SAMSUNG

Rain Prediction

Together for Tomorrow!
Enabling People
Education for Future Generations

Team Thunder

Salama Almansoori, Althaf Abdul Wahab

AI Course

Together for Tomorrow!
Enabling People
Education for Future Generations

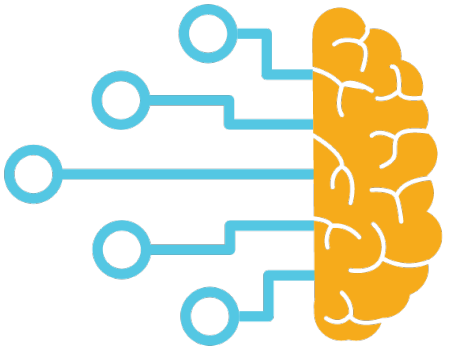
Rain Prediction

Rain Prediction

Rain prediction is an important part of weather forecasting because it assists people and organizations in reducing rain-related losses, increasing preparation, ensure water security and enhancing social benefits.

Aim

- We intend to train our AI model with 10 years of daily weather observations data so that we can anticipate rain for the next day.



About The Data

Public weather data from the Commonwealth of Australia's Bureau of Meteorology

- 10 years of daily Data
- 49 Locations across Australia
- 23 attributes including the target variable "RainTomorrow", indicating whether or not it will rain the next day

There are 23 attributes that are the daily weather parameters such as Max-Min temperature, evaporation, wind speed, rainfall, humidity, pressure, etc.

- Number of records: 145,460

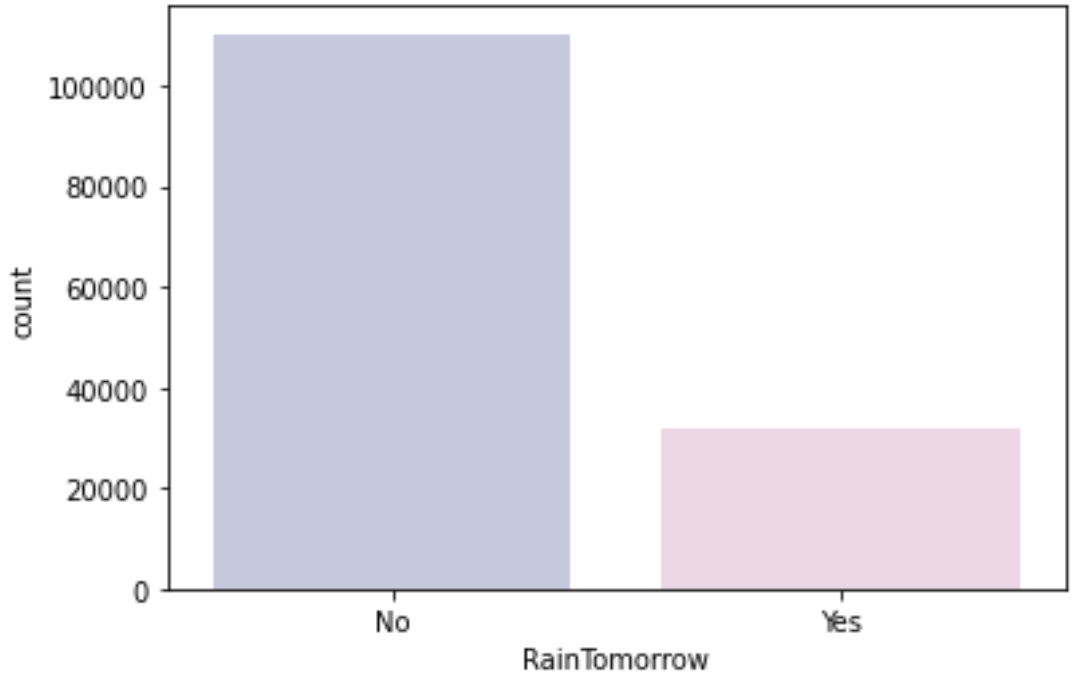
```
Data columns (total 23 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   Date           145460 non-null  object
 1   Location       145460 non-null  object
 2   MinTemp        143975 non-null  float64
 3   MaxTemp        144199 non-null  float64
 4   Rainfall       142199 non-null  float64
 5   Evaporation    82670 non-null   float64
 6   Sunshine       75625 non-null   float64
 7   WindGustDir    135134 non-null  object
 8   WindGustSpeed  135197 non-null  float64
 9   WindDir9am     134894 non-null  object
 10  WindDir3pm     141232 non-null  object
 11  WindSpeed9am   143693 non-null  float64
 12  WindSpeed3pm   142398 non-null  float64
 13  Humidity9am    142806 non-null  float64
 14  Humidity3pm    140953 non-null  float64
 15  Pressure9am    130395 non-null  float64
 16  Pressure3pm    130432 non-null  float64
 17  Cloud9am       89572 non-null   float64
 18  Cloud3pm       86102 non-null   float64
 19  Temp9am        143693 non-null  float64
 20  Temp3pm        141851 non-null  float64
 21  RainToday      142199 non-null  object
 22  RainTomorrow   142193 non-null  object
dtypes: float64(16), object(7)
```

Data head

index	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	WNW	20.0	24.0	71.0	22.0	1007.7	1007.1	8.0	NaN	16.9
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	WSW	4.0	22.0	44.0	25.0	1010.6	1007.8	NaN	NaN	17.2
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	WSW	19.0	26.0	38.0	30.0	1007.6	1008.7	NaN	2.0	21.0
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	E	11.0	9.0	45.0	16.0	1017.6	1012.8	NaN	NaN	18.1
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	NW	7.0	20.0	82.0	33.0	1010.8	1006.0	7.0	8.0	17.8

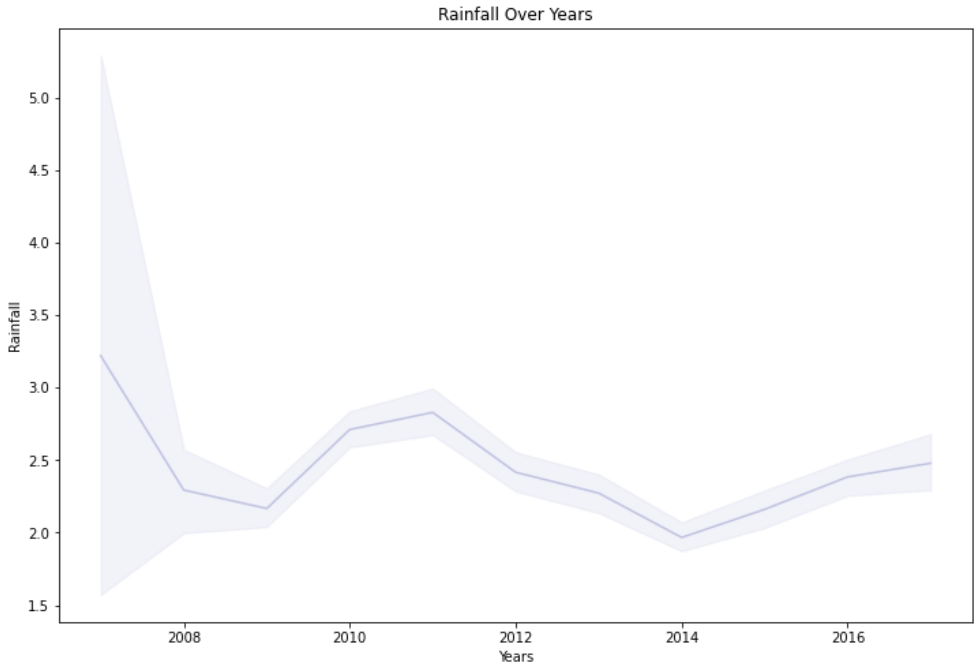
- Missing Values
- Non-Numerical Values
- Some correlation between the data such as evaporation & humidity
- Large Values (scaling)

Checking for Data imbalance



Rainfall over years

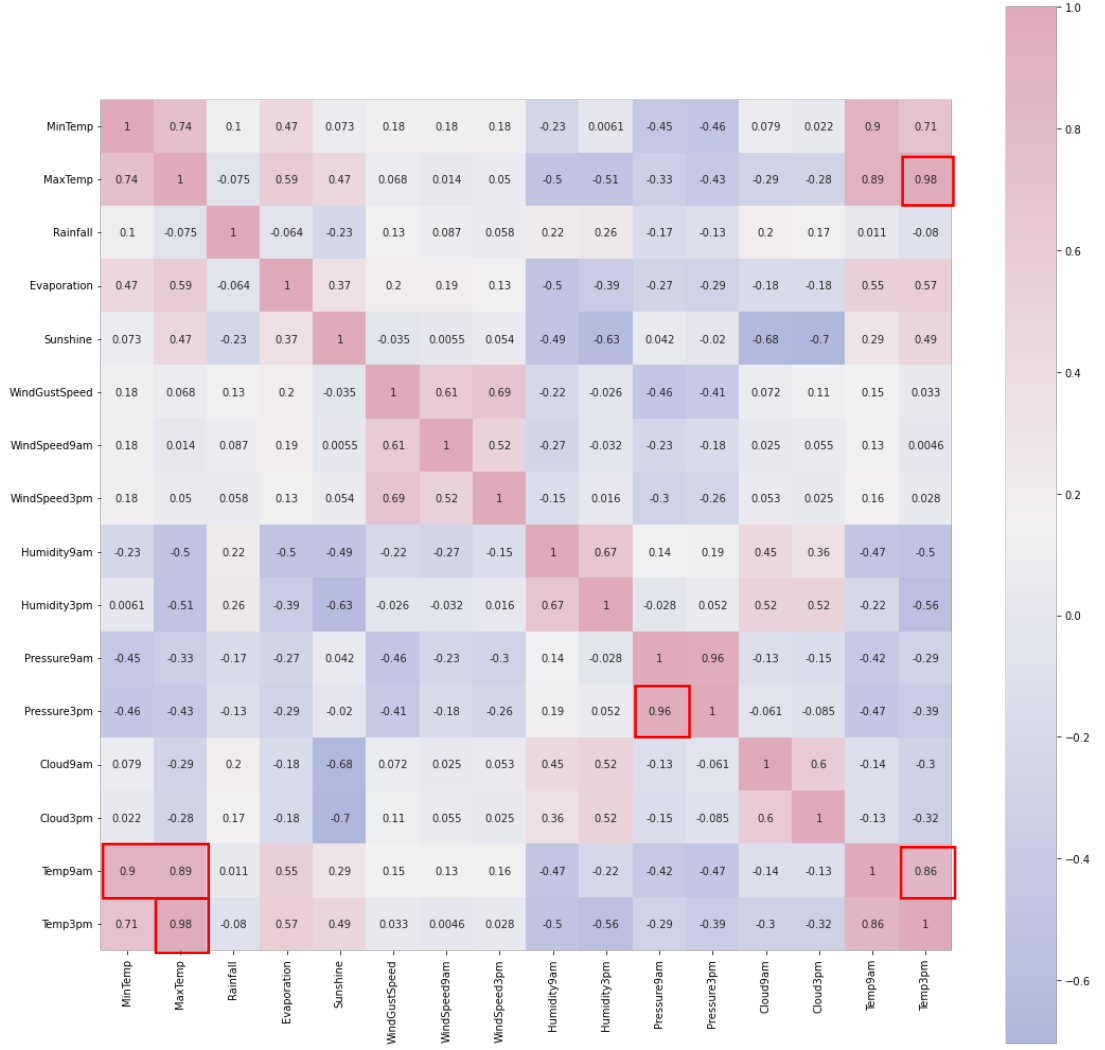
Parse dates into date times, by creating columns from date as year month and day of month



Data Correlation Matrix

We have highly correlated values >89%
Need to avoid multi-collinearity

- MaxTemp vs Temp9am
- Temp3pm vs Temp9am
- Pressure3pm vs Pressure9am
- MaxTemp vs Temp3pm

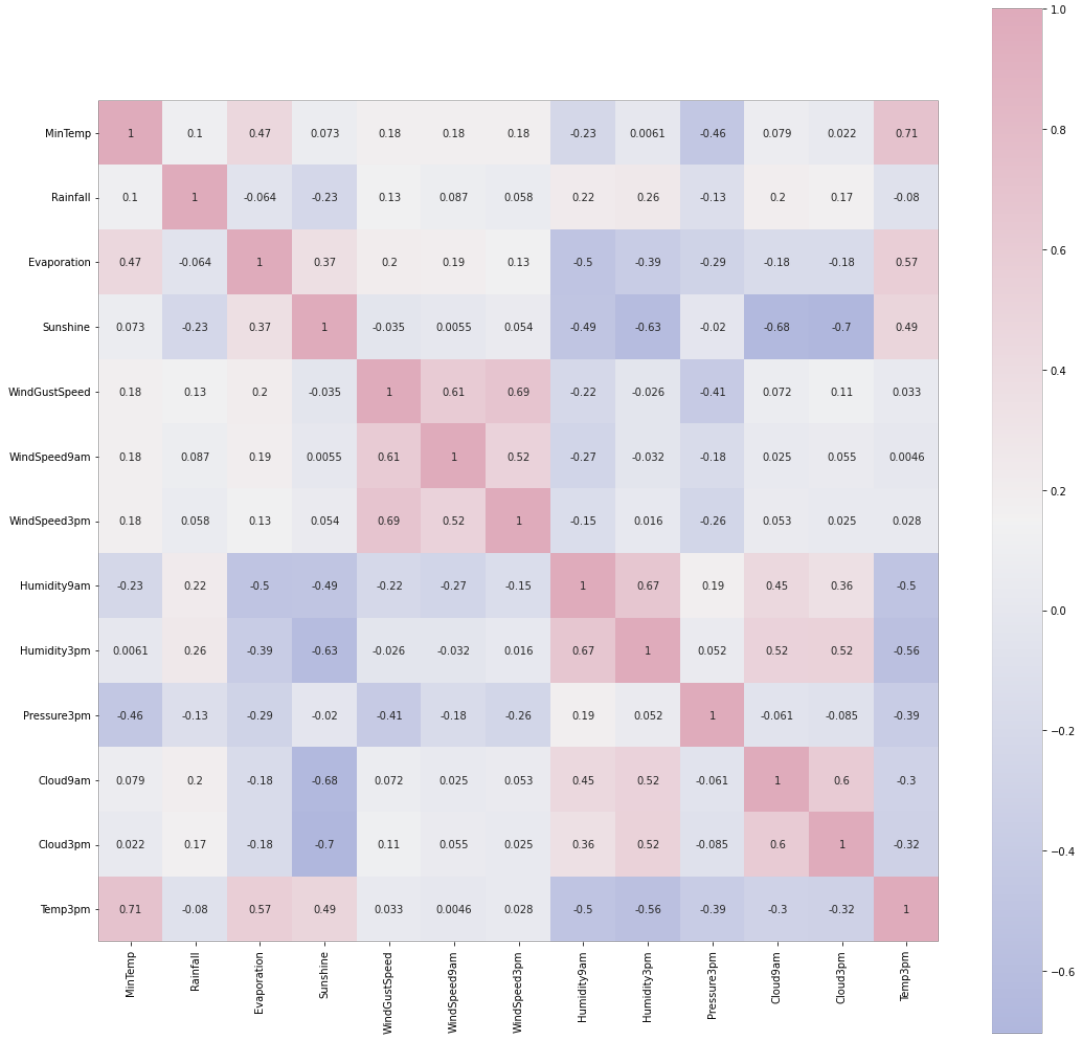


Data Preprocessing

- Removing Highly correlated data
- Removing columns with 40% missing values.
- Fill the missing or Nan values with mean and mode.
- Label encoding the categorial variables
- Preparing the data for scaling, removing outliers

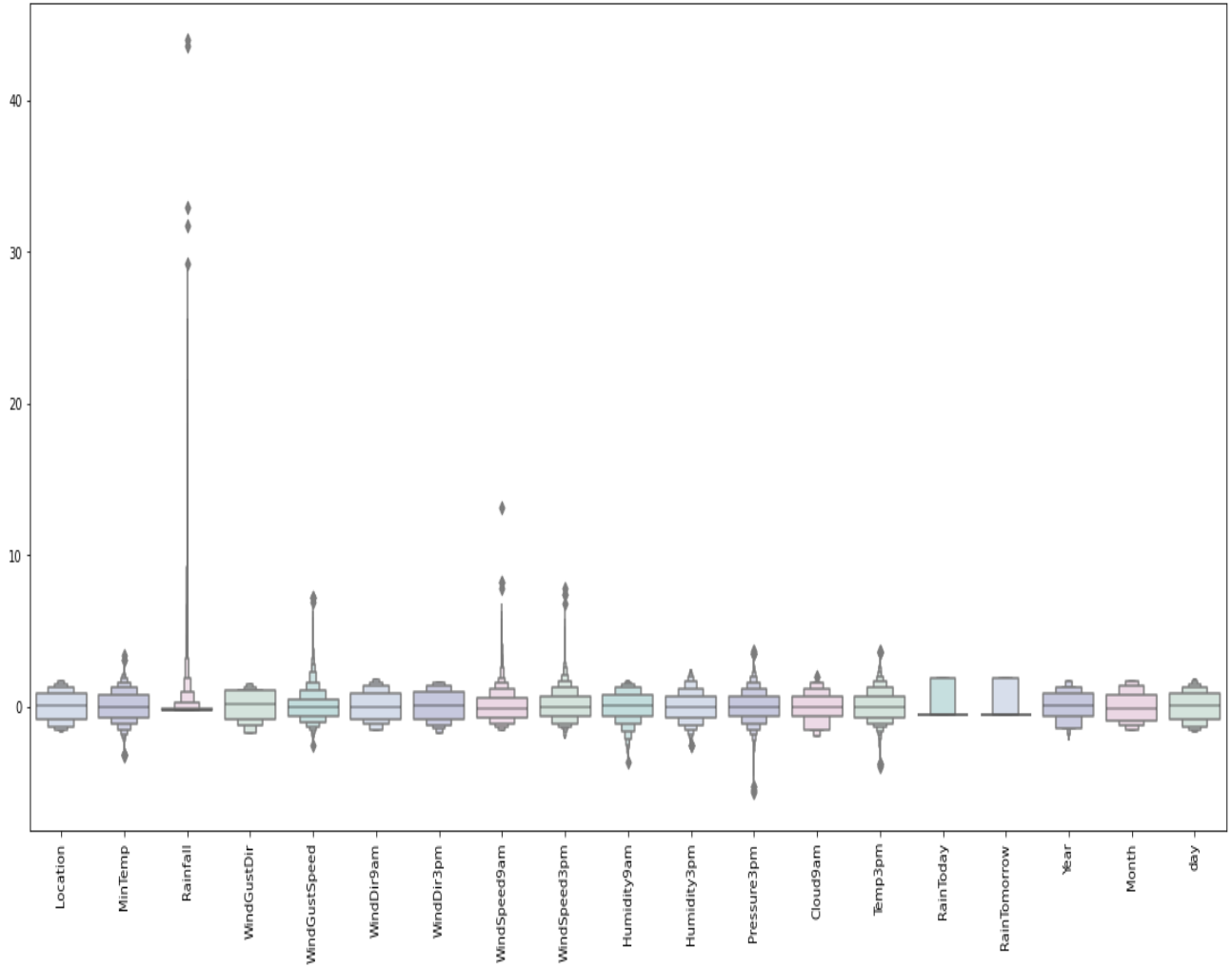
Data Correlation

- Removing the highly correlated values >=0.89:
  - Temp9am
  - Pressure9am
  - Temp3pm



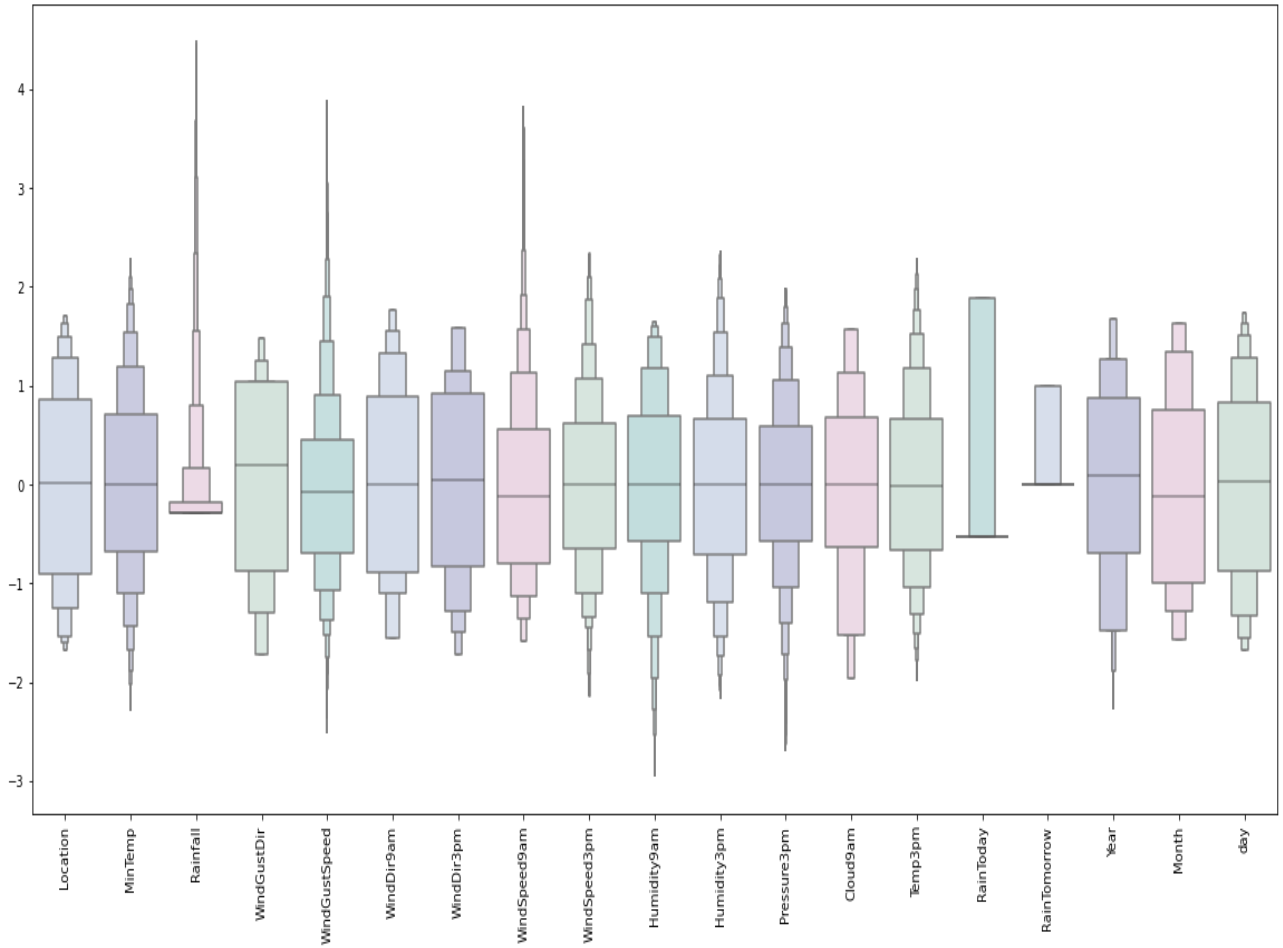
Data Preprocessing

- Outliers present



Data Preprocessing

- After removing the outliers



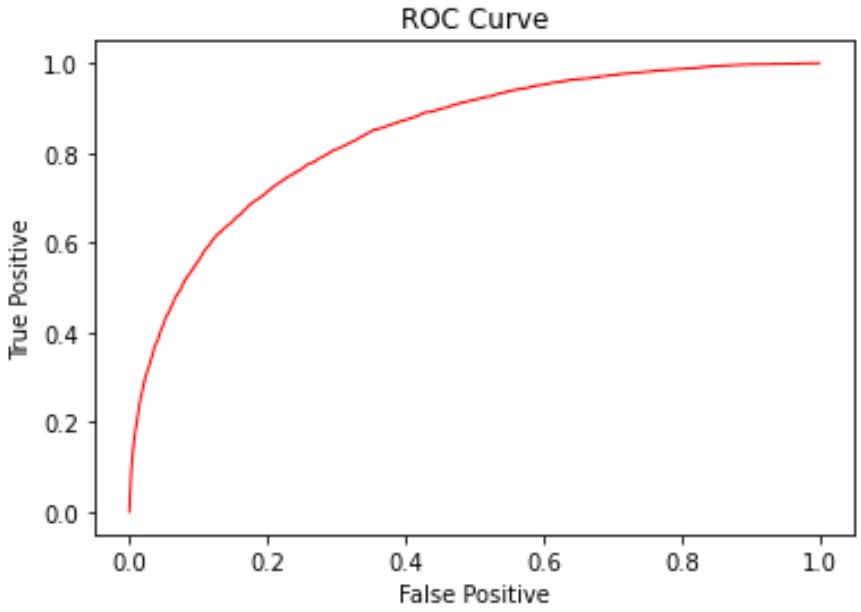
Model Building

- Linear Regression Model
- KNN Neighbors
- Sequential Neural Network

Linear Regression

Following steps are involved in the model building

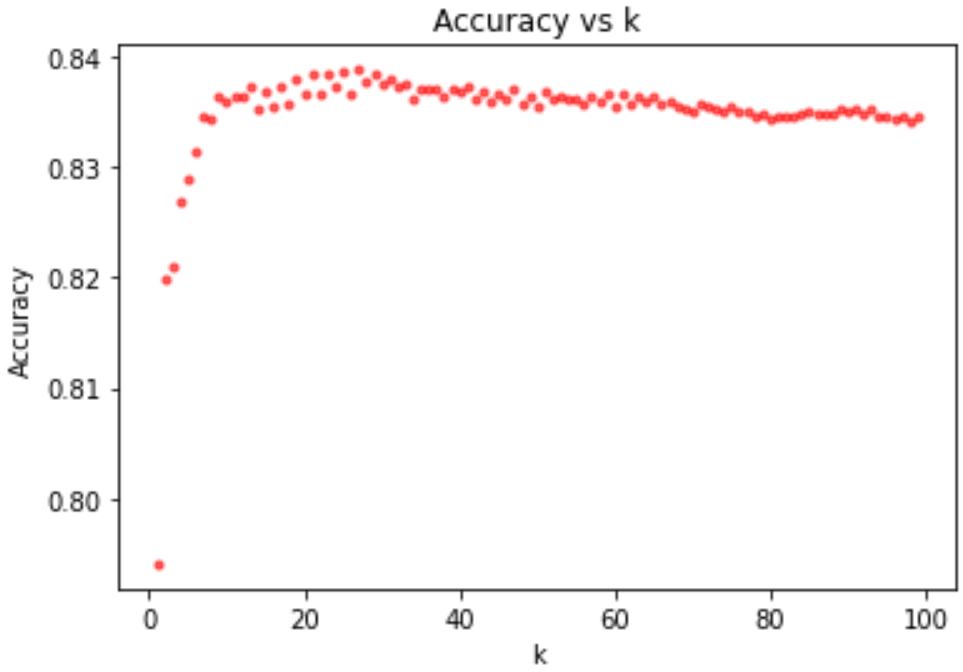
- Assigning X and y the status of attributes and tags
- Splitting test and training sets
- Train and predict
- Accuracy, Sensitivity, Specificity and Precision using the confusion matrix
- ROC curve
- **Accuracy = 0.839**
- **Sensitivity = 0.429**
- **Precision = 0.688**
- **AUC = 0.841**



KNN Neighbor

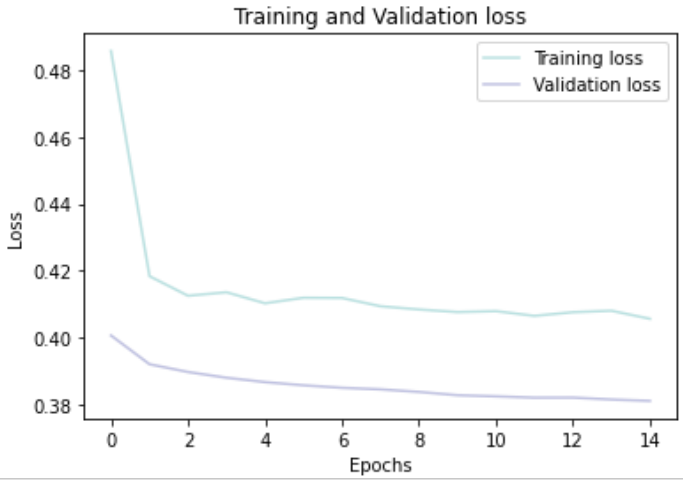
Following steps are involved in the KNN model building

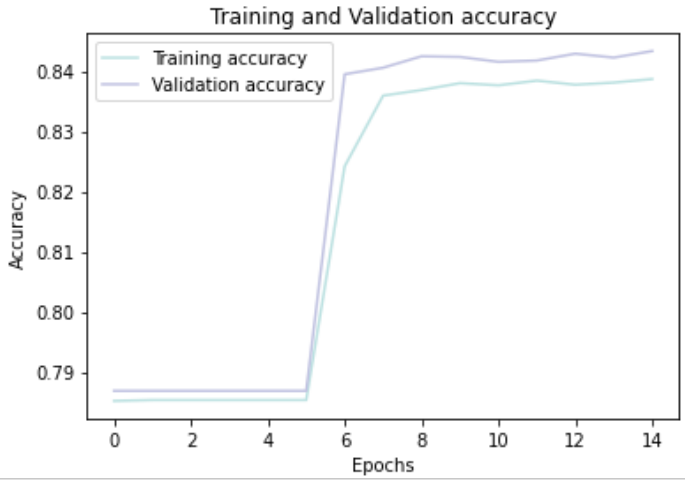
- KNN bias-Variance tradeoff as function of k
- KNN hyperparameter optimization
- **Best k : 25**
- **Best Accuracy : 0.839**



Sequential Neural Network

- Following steps are involved in the Neural Network model building
  - Initializing the neural network using **Sequential API Keras model**
  - Defining by adding layers, ReLu activation function, Binary Cross entropy loss function and Adam Gradient descent algorithm to optimize
  - Compiling the neural network
  - Train the neural network



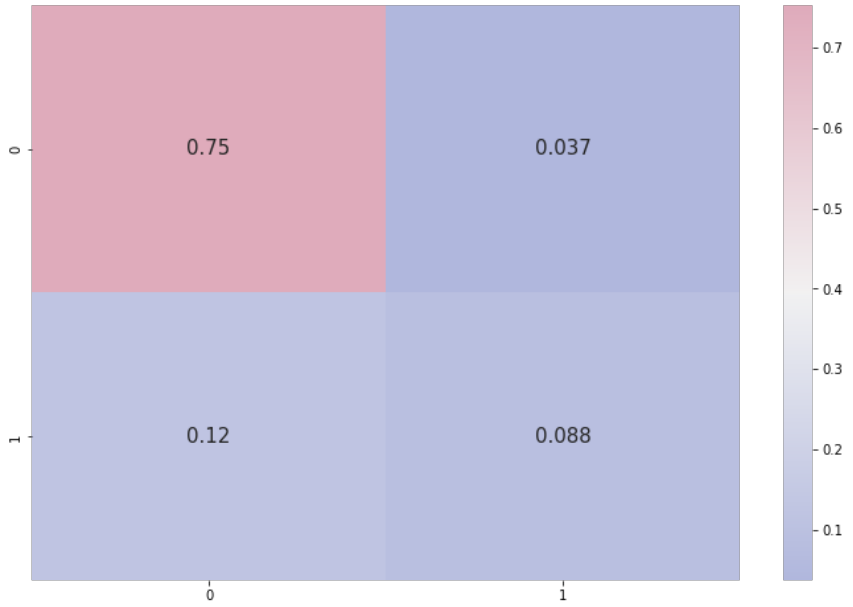


Concluding the NN:

- Testing on the test set
- Evaluating the confusion matrix
- Evaluating the classification report
- **Accuracy = 0 . 84058**

	precision	recall	f1-score	support
0	0.86	0.95	0.90	31192
1	0.70	0.42	0.52	8291
accuracy			0.84	39483
macro avg	0.78	0.69	0.71	39483
weighted avg	0.83	0.84	0.82	39483

Classification report



Confusion Matrix

Projected Impact

We have achieved our objective set out to create an **AI model for rain predictions** and hope to improve it with more data as per user case. We also managed to illustrate and compare logistic regression model and KNN neighbor model.

We hope to work on and create the below for a comprehensive sustainable solution.
1. Cloud detection and segmentation from satellite images
2. Algorithm to predict and detect moisture-laden clouds and deploy cloud precipitation resources
3. Real-time prediction of hostile weather using big data

SAMSUNG

Together for Tomorrow!

Enabling People

Education for Future Generations