Table of Contents

1 Business Understanding

1.1 Business Problem

Global warming affects the ecosystem and the human communities across the globe are losing natural resources and cultural significance. Australia is one of the worst-hit countries that is experiencing high temperatures, rising sea levels, coral bleaching in The Great Barrier Reef and unseasonal rain. One such example is the "Black Summer" bushfire in 2019 where more than 55 million acres of land were burnt. However, food security is a major issue caused by climate change. Agriculture largely depends on rain and hence its forecast becomes imperative. This project mainly focuses on the prediction of rain the next day. Rain tomorrow is the target variable of the data. Multiple attributes like Temperature, Windspeed, Humidity, and Pressure are considered to predict the rain on the next day. The information in the dataset is gathered from automated equipment from various weather stations. The goal of the project is to build an efficient classification model which would help in building budget-wise rainfall forecast applications.

1.2 Dataset

The Dataset we used for predicting the rainfall prediction in Australia from Kaggle. This dataset consists of 23 features and 145461 rows. The "RainTomorrow" attribute is considered the target variable. Below is the list of features and their descriptions:

1. Date: -The date of observation
2. Location: -The name of the location of the weather station
3. MinTemp: -The minimum temperature in degrees Celsius
4. Max Temp: -The maximum temperature in degrees Celsius
5. Rainfall: -The rainfall recorded for the day
6. Evaporation: -The evaporation (mm) during the day
7. Sunshine: -The number of hours of sunshine in the day
8. WindGustDir: -The direction of the wind gust
9. WindGustSpeed: -The speed of the wind gust
10. WindDir9am: -Direction of wind at 9am
11. WindDir3pm: -Direction of wind at 3 pm

12.WindSpeed9am: -Wind speed at 9am(km/hr)

13.WindSpeed3pm: -Wind speed at 3pm(km/hr)

14.Humidity9am: -Humidity at 9 am

15.Humidity3pm: -Humidity at 3 pm

16.Pressure9am: -Atmospheric pressure at 9 am

17.Pressure3pm: -Atmospheric pressure at 3 pm

18.Cloud9am: - Fraction of sky obscured by cloud at 9 am

19.Cloud3pm: - Fraction of sky obscured by cloud at 3 pm

20.Temp9am: - Temperature (degree celsius) at 9 am

21.Temp3pm: -Temperature (degree celsius) at 3pm

22.Rain Today: -Did, did it rain today?

23.Rain Tomorrow: -Target variable (Did it rain Tomorrow?)

1.3 Proposed Analytics Solution

1. Gathering data: For the project, we looked at a variety of online data sources before settling on a Kaggle dataset with many of the features mentioned above for rainfall prediction.

2. Data Analysis: : In the first stage, we analyze the data to better understand each element of the dataset, which allows us to better comprehend the data and identify important features and trends that could be beneficial in model building.

3. Data Preprocessing: : Through the data analysis step, we will handle the data quality issues using different approaches suitable for the issue (such as imputation for missing values and clamp transformation for outliers).

4. Feature Selection: Selecting the features and implementing dimensionality reduction using Chi square test, PCA and Recursive Feature Elimination methods.

2 Data Exploration and Preprocessing

2.1 Data Quality Report

We have -- continuous variables and -- categorical variables. Data Quality Report shows the summary of each feature in the dataset.

The metrics for measuring data quality in Categorical Variables:

Count - Total number of records.

% Miss - Number of missing values.

Card - Cardinality of the feature i.e. number of unique values.

Mode - Frequently repeated value.

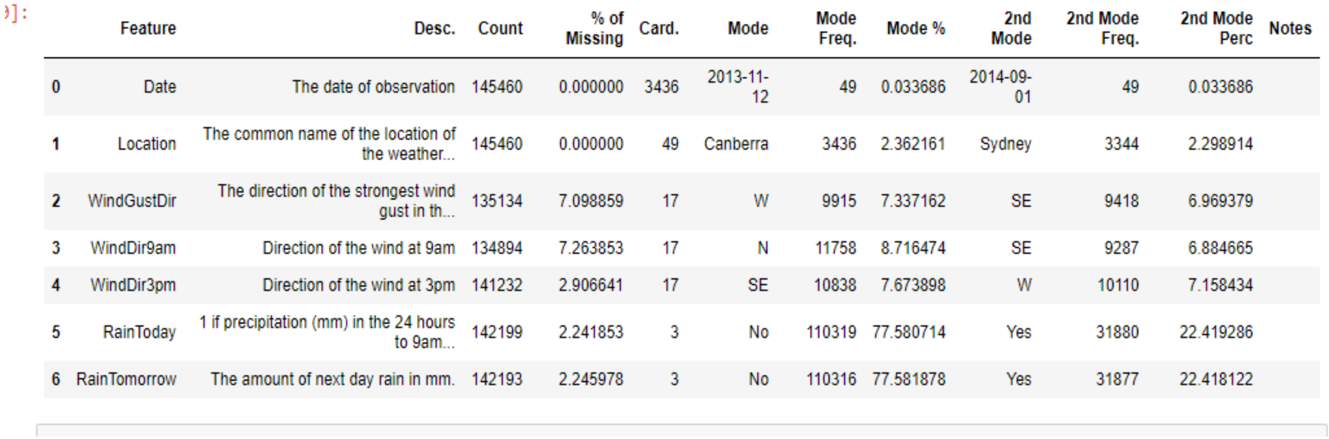
Mode Freq - Frequency of the mode value in the dataset.

Mode % - The percentage of repeated value in the dataset.

2nd Mode - Second most frequent value.

2nd Mode Freq - Frequency of the 2nd mode.

2nd Mode% - The percentage of second-most repeated value in the dataset.



	Feature	Desc.	Count	% of Missing	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode Perc	Notes
0	Date	The date of observation	145460	0.000000	3436	2013-11-12	49	0.033686	2014-09-01	49	0.033686	
1	Location	The common name of the location of the weather...	145460	0.000000	49	Canberra	3436	2.362161	Sydney	3344	2.298914	
2	WindGustDir	The direction of the strongest wind gust in th...	135134	7.098859	17	W	9915	7.337162	SE	9418	6.969379	
3	WindDir9am	Direction of the wind at 9am	134894	7.263853	17	N	11758	8.716474	SE	9287	6.884665	
4	WindDir3pm	Direction of the wind at 3pm	141232	2.906641	17	SE	10838	7.673898	W	10110	7.158434	
5	RainToday	1 if precipitation (mm) in the 24 hours to 9am...	142199	2.241853	3	No	110319	77.580714	Yes	31880	22.419286	
6	RainTomorrow	The amount of next day rain in mm.	142193	2.245978	3	No	110316	77.581878	Yes	31877	22.418122	

Fig 1. Data Quality report for Categorical features

The metrics for measuring data quality in Continuous Variables:

Count - Total number of records

% Miss - Number of missing values.

Card - Cardinality of the feature Min - Minimum value in the column.

1st Qrt. - Q1 First quartile - Values under 25% of data points.

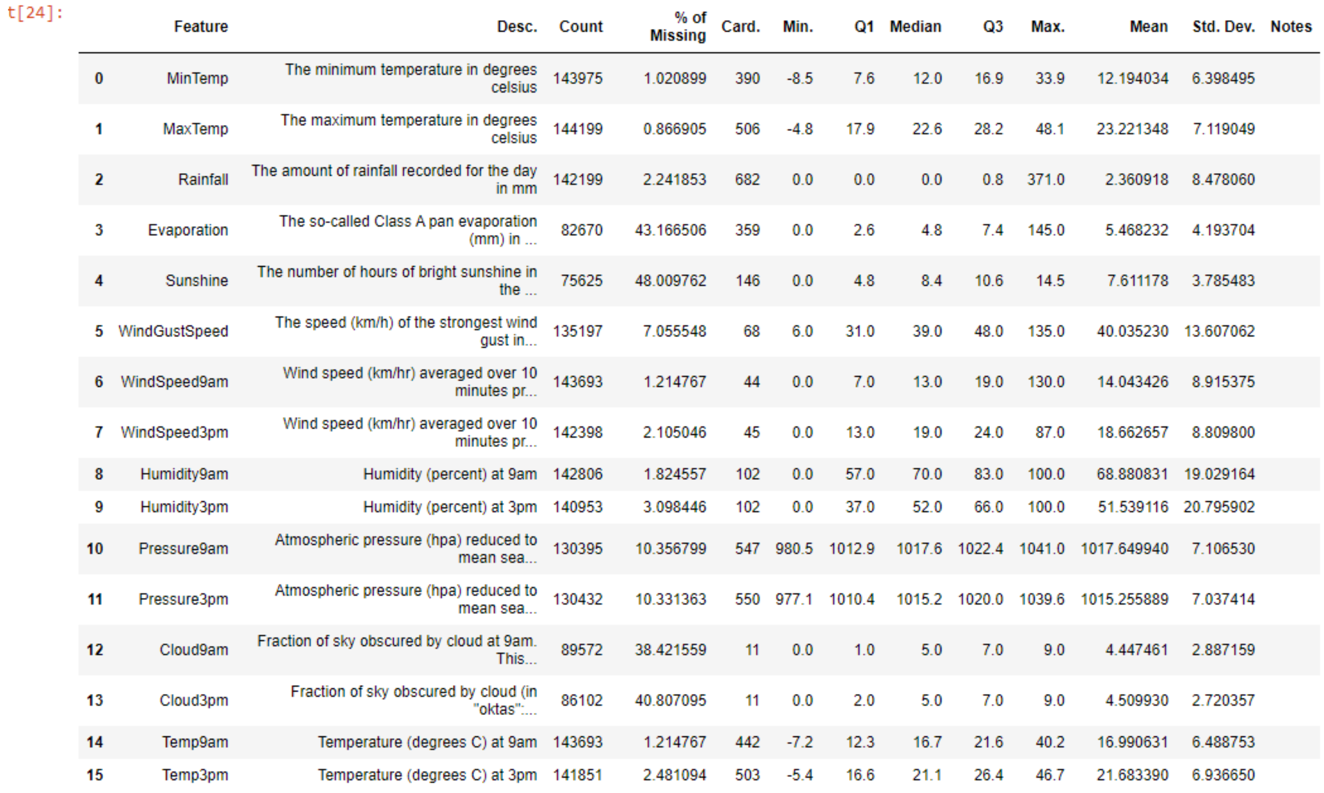
Mean - Mean value of each feature.

Median - Median value of each feature.

3rd Qrt. - Q3 Third quartile - Values over 75% of data points.

Max - Maximum value from all the data points in the column.

Std. Dev. - Standard deviation



t[24]:

	Feature	Desc.	Count	% of Missing	Card.	Min.	Q1	Median	Q3	Max.	Mean	Std. Dev.	Notes
0	MinTemp	The minimum temperature in degrees celsius	143975	1.020899	390	-8.5	7.6	12.0	16.9	33.9	12.194034	6.398495	
1	MaxTemp	The maximum temperature in degrees celsius	144199	0.866905	506	-4.8	17.9	22.6	28.2	48.1	23.221348	7.119049	
2	Rainfall	The amount of rainfall recorded for the day in mm	142199	2.241853	682	0.0	0.0	0.0	0.8	371.0	2.360918	8.478060	
3	Evaporation	The so-called Class A pan evaporation (mm) in ...	82670	43.166506	359	0.0	2.6	4.8	7.4	145.0	5.468232	4.193704	
4	Sunshine	The number of hours of bright sunshine in the ...	75625	48.009762	146	0.0	4.8	8.4	10.6	14.5	7.611178	3.785483	
5	WindGustSpeed	The speed (km/h) of the strongest wind gust in...	135197	7.055548	68	6.0	31.0	39.0	48.0	135.0	40.035230	13.607062	
6	WindSpeed9am	Wind speed (km/hr) averaged over 10 minutes pr...	143693	1.214767	44	0.0	7.0	13.0	19.0	130.0	14.043426	8.915375	
7	WindSpeed3pm	Wind speed (km/hr) averaged over 10 minutes pr...	142398	2.105046	45	0.0	13.0	19.0	24.0	87.0	18.662657	8.809800	
8	Humidity9am	Humidity (percent) at 9am	142806	1.824557	102	0.0	57.0	70.0	83.0	100.0	68.880831	19.029164	
9	Humidity3pm	Humidity (percent) at 3pm	140953	3.098446	102	0.0	37.0	52.0	66.0	100.0	51.539116	20.795902	
10	Pressure9am	Atmospheric pressure (hpa) reduced to mean sea...	130395	10.356799	547	980.5	1012.9	1017.6	1022.4	1041.0	1017.649940	7.106530	
11	Pressure3pm	Atmospheric pressure (hpa) reduced to mean sea...	130432	10.331363	550	977.1	1010.4	1015.2	1020.0	1039.6	1015.255889	7.037414	
12	Cloud9am	Fraction of sky obscured by cloud at 9am. This...	89572	38.421559	11	0.0	1.0	5.0	7.0	9.0	4.447461	2.887159	
13	Cloud3pm	Fraction of sky obscured by cloud (in "oktas":...	86102	40.807095	11	0.0	2.0	5.0	7.0	9.0	4.509930	2.720357	
14	Temp9am	Temperature (degrees C) at 9am	143693	1.214767	442	-7.2	12.3	16.7	21.6	40.2	16.990631	6.488753	
15	Temp3pm	Temperature (degrees C) at 3pm	141851	2.481094	503	-5.4	16.6	21.1	26.4	46.7	21.683390	6.936650	

Fig 2:- Data Quality report for Continuous features

2.2 Missing Values and Outliers

The Data Quality Report identified the missing values and their percentage in the dataset. The features WindGustDir, WindDir9am, WindDir3pm, Rain Today, RainTomorrow, MinTemp, MaxTemp, Rainfall, Evaporation, Sunshine, Humidity9am, Humidity3pm, Pressure 9 am, Pressure 3 pm, Cloud9am, Cloud3pm, Temp9am, Temp3pm all have missing values as shown in fig. So, we replaced all the missing values in the continuous data variables with the mean of each column as the variables show skewed distribution. In contrast, the categorical missing values are replaced with the mode of that particular column. Outliers are identified by the boxplots generated for each feature and by calculating the IQR (InterQuartile Range). The values above and below the range are considered outliers.

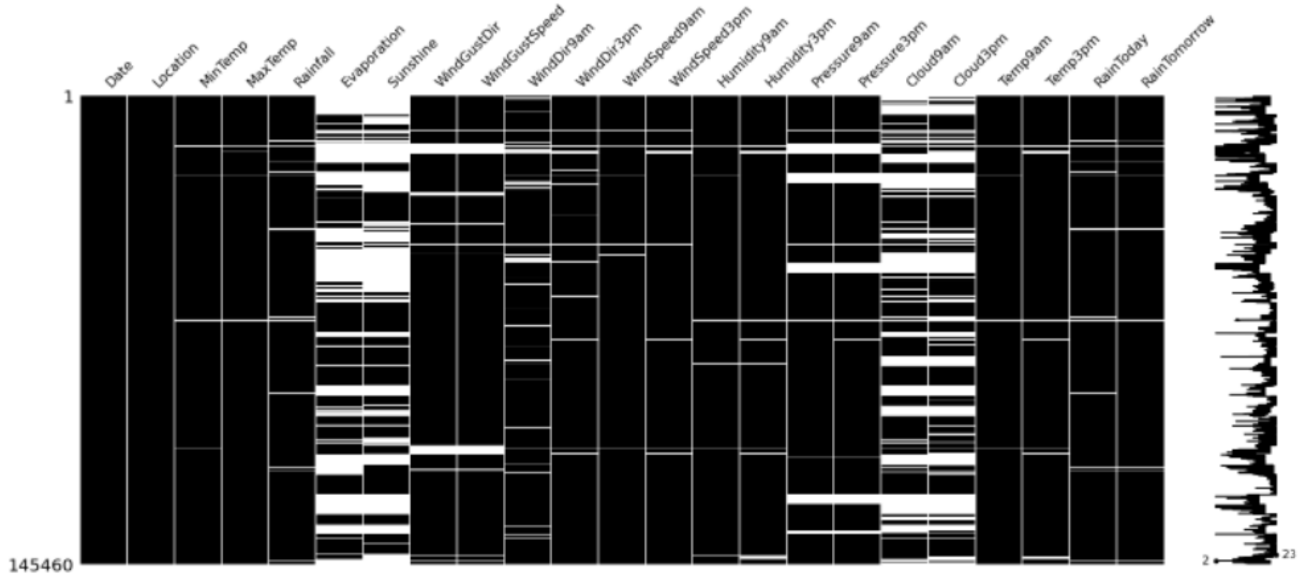


Fig 3:- Visualization of the missing values in the dataset

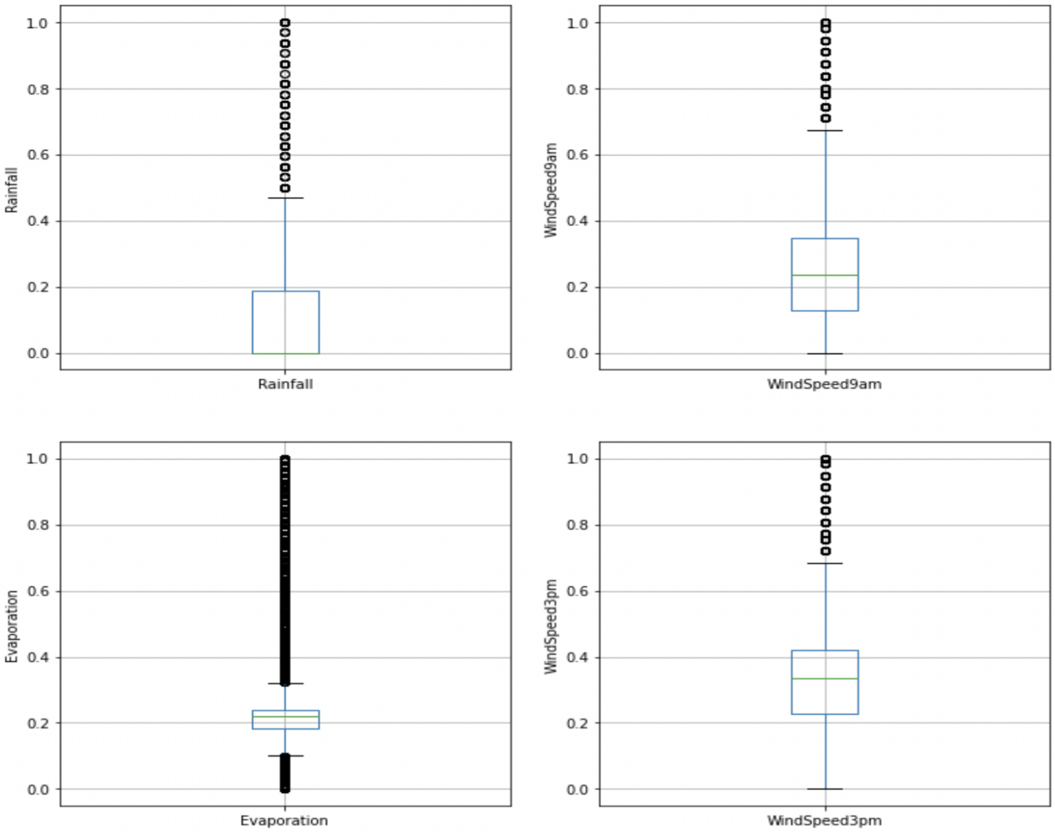


Fig 4:- Box Plot Visualization of the outliers in the dataset

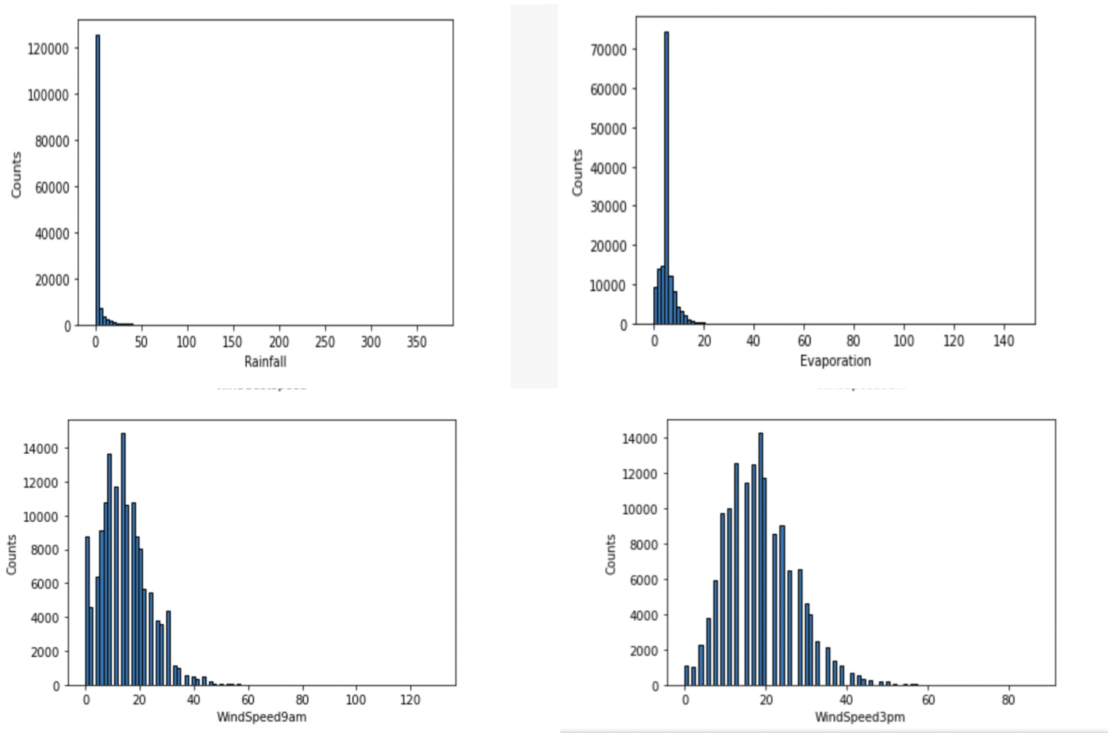


Fig 5:- Histogram Visualization of the outliers in the dataset

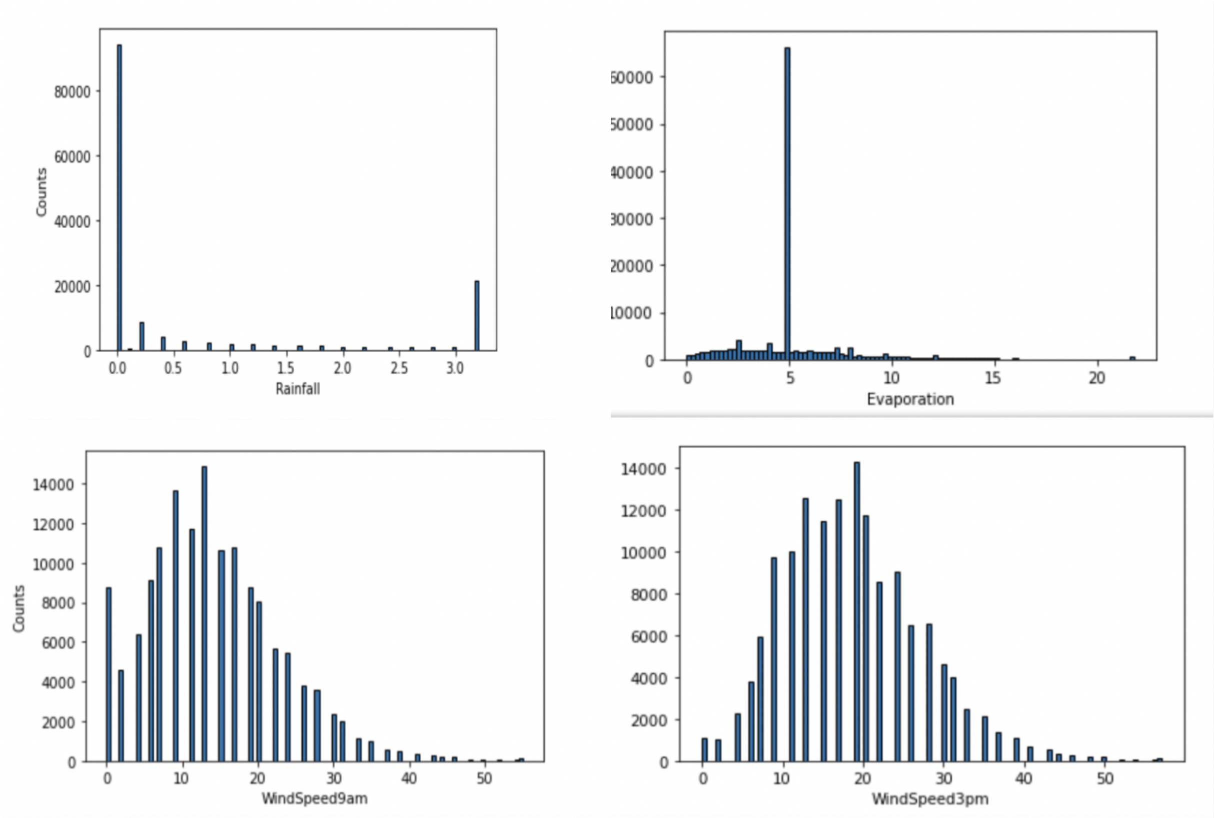


Fig 6 :- Histogram Visualization of the features after removing the outliers

2.3 Normalization

Data normalization is used to change the continuous features to fall within a specified range while maintaining the relative differences between the values for the feature. So for our data, we have used Min-Max Normalization. By using this method all the values of each feature will be transformed into the range [0,1] meaning that the minimum and maximum values of the feature will be 0 and 1. For performing this Min-Max normalization we have used the scikit-learn library.There are a total of 16 different continuous features in the dataset. The range varies for different features.

2.4 Transformations

The One-hot encoding method is used to convert categorical data variables for improving predictions.Most of the machine learning algorithms cannot directly work with categorical data(i.e labels) instead they require inputs and output variables to be numeric values which means we have to map each label to a numeric value.All the categorical columns in our dataset are converted into numerical data using this One-hot encoding method. Categorical columns like WindGustDir and WindDir3pm are converted using the One-hot encoding method.

2.4 Feature Selection

We preferred the features with the highest score for the Chi-Square test and found a list in descending order for the same. For PCA, we plotted the explained variance ratio, which is the percentage of variance attributed by each selected component. So, according to it, we should use around ten features which would be optimal. For RFE, we tried to select ten features, and after eliminating the features at each step we got these ten features : Rainfall, MaxTemp, Sunshine, WindGustSpeed, WindSpeed3pm, Humidity3pm, Pressure9am, Pressure3pm, Cloud3pm and Cloud9am.

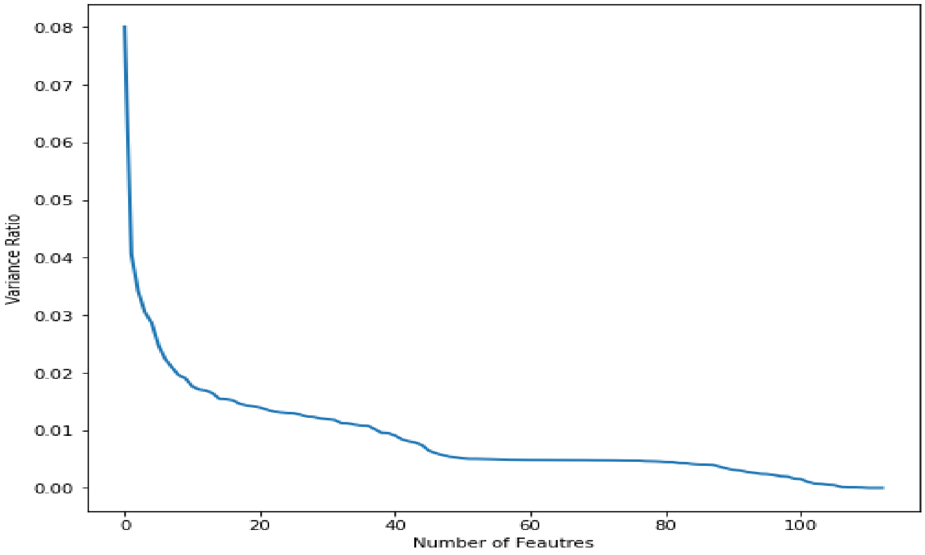


Fig 7 :- Graph Visualization of the Explained Variance Ratio of PCA Analysis

3 Model Selection

3.1 Logistic Regression

Logistic Regression is a statistical model to model the relationship between input variables and output variable. This model helped us in identifying the features that helped in predicting Rain Tomorrow. We have included all the relevant variables and our model predicted an accuracy score of 85.03%. We then created a confusion matrix and below are the readings,

True Positives(TP) - 35434

True Negatives(TN) - 5148

False Positives(FP) - 2029

False Negatives(FN) - 5391

Classification report -

	Prediction	**Recall**	**F1-score**	**Support**
0	0.87	0.95	0.91	37463
1	0.72	0.49	0.58	10539
Accuracy			0.85	48002
Macro avg	0.79	0.72	0.74	48002
Weighted avg	0.83	0.85	0.83	48002

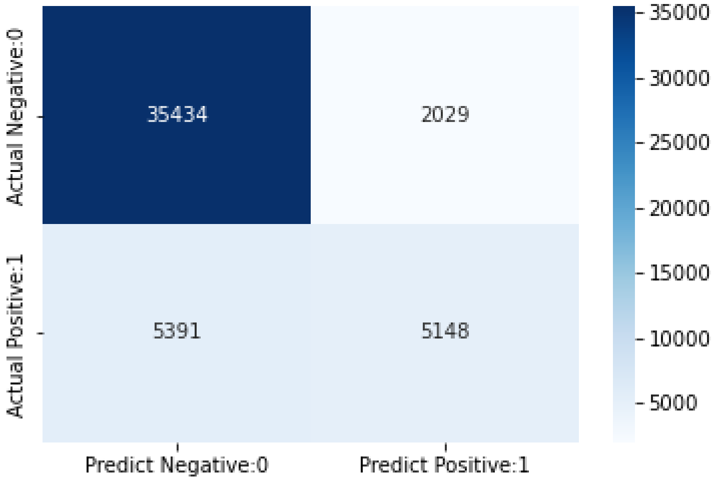


Fig 8 :- Visualization of Confusion Matrix with seaborn heatmap

Hyper Parameter Optimization : Hyperparameter thing is the process of choosing a set of hyperparameters.A good choice of such parameters makes a model succeed and optimize it. GridSearchCV is a member of sklearn model selection package.This gives the best parameter by fitting the model on the training set. This was performed to loop through predefined parameters with 'logreg' as estimator. We received {'penalty' : '12'} as the parameter that gives the best result .

Estimator chosen by the search is : LogisticRegression(random_state=0, solver='liblinear')

GridSearchCV score on test set is : 0.8454

3.2 Random Forest

A Random Forest algorithm is used to provide accuracy by handling a large proportion of data and it could work well with both categorical and numerical data without any normalization. We performed a Random Forest classifier with max_depth=2, random state=0 and received a Model accuracy score of 0.7811 and Training set score 0.8503.

Below are the confusion matrix readings,

True Positives(TP) - 37462

True Negatives(TN) - 34

False Positives(FP) - 1

False Negatives(FN) - 10505

Classification report -

	Prediction	**Recall**	**F1-score**	**Support**
0	0.78	1.00	0.88	37463
1	0.97	0.00	0.01	10539
Accuracy			0.78	48002
Macro avg	0.88	0.50	0.44	48002
Weighted avg	0.82	0.78	0.69	48002

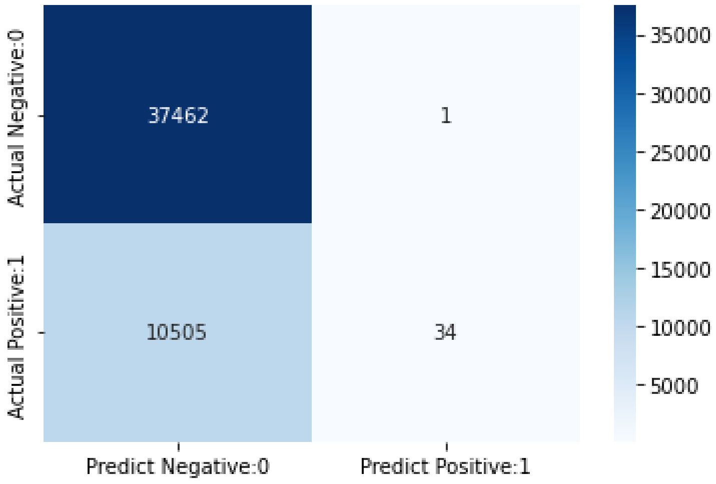


Fig 9 :- Visualization of Confusion Matrix with seaborn heatmap

We performed GridSearchCV with estimator max_depth=2, random_state=0 and received an accuracy score of 0.7811 and Training set score 0.7817.

3.3 KNN Classifier

KNN classifier is a supervised machine learning algorithm and doesn't make any assumptions about the underlying data distribution. We performed this model with n_neighbors= 3 and received a Model Accuracy score for testing data as 0.7923 and Training test score 0.8805.

Below are the confusion matrix readings,

True Positives(TP) - 34043

True Negatives(TN) - 3987

False Positives(FP) - 3420

False Negatives(FN) - 6552

Classification report -

	Prediction	**Recall**	**F1-score**	**Support**
0	0.87	0.95	0.91	37463
1	0.72	0.49	0.58	10539
Accuracy			0.85	48002

Macro avg	0.79	0.72	0.74	48002
Weighted avg	0.83	0.85	0.83	48002

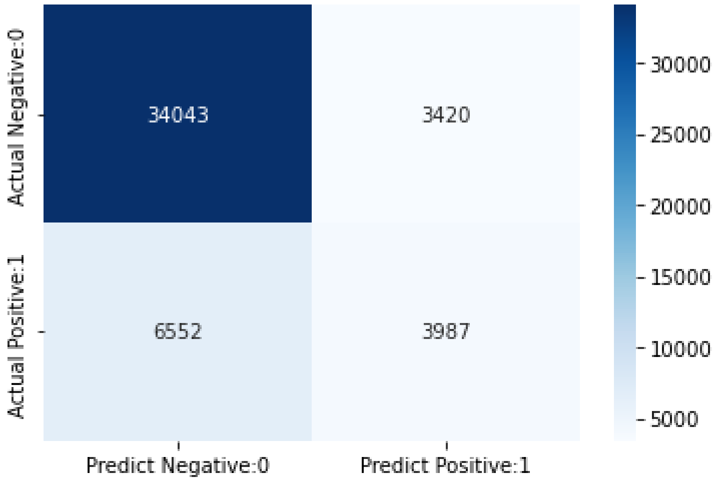


Fig 9 :- Visualization of Confusion Matrix with seaborn heatmap

3.4 Naive Bayes Classifier

Naive Bayes Classifier is one of the classification algorithms, to build models in order to make quick predictions. Assuming the dependent feature follows a normal distribution, we created a predict variable y_pred and used this function to make predictions. We received a model accuracy score of 0.7811 and Training Test score 0.7817.

Below are the confusion matrix readings,

True Positives(TP) - 37462

True Negatives(TN) - 34

False Positives(FP) - 1

False Negatives(FN) - 10505

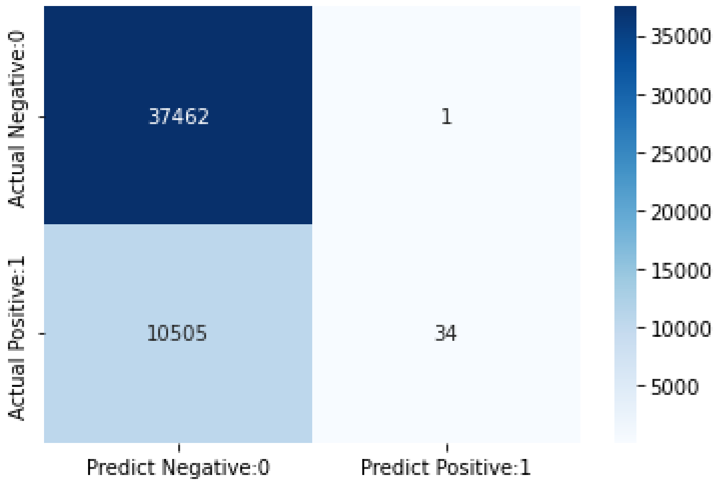


Fig 10 :- Visualization of Confusion Matrix with seaborn heatmap

3.5 AdaBoost Classifier

Adaptive Boosting is one of the boosting techniques used as an Ensemble method and to boost the performance of the model.With n_estimators=100 , we received a model accuracy score of 0.8447 and Testing score of 0.8485.

Below are the confusion matrix readings,

True Positives(TP) - 35438

True Negatives(TN) - 5107

False Positives(FP) - 2025

False Negatives(FN) - 5432

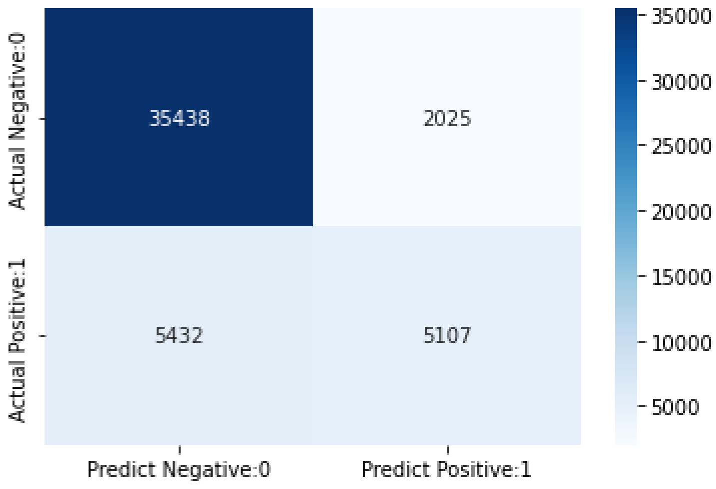


Fig 11 :- Visualization of Confusion Matrix with seaborn heatmap

3.6 Gradient Boosting Classifier

Gradient Boosting is a machine learning algorithm used to combine different weak learning models to make a strong predictive model. With n_estimators=100, we received a model accuracy of 0.8462 and a training set score of 0.8502.

Below are the confusion matrix readings,

True Positives(TP) - 35514

True Negatives(TN) - 5103

False Positives(FP) - 1949

False Negatives(FN) - 5436

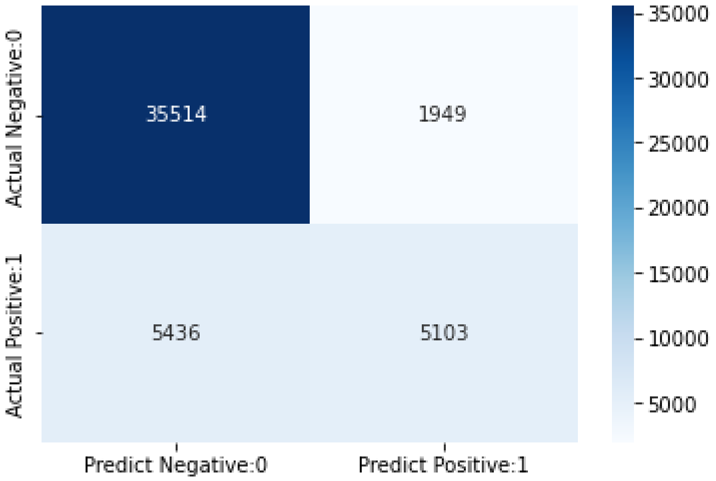


Fig 12 :- Visualization of Confusion Matrix with seaborn heatmap

3.7 XGBoost Classifier

Extreme Gradient Boosting is a decision tree-based ensemble method that uses a gradient boosting framework to boost the performance of a model. With n_estimators=100, we received a model accuracy score of 0.8562 and Testing score of 0.8980.

Below are the confusion matrix readings,

True Positives(TP) - 35410

True Negatives(TN) - 5691

False Positives(FP) - 2053

False Negatives(FN) - 4848

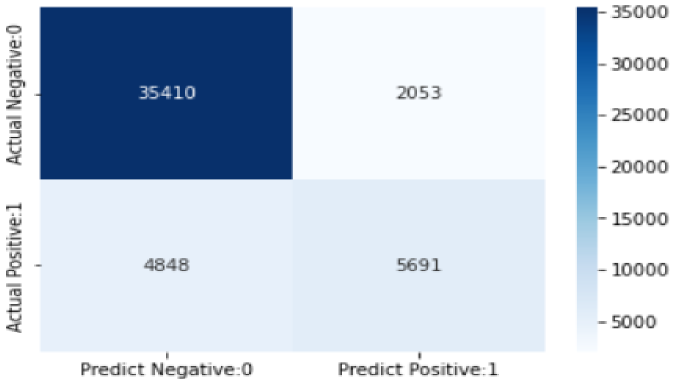


Fig 13 :- Visualization of Confusion Matrix with seaborn heatmap

4. Evaluation

4.1 Accuracy

Accuracy is an evaluation metric that measures the total number of predictions that a model gets right.

Accuracy = Correct Predictions / Total Predictions

4.2 Sensitivity

Sensitivity is the true positive rate that summarizes how well the positive class is predicted.

Sensitivity = TruePositive / (TruePositive + FalseNegative)

4.3 Specificity

Specificity is the true negative rate that summarizes how well the negative class is predicted.

Specificity = TrueNegative / (FalsePositive + TrueNegative)

4.3 Precision Score

Precision refers to the number of positive class predictions that actually belong to the positive class. It can be seen as a measure of quality.

Precision = TruePositive / (TruePositive + FalsePositive)

4.4 Recall

Recall refers to the percentage of total relevant results correctly classified by the algorithm. It can be seen as a measure of quantity.

Recall = TruePositive / (TruePositive + FalseNegative)

4.5 FNR

False Negative Rate

FNR = False Negative / (True Positive + False Negative)

4.5 Youden’s Index

Youden’s J statistic or Youden’s index is a single statistic that captures the performance of a dichotomous diagnostic test.

Youden’s Index = Sensitivity + Specificity - 1

4.6 Discriminant Power

This metric evaluates how well the classification model distinguishes between positive and negative samples.

$$DP = \frac{\sqrt{3}}{\pi}\left(log\left(\frac{TPR}{1-TNR}\right) + log\left(\frac{TNR}{1-TPR}\right)\right)$$

4.7 Balanced Classification Rate

This metric combines the sensitivity and specificity metrics and it is calculated using:

BCR = ½ (TPR + TNR)

4.8 Geometric Mean

The Geometric Mean (GM) is the average value or mean that represents the central tendency of a set of numbers by calculating the product of their values. Essentially, we multiply the numbers together and then take the nth root of the resulting numbers, where n is the total number of data values.

$$GM = \sqrt{TPR \times TNR} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}$$

5. RESULTS

	Accuracy	Sensitivity	Precision Score		False Negative rate	Youden’s Index	Discrimination Power	Balanced Classification Rate	Geometric Mean
Logistic Regression	0.8503	0.72	0.79		0.13	0.59	1.55	0.79	0.79
Random Forest	0.7811	0.66	0.82		0.16	0.64	1.67	0.82	0.82
KNN	0.7923	0.64	0.74		0.16	0.47	1.2	0.74	0.73
Naive Bayes	0.7811	0.50	0.88		0.22	0.75	2.65	0.88	0.87
AdaBoost	0.8447	0.72	0.79		0.13	0.58	1.54	0.79	0.79
Gradient Boosting	0.8462	0.72	0.80		0.13	0.59	1.57	0.80	0.79
XGBoost	0.8562	0.74	0.81		0.12	0.62	1.67	0.81	0.81

6. CONCLUSION

In conclusion, we have used Multiple classification machine learning models to predict the rain in Australia. We have performed all the pre-processing steps -Normalization, Transformations, Handled missing values, and Outliers. After pre-processing, we used this pre-processed dataset to implement our machine learning models. In this project, we have Implemented seven machine-learning algorithms to predict the rainfall and used nine metrics Sensitivity, Specificity, Precision score, Recall, F-measure, Discriminant power balanced, Classification Rate, Geometric Mean, Youden's Index to evaluate the performance of each of our models. Among the algorithms implemented, we observed that the XGBoost classifier performed the best and Naive Bayes showed the least efficiency on our dataset. So XGBoost classifier can be used to predict the rainfall the next day, and it would help build budget-wise rainfall forecast applications.