# **Problem Definition**

Weather forecasting is a crucial task with numerous applications ranging from agriculture to disaster management. Accurate predictions can help in making informed decisions, minimizing risks, and optimizing operations.

Weather forecasting involves using scientific principles and technology to predict atmospheric conditions for specific locations and times. This process includes gathering quantitative data about the current atmospheric state at a particular place and employing meteorological methods to forecast how the atmosphere will evolve.

In this project, we aim to design predictive models to forecast:

1. **Whether it will rain tomorrow.**
2. **The amount of rainfall expected.**

Using a dataset with 10 years of daily weather observations from various locations in Australia, we will develop machine learning models to achieve these objectives.

# **Data Analysis**

In this section, we examined and interpreted the dataset to understand its structure, quality, and the relationships between different variables. The detailed outline for this section are as follows:

1. We need to design a predictive model with the use of machine learning algorithms to forecast whether or not it will rain tomorrow. Hence RainTomorrow column is our target variable.
2. We also need to design a predictive model with the use of machine learning algorithms to predict how much rainfall could there be. Hence Rainfall is our other target variable.
3. The dataset has 8425 rows, and 23 columns present in the dataset.
4. We observed only 2 types of data present in the dataset i.e. float and object.
5. There are 16 columns of float data type and 7 columns of object data type.
6. Non-null count is not equal for all columns. Hence null values are present in the dataset.
7. We observe that RainTomorrow is categorical column with 2 categories i.e Yes and No. Also null values are present in the column. Hence, we need to build a classification model to predict 'RainTomorrow'.
8. We observe that Rainfall is continuous column with float data type. Hence, we need to build a regression model to predict 'Rainfall'.
9. We observe all the columns except Date and Location has null values present.
10. The columns Rainfall and RainTomorrow both has null values present. Since both are target columns hence, we have removed all the null values and its respective rows from the dataset.
11. Columns Evaporation and Sunshine has high presence of null values with approximately 41% and 47% respectively.
12. Columns Cloud9am and Cloud3pm has approximately 28% and 29% of null values present respectively.
13. Each column WindGustDir and WindGustSpeed has approximately 11.76% null values present.
14. There are other columns where null value are present. However, their percentage is less than 10%.
15. We have extracted Month from Date column and created a new column with name Month.
16. We have removed null values from columns MinTemp, MaxTemp, WindDir9am, WindDir3pm, Temp9am and Temp3pm using mode method based on each column’s locations and for that month.
17. We have removed null values from columns MinTemp, MaxTemp, WindDir9am, WindDir3pm, Temp9am and Temp3pm using mode method based on each column’s locations and for that month by grouping Location and Month columns.
18. We have removed null values from columns Evaporation, Sunshine, WindGustDir, WindGustSpeed, Pressure9am, Pressure3pm, Cloud9am and Cloud3pm using mode method based on each column’s month by grouping Month column.
19. After we have removed all the null values, the Month column was dropped from the dataset.
20. We have separated numerical and categorical columns.
21. We have checked the unique values of each categorical column.

* 12 unique locations available in Location column.
* 16 unique values present in WindGustDir column.
* 16 unique values present in WindDir9am column.
* 16 unique values present in WindDir3pm column.
* 2 unique values present in RainToday column

1. Statistical Analysis of numerical columns using describe method

* We can observe that high difference between 75 percentile and max in columns Rainfall, Evaporation, WindGustSpeed, WindSpeed9am and WindSpeed3pm. Hence, outliers are present in these columns.
* Mean is greater than median in columns Rainfall, Evaporation, WindGustSpeed. Hence these columns are right skewed.
* Median is greater than mean in columns Sunshine, Cloud9am. Hence, these columns are left skewed.

1. Statistical Analysis of categorical columns using describe method

* We can observe that Melbourne has the highest count in Location.
* Most of the WindGustDir is from North.
* Most of the WindDir9am is from North.
* Most of the WindDir3pm is from SouthEast.
* RainToday has 6120 counts of No rain. The RainToday column is highly biased.
* RainTomorrow has 6155 counts of No rain. The The RainToday column is highly biased.column is highly biased.

1. Univariate analysis was performed.
2. Visualizing categorical columns using count plot.

* Locations Melbounrne, WilliamTown, PerthAirport has high counts. Location Uluru has the lowest count in the dataset. There are other locations which has moderate counts.
* We can observe that the direction of the strongest wind gust is highest from North. Also, wind gust from all other directions are moderate. Direction of wind from NNW is the lowest
* Wind direction at 9AM from North and NorthWest is very high. Wind direction at 9AM from all other direction is moderate.
* Wind direction at 3PM from SouthEast very high. Wind direction at 3AM from all other direction is moderate.
* In RainToday column, No category count is very high.
* In RainTomorrow column, which is our target column, the No category is very high compared to Yes category. Hence the column is very biased.

1. Visualizing numerical columns using hist plot

* We can observe skewness in Rainfall column. However, since rainfall is one of the target columns hence no skewness check is required.
* Column Evaporation has high right skewness.
* Not much skewness is observed in any other columns.

1. Bivariate analysis is performed.
2. Visualizing categorical columns with Rainfall

* Melbourne has the highest count of rain among all categories. WilliamTown, PerthAirport also has high record of rainfall. Location Uluru has almost no record rainfall. However, the count records of Location Uluru in the dataset is very low.
* Most of the rainfall is observed when wind gust direction is from N (North). Moderate rainfall can also be observed from all other directions.
* Rainfall is observed highest when wind direction is from N (North) in column WinDir9am. Moderate rainfall can also be observed from all other directions.
* Rainfall is observed highest when wind direction is from SE (South-East) in column WinDir3pm. Moderate rainfall can also be observed from all other directions.
* In column RainToday, we observe that there is almost equal count of Yes for both categories of RainToday.

1. Visualizing numerical columns using hist plot with RainTomorrow

* We can observe that in all the columns No rain is greater than Yes rain.
* In can observe that in Humidity3pm column, with more humidity rainfall is more.

1. Visualizing numerical columns using regplot with Rainfall

* Not much correlation is observed among columns with Rainfall.

1. Multivariate analysis is performed.

* We observed Temp3pm Is highly correlated with MinTemp, MaxTemp and Temp9am and negatively correlated with Humidity9am and Humidity3pm.
* Temp9am Is highly correlated with MinTemp, MaxTemp and Temp3am and negatively correlated with Humidity9am, Humidity3pm, Pressure9am and Pressure3pm.
* Pressure9am is highly correlated with Pressure3pm. Pressure9am and Pressure3pm has negative correlation with MinTemp, WindGustSpeed.
* The target Rainfall has not much correlation with other columns. Low to moderate correlation can be observed with RainToday.

# **EDA Concluding Remarks**

The Exploratory Data Analysis (EDA) phase provided valuable insights into the structure, quality, and relationships within the weather dataset. Key findings include the significant presence of missing values in several columns, which were addressed through appropriate imputation techniques. The analysis revealed clear seasonal patterns in temperature and humidity, with summer months showing higher temperatures and lower humidity levels.

Geographical analysis underscored the variability of rainfall across different regions. Correlation analysis highlighted strong relationships between variables such as temperature and humidity, which will be critical in building accurate predictive models.

# **Rainfall Prediction – Regression analysis**

# **Pre-processing Pipeline**

Below are the steps followed in the pre-processing:

**Checking Outliers**

* 1. We have observed outliers in columns MaxTemp, Evaporation, WindGustSpeed, WindSpeed9am, WindSpeed9am, WindSpeed3pm, Humidity9am, Pressure9am, Pressure3pm, Temp9am, Temp3pm using boxplot.
  2. We have checked outliers using zscore and total data loss is 295 rows. The total data lost percentage is 3.65% which is within the acceptable range.
  3. We have checked outliers using IRQ method and total data loss found is 591 rows. Total percentage of data lost is 7.32% which is also within the acceptable range. However, we have applied zscore to remove outliers as it has minimum data loss
  4. The shape of the data after removing outliers is 7784 rows and 23 columns.

**Checking skewness**

1. Skewness is observed in Rainfall. However, since the column Rainfall is our target for regression analysis, hence, we will not remove skewness.
2. Low skewness is observed in Evaporation, Sunshine, WindSpeed9am. Hence, we will continue without removing skewness.

**Feature Selection**

1. We have dropped the column Date as it does not have any role in predicting rainfall.

**Label Encoding of categorical columns**

1. We have performed label encoding on categorical columns.

**Scaling the dataset**

1. We have separated the target and independent columns
2. We have performed scaling using StandardScaler

**Checking the Variance Inflation Factor**

1. High variance inflation is present in columns Pressure9am, Pressure3pm, Temp3pm, MaxTemp, Temp9am, MinTemp, Humidity9am, Humidity3pm, WindGustSpeed.
2. Hence, we have dropped columns Pressure9am, Temp3pm, Temp9am, Pressure3pm to limit the variance inflation factor.
3. Although VIF is high in columns MinTemp, MaxTemp, WindGustSpeed, Humidity9am, Humidity3pm, we will continue with modelling as these columns are essential in prediction.

# **Building Machine Learning Models for Rainfall using Regression analysis**

1. We have calculated the best random state and score using Linear Regression. The R2 Score is 0.34829415464217195 and best random state is151.
2. We have built the following models with random\_state 151. Following is the list of scores of each model.
   1. **Linear Regression Model**

| **Linear Regression** | **Scores** |
| --- | --- |
| **0** | R2 Score | 34.829415 |
| **1** | R2 Score on Training Data | 20.286998 |
| **2** | Mean Absolute Error | 2.878344 |
| **3** | Mean Squared Error | 41.469970 |
| **4** | Root Mean Squared Error | 6.439718 |

* 1. **Lasso Regression**

| **Lasso** | **Scores** |
| --- | --- |
| **0** | R2 Score | 23.650710 |
| **1** | R2 Score on Training Data | 14.658499 |
| **2** | Mean Absolute Error | 3.381000 |
| **3** | Mean Squared Error | 48.583311 |
| **4** | Root Mean Squared Error | 6.970173 |

* 1. **Ridge Regression**

| **Ridge** | **Scores** |
| --- | --- |
| **0** | R2 Score | 34.828417 |
| **1** | R2 Score on Training Data | 20.286980 |
| **2** | Mean Absolute Error | 2.878406 |
| **3** | Mean Squared Error | 41.470606 |
| **4** | Root Mean Squared Error | 6.439768 |

* 1. **Random Forest**

| **Random Forest** | **Scores** |
| --- | --- |
| **0** | R2 Score | 32.136520 |
| **1** | R2 Score on Training Data | 90.782937 |
| **2** | Mean Absolute Error | 1.830664 |
| **3** | Mean Squared Error | 43.183539 |
| **4** | Root Mean Squared Error | 6.571418 |

* 1. **KNN Regression**

| **KNN** | **Scores** |
| --- | --- |
| **0** | R2 Score | -5.984630 |
| **1** | R2 Score on Training Data | 39.551539 |
| **2** | Mean Absolute Error | 3.232397 |
| **3** | Mean Squared Error | 67.441154 |
| **4** | Root Mean Squared Error | 8.212256 |

* 1. **SVR Regression**

| **SVR** | **Scores** |
| --- | --- |
| **0** | R2 Score | 4.069440 |
| **1** | R2 Score on Training Data | 2.766063 |
| **2** | Mean Absolute Error | 2.426479 |
| **3** | Mean Squared Error | 61.043452 |
| **4** | Root Mean Squared Error | 7.813031 |

* 1. **Decision Tree Regressor**

| **Decision Tree** | **Scores** |
| --- | --- |
| **0** | R2 Score | -64.311125 |
| **1** | R2 Score on Training Data | 100.000000 |
| **2** | Mean Absolute Error | 2.124743 |
| **3** | Mean Squared Error | 104.556027 |
| **4** | Root Mean Squared Error | 10.225264 |

* 1. **Gradient Boosting Regressor**

| **Gradient Boosting Regressor** | **Scores** |
| --- | --- |
| **0** | R2 Score | 29.186202 |
| **1** | R2 Score on Training Data | 67.513837 |
| **2** | Mean Absolute Error | 2.253978 |
| **3** | Mean Squared Error | 45.060914 |
| **4** | Root Mean Squared Error | 6.712743 |

* 1. **ExtraTrees Regression**

| **Extra Trees Regressor** | **Scores** |
| --- | --- |
| **0** | R2 Score | 49.183987 |
| **1** | R2 Score on Training Data | 100.000000 |
| **2** | Mean Absolute Error | 1.501315 |
| **3** | Mean Squared Error | 32.335732 |
| **4** | Root Mean Squared Error | 5.686452 |

* 1. **AdaBoost Regressor**

| **AdaBoostRegressor** | **Scores** |
| --- | --- |
| **0** | R2 Score | -126.417351 |
| **1** | R2 Score on Training Data | -11.193961 |
| **2** | Mean Absolute Error | 10.381779 |
| **3** | Mean Squared Error | 144.076055 |
| **4** | Root Mean Squared Error | 12.003169 |

* 1. **SGD Regressor**

| **SGDRegressor** | **Scores** |
| --- | --- |
| **0** | R2 Score | -1.879778e+23 |
| **1** | R2 Score on Training Data | -9.635286e+22 |
| **2** | Mean Absolute Error | 2.783809e+11 |
| **3** | Mean Squared Error | 1.196158e+23 |
| **4** | Root Mean Squared Error | 3.458552e+11 |

* 1. **Poisson Regressor**

| **PoissonRegressor** | **Scores** |
| --- | --- |
| **0** | R2 Score | 25.918947 |
| **1** | R2 Score on Training Data | 20.595619 |
| **2** | Mean Absolute Error | 2.807489 |
| **3** | Mean Squared Error | 47.139964 |
| **4** | Root Mean Squared Error | 6.865855 |

1. We have performed cross validation for each model. The score is mentioned below.

|  | **Regression Name** | **Regression Score** | **Cross Val Score** | **Difference** |
| --- | --- | --- | --- | --- |
| **0** | LinearRegression | 3.482942e-01 | 2.448421e-01 | 1.034521e-01 |
| **1** | Lasso | 2.365071e-01 | 1.696872e-01 | 6.681992e-02 |
| **2** | Ridge | 3.482842e-01 | 2.448445e-01 | 1.034397e-01 |
| **3** | Random Forest | 3.213652e-01 | 3.143096e-01 | 7.055554e-03 |
| **4** | KNN | -5.984630e-02 | 8.166208e-02 | -1.415084e-01 |
| **5** | SVR | 4.069440e-02 | 3.596111e-02 | 4.733289e-03 |
| **6** | Decision Tree | -6.431113e-01 | -1.399271e+00 | 7.561596e-01 |
| **7** | GradientBoost | 2.918620e-01 | 3.157532e-01 | -2.389115e-02 |
| **8** | ExtraTreesRegressor | 4.918399e-01 | 4.283901e-01 | 6.344972e-02 |
| **9** | AdaBoostRegressor | -1.264174e+00 | -1.525008e+00 | 2.608347e-01 |
| **10** | SGDRegressor | -1.879778e+21 | -3.842889e+22 | 3.654911e+22 |
| **11** | PoissonRegressor | 2.591895e-01 | 1.922566e-01 | 6.693291e-02 |

* ExtraTreesRegressor has the highest scores among all models and a small difference, indicating strong performance.
* The ExtraTreesRegressor is the best-performing model overall with the highest scores and a small difference between the training and cross-validation scores.

1. We have performed Hyperparameter Tuning with ExtraTreesRegressor to check the best parameters and created the final model as mentioned below:

**ExtraTreesRegressor(criterion='poisson',n\_estimators= 150, max\_features= 'sqrt', max\_depth=None, random\_state= 100)**

**The final R2 Score 0.555350**

1. We have saved the model using pickle

# **RainTomorrow Prediction - Classification**

# **Pre-processing Pipeline**

Below are the steps followed in the pre-processing:

**Checking Outliers**

1. We have observed outliers in columns MaxTemp, Evaporation, WindGustSpeed, WindSpeed9am, WindSpeed9am, WindSpeed3pm, Humidity9am, Pressure9am, Pressure3pm, Temp9am, Temp3pm and Rainfall using boxplot.
2. We have checked outliers using zscore and total data loss is 401 rows. The total data lost percentage is 4.96% which is within the acceptable range.
   1. We have checked outliers using IRQ method and total data loss found is 2006 rows. Total percentage of data lost is 24.82% which is which is very high then acceptable range. Hence, we have applied zscore to remove outliers as it has minimum data loss
   2. The shape of the data after removing outliers is 7678 rows and 23 columns.

**Checking skewness**

1. Skewness is observed in Rainfall, Evaporation, WindGustSpeed, WindSpeed9am.
2. We have checked skewness using sqrt, cbrt and log.
3. We have removed skewness from WindGustSpeed and WindSpeed9am applying sqrt
4. We have removed skewness from Evaporation applying cbrt
5. We have removed skewness from Rainfall applying log

**Feature Selection**

1. We have dropped the column Date as it does not have any role in predicting rainfall.

**Label Encoding of categorical columns**

1. We have performed label encoding on categorical columns.

**Scaling the dataset**

1. We have separated the target and independent columns
2. We have performed scaling using StandardScaler

**Checking the Variance Inflation Factor**

1. Columns Pressure9am, Pressure3pm has very high variance inflation.
2. We can observe high variance in columns Temp3pm, MaxTemp, Temp9am, WindGustSpeed, Hence we have dropped columns Pressure9am, Temp3pm, Temp9am, Pressure3pm to limit the variance inflation factor.
3. We have dropped columns Pressure9am, Temp3pm, Temp9am, Pressure3pm, MaxTemp, WindGustSpeed, Evaporation and Humidity9am.
4. After dropping the columns variance inflation is within the acceptable range

#### **We have checked for Imbalance dataset**

1. We can observe that the dataset is imbalanced. Hence we need to apply SMOTE to make the dataset balanced and ready for modelling

# **Building Machine Learning Models for RainTomorrow using Classification**

1. We have evaluated the best random\_state using Logistic regression. The Accuracy score is 0.768762677484787 and Random state: 155
2. We have built the following models with random\_state 155. Following is the scores of each model.
3. **Logistic regression model**

The accuracy using Logistic regression is: 76.816901%

[[1376 403]

[ 420 1351]]

precision recall f1-score support

0 0.77 0.77 0.77 1779

1 0.77 0.76 0.77 1771

accuracy 0.77 3550

macro avg 0.77 0.77 0.77 3550

weighted avg 0.77 0.77 0.77 3550

1. **Support Vector Classification**

The accuracy using SVC is: 78.000000%

[[1450 329]

[ 452 1319]]

precision recall f1-score support

0 0.76 0.82 0.79 1779

1 0.80 0.74 0.77 1771

accuracy 0.78 3550

macro avg 0.78 0.78 0.78 3550

weighted avg 0.78 0.78 0.78 3550

1. **Random Forest Classifier**

The accuracy using Random Forest is: 92.197183%

[[1656 123]

[ 154 1617]]

precision recall f1-score support

0 0.91 0.93 0.92 1779

1 0.93 0.91 0.92 1771

accuracy 0.92 3550

macro avg 0.92 0.92 0.92 3550

weighted avg 0.92 0.92 0.92 3550

1. **Adaboost Classifier**

The accuracy using Adaboost is: 81.549296%

[[1478 301]

[ 354 1417]]

precision recall f1-score support

0 0.81 0.83 0.82 1779

1 0.82 0.80 0.81 1771

accuracy 0.82 3550

macro avg 0.82 0.82 0.82 3550

weighted avg 0.82 0.82 0.82 3550

1. **Gradient Boost Classifier**

The accuracy using Gradientboost is: 86.281690%

[[1572 207]

[ 280 1491]]

precision recall f1-score support

0 0.85 0.88 0.87 1779

1 0.88 0.84 0.86 1771

accuracy 0.86 3550

macro avg 0.86 0.86 0.86 3550

weighted avg 0.86 0.86 0.86 3550

1. **Bagging classifier**

The accuracy using Bagging classfier is: 90.309859%

[[1655 124]

[ 220 1551]]

precision recall f1-score support

0 0.88 0.93 0.91 1779

1 0.93 0.88 0.90 1771

accuracy 0.90 3550

macro avg 0.90 0.90 0.90 3550

weighted avg 0.90 0.90 0.90 3550

1. **ExtraTree classifier**

The accuracy using Extratrees is: 93.239437%

[[1660 119]

[ 121 1650]]

precision recall f1-score support

0 0.93 0.93 0.93 1779

1 0.93 0.93 0.93 1771

accuracy 0.93 3550

macro avg 0.93 0.93 0.93 3550

weighted avg 0.93 0.93 0.93 3550

1. **DecisionTree classifier**

The accuracy using Decisiontree is: 85.802817%

[[1530 249]

[ 255 1516]]

precision recall f1-score support

0 0.86 0.86 0.86 1779

1 0.86 0.86 0.86 1771

accuracy 0.86 3550

macro avg 0.86 0.86 0.86 3550

weighted avg 0.86 0.86 0.86 3550

1. **KNN classifier**

The accuracy using KNN is: 86.084507%

[[1364 415]

[ 79 1692]]

precision recall f1-score support

0 0.95 0.77 0.85 1779

1 0.80 0.96 0.87 1771

accuracy 0.86 3550

macro avg 0.87 0.86 0.86 3550

weighted avg 0.87 0.86 0.86 3550

1. **Ridge Classifier**

The accuracy using Ridge is: 76.788732%

[[1379 400]

[ 424 1347]]

precision recall f1-score support

0 0.76 0.78 0.77 1779

1 0.77 0.76 0.77 1771

accuracy 0.77 3550

macro avg 0.77 0.77 0.77 3550

weighted avg 0.77 0.77 0.77 3550

1. We have performed cross validation for each model. The score is mentioned below.

| **Classification Name** | **Classification Score** | **Cross Val Score** | **Difference** |
| --- | --- | --- | --- |
| **0** | Logistic Regression | 0.768169 | 0.699208 | 0.068961 |
| **1** | SVC | 0.780000 | 0.721519 | 0.058481 |
| **2** | RandomForestClassifier | 0.921972 | 0.809088 | 0.112884 |
| **3** | AdaBoostClassifier | 0.815493 | 0.662279 | 0.153214 |
| **4** | GradientBoostingClassifier | 0.862817 | 0.686286 | 0.176531 |
| **5** | BaggingClassifier | 0.903099 | 0.795480 | 0.107618 |
| **6** | ExtraTreesClassifier | 0.932394 | 0.822860 | 0.109534 |
| **7** | DecisionTreeClassifier | 0.858028 | 0.757872 | 0.100156 |
| **8** | KNeighborsClassifier | 0.860845 | 0.792009 | 0.068837 |
| **9** | RidgeClassfier | 0.767887 | 0.704279 | 0.063608 |

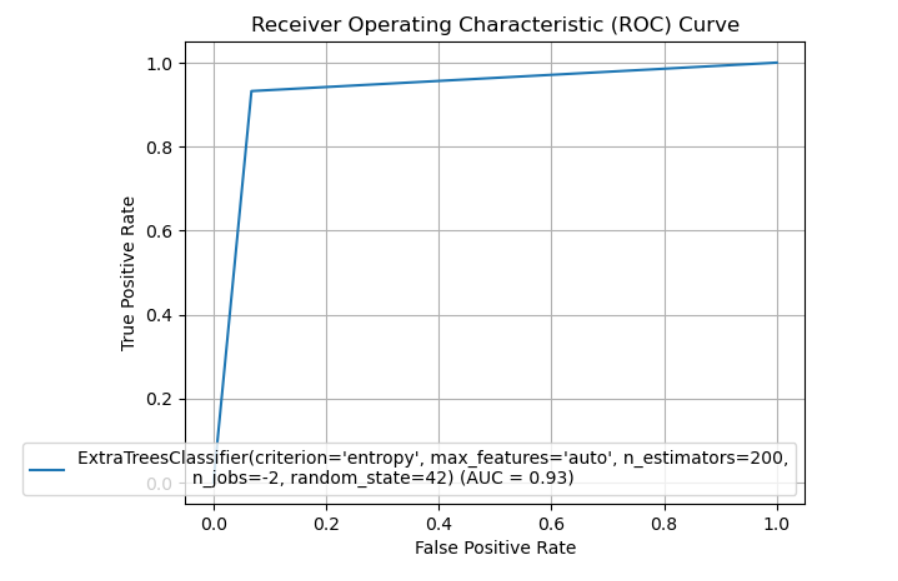
* ExtraTreesClassifier has the highest cross-validation scores (0.830214), indicating strong performance on unseen data.
* It also has a high training score (0.931549), suggesting it is capable of learning complex patterns in the data.
* The ExtraTreesClassifier is the best model due to its high cross-validation score and overall performance.

1. We have performed Hyperparameter Tuning with ExtraTreesClassifier to check the best parameters and created the final model as mentioned below

**ExtraTreesClassifier(criterion='entropy',max\_depth=None,max\_features='auto',n\_estimators=200,n\_jobs=-2,random\_state=42).**

**The final accuracy is: 93.24%**

1. We have also plotted ROC curve and comparing AUC.



1. We have saved the model using pickle

# **Concluding Remarks**

The project demonstrates the application of machine learning techniques in weather forecasting. By carefully analyzing the data, handling null values, addressing outliers and skewness, and selecting appropriate features, we were able to build robust predictive models. The ExtraTreesRegressor for Rainfall prediction and RandomForestClassifier for RainTomorrow prediction were identified as the best-performing models. These models can aid in making informed decisions and planning based on weather forecasts.

Future work can focus on further improving the model accuracy by incorporating more advanced techniques and exploring additional features.