayes is a probabilistic classifier that applies Bayes' theorem with strong independence assumptions between the features. It is particularly effective in rain prediction when dealing with large datasets, offering fast and efficient predictions.
- NB Accuracy: 0.721502443459484

NAÏVE BAYES

```
Confusion Matrix:
 [[6041 2755]
 [2146 6656]]
True Positives(TP) = 6656
True Negatives(TN) = 6041
False Positives(FP) = 2755
False Negatives(FN) = 2146
```

classification report:

	precision	recall	f1-score	support
0	0.74	0.69	0.71	8796
1	0.71	0.76	0.73	8802
accuracy			0.72	17598
macro avg	0.72	0.72	0.72	17598
weighted avg	0.72	0.72	0.72	17598

RANDOM FOREST

- The Random Forest model is a powerful ensemble learning technique that builds multiple decision trees during training and predicts the outcome based on the mode of the results from individual trees.

- This approach enhances robustness and accuracy, outperforming single decision trees, particularly with complex datasets. It's well-suited for rain prediction as it effectively handles large datasets with multiple variables and captures complex, nonlinear relationships within weather data.

- Random Forests also mitigate overfitting risks, providing reliable and generalizable predictions for varied meteorological conditions.

- RF Accuracy: 0.7923059438572565

RANDOM FOREST

```
Confusion Matrix:
 [[6714 2082]
 [1573 7229]]
True Positives(TP) = 7229
True Negatives(TN) = 6714
False Positives(FP) = 2082
False Negatives(FN) = 1573
```

classification report:

	precision	recall	f1-score	support
0	0.81	0.76	0.79	8796
1	0.78	0.82	0.80	8802
accuracy			0.79	17598
macro avg	0.79	0.79	0.79	17598
weighted avg	0.79	0.79	0.79	17598

DECISION TREE

- The Decision Tree model uses a tree-like graph of decisions and their possible consequences. It is intuitive and easy to understand, making it a popular choice for rain prediction. Decision Trees are particularly useful for visualizing the decision-making process.

- Decision Tree Accuracy: 0.7287759972724174

DECISION TREE

```
Confusion Matrix:
 [[6469 2327]
 [2446 6356]]
True Positives(TP) = 6356
True Negatives(TN) = 6469
False Positives(FP) = 2327
False Negatives(FN) = 2446
```

classification report:

	precision	recall	f1-score	support
0	0.73	0.74	0.73	8796
1	0.73	0.72	0.73	8802
accuracy			0.73	17598
macro avg	0.73	0.73	0.73	17598
weighted avg	0.73	0.73	0.73	17598

xgBoost Classifier

- The xgboostClassifier is a classification algorithm based on the decision tree model. It is an implementation of the XGBoost algorithm, which is an optimized and scalable gradient boosting library that uses decision trees as base learners.

- xgBoost Accuracy: 0.8031026252983293

xgBoost Classifier

```
Confusion Matrix:
 [[6970 1826]
 [1639 7163]]
True Positives(TP) = 7163
True Negatives(TN) = 6970
False Positives(FP) = 1826
False Negatives(FN) = 1639
```

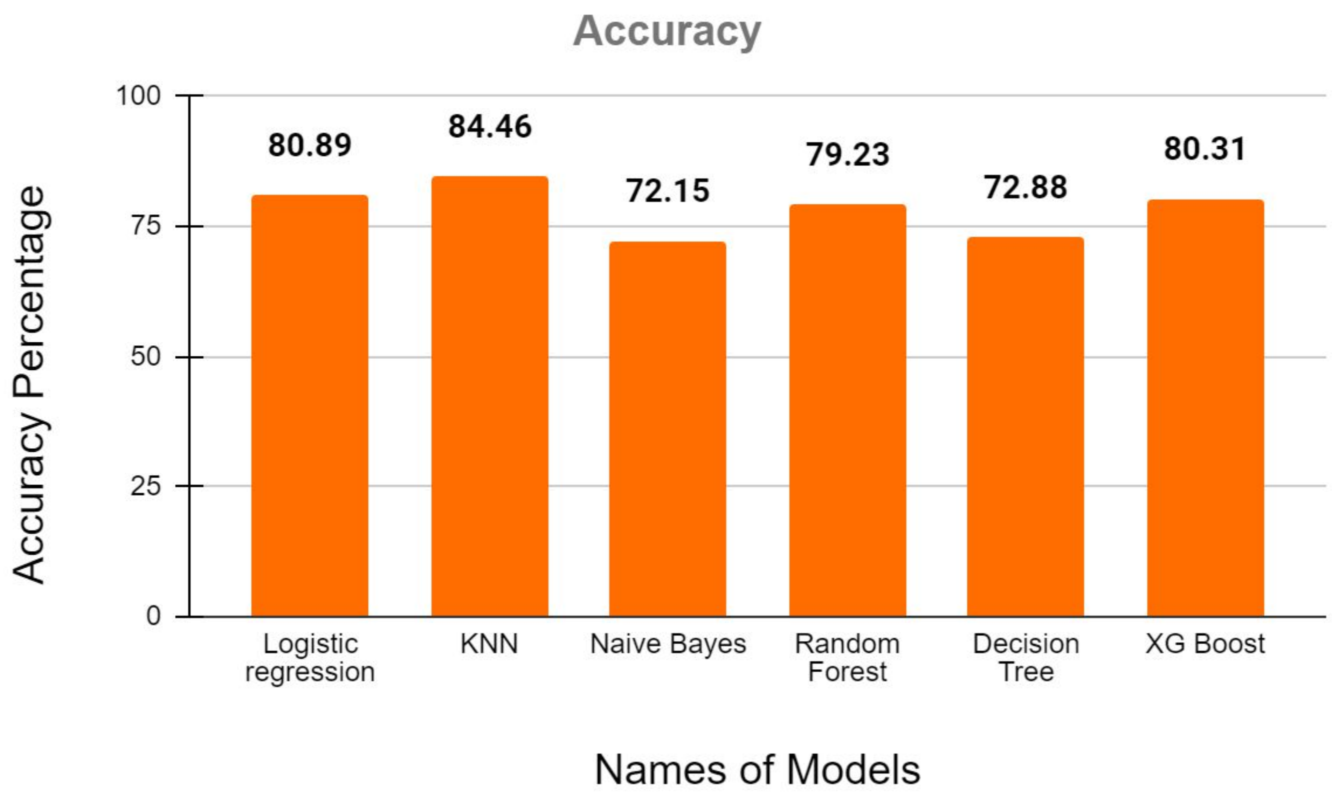
```
classification report:
              precision    recall  f1-score   support

           0       0.81      0.79      0.80      8796
           1       0.80      0.81      0.81      8802

    accuracy                           0.80     17598
   macro avg       0.80      0.80      0.80     17598
weighted avg       0.80      0.80      0.80     17598
```

CONCLUSION

- As per the classification models used and considering the accuracy of each and every model, the best classification model is K-Nearest Neighbor (k=3) with 84.46% accuracy.
- The following graph compares the accuracy of the models used



THANK YOU