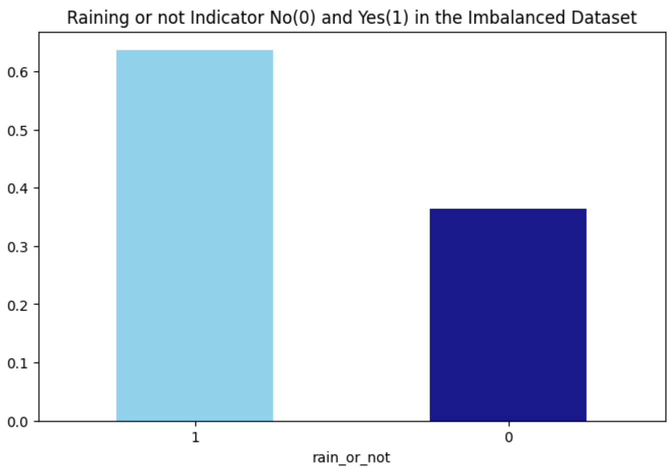
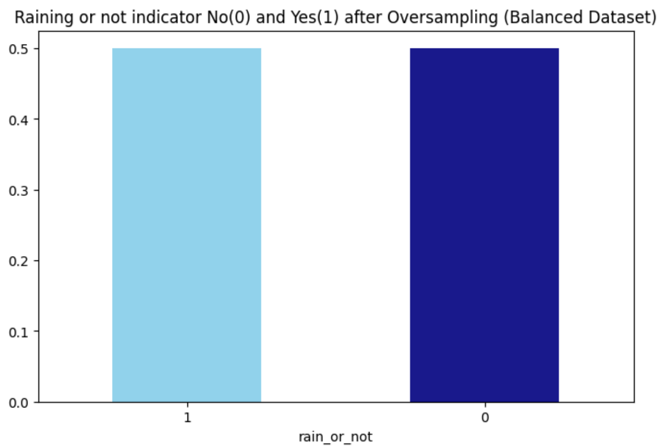
Task 01 - Weather Forecasting
Team Duo Dynamics

	avg_temperature	humidity	avg_wind_speed	cloud_cover	pressure
count	296.000000	296.000000	296.000000	296.000000	311.000000
mean	25.983840	55.041385	7.556636	49.834827	1001.059119
std	6.802475	19.220133	5.344683	29.009459	28.835595
min	15.000000	30.000000	0.069480	0.321826	951.240404
25%	20.265692	34.280826	3.550354	24.530951	975.757545
50%	27.177958	56.759806	7.326421	50.725120	1001.938586
75%	32.204599	72.189837	11.050627	76.046506	1026.578884
max	35.000000	90.000000	56.636041	99.834751	1049.543752

Handling class imbalancing

Upsampled the class with least number of labels



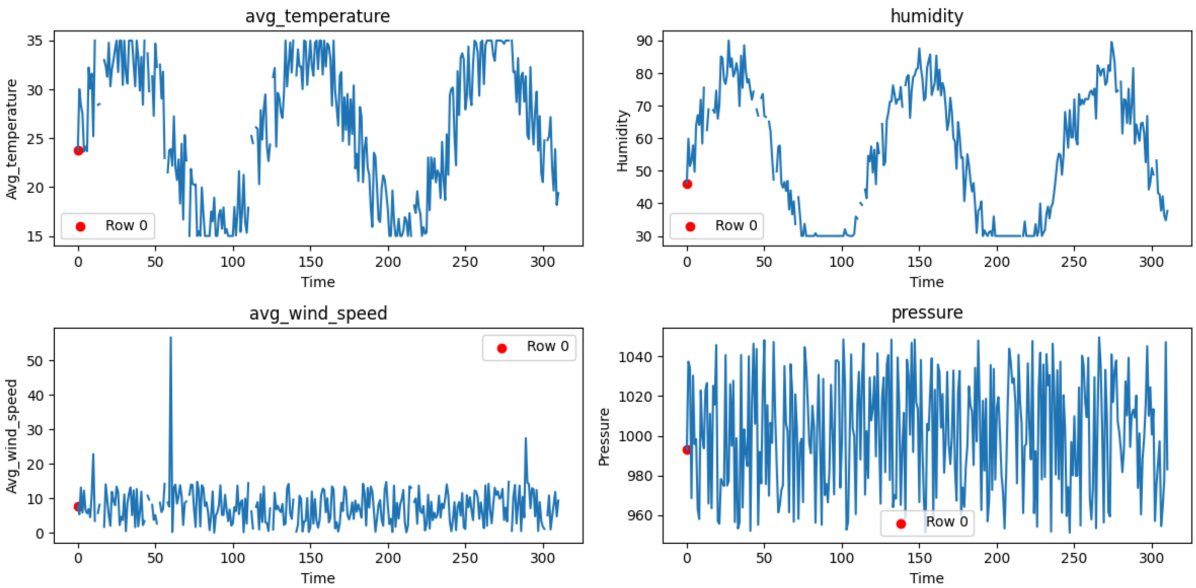


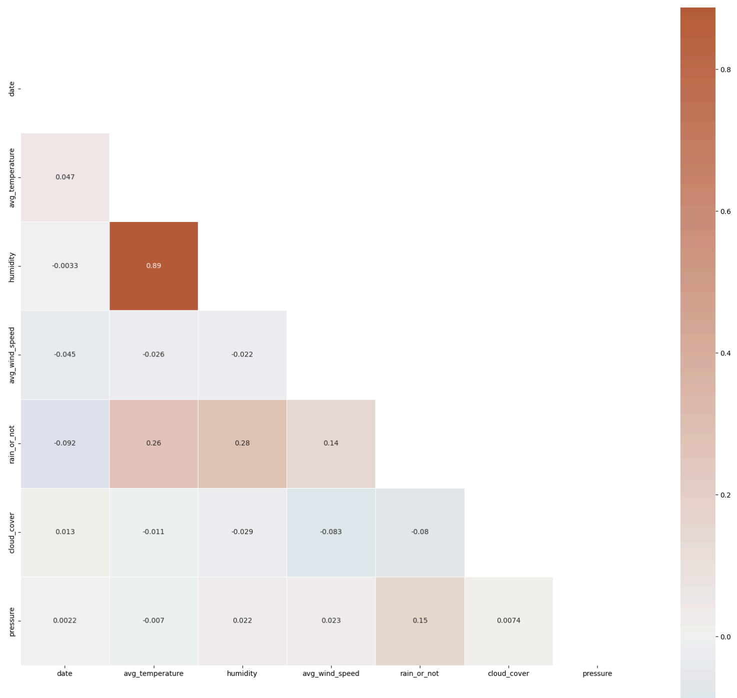
Handling missing values

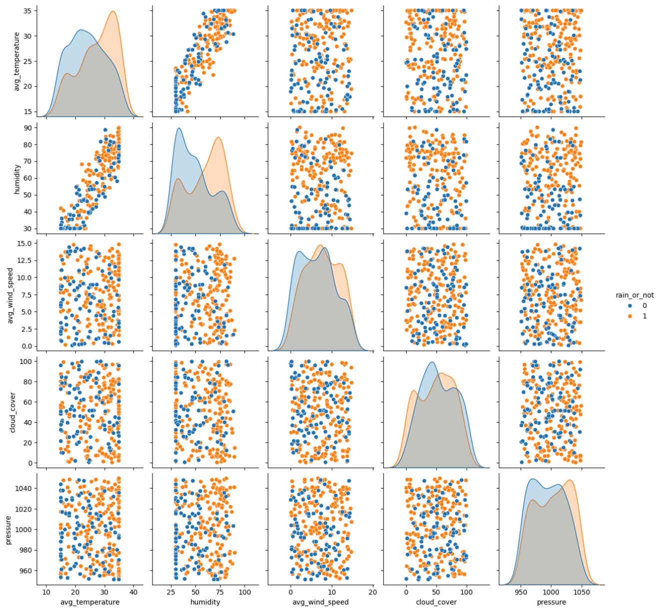
- Imputed categorical variable ‘date’ with mode
- Used Multiple Imputation by Chained Equations (MICE) is a statistical method for handling missing data by creating multiple plausible replacements for each missing

value. It uses a sequence of regression models, where each missing value is predicted based on the observed and currently imputed values of other variables.

- This "chained" approach accounts for relationships between variables, leading to more accurate and robust estimates compared to simpler imputation methods.
- By generating multiple imputed datasets and analyzing them separately, MICE incorporates the uncertainty associated with missing data, resulting in more reliable and valid statistical inferences.







No linear correlation was shown by the data as appears in the pairplot and the correlation matrix.

Data Preprocessing

Other Data preprocessing steps that we have taken

- The categorical target variable rain_or_not (values: 'No Rain', 'Rain') was converted into numerical format (0, 1) using the `replace()` function.

- To address class imbalance, oversampling was applied to ensure a balanced distribution bet
- Categorical variables were transformed into numerical representations using Label Encoding, making them suitable for model training.
- The Interquartile Range (IQR) method was employed to identify and remove outliers, preventing them from negatively impacting model performance.
- MinMaxScaler was used to standardize numerical features, ensuring all variables were on a similar scale, which enhanced model stability and convergence.

Model Selection

Three machine learning algorithms were considered for weather prediction due to their strong performance in classification tasks:

- **Random Forest**
- **CatBoost**
- **XGBoost**

Model Evaluation Metrics

To assess model performance, the following metrics were used:

1. **Accuracy** – Measures the proportion of correctly classified instances.
2. **ROC AUC** – Evaluates the model's ability to distinguish between classes.
3. **Cohen's Kappa** – Assesses agreement between predicted and actual classifications.
4. **Training Time –** Records the time taken for model training and prediction.
5. **Classification Report** – Includes precision, recall, F1-score, and support for each class.
6. **Confusion Matrix** – Provides a visual representation of model predictions against actual values

After evaluating all models, **CatBoost achieved the highest evaluation metrics** across accuracy, ROC AUC, Cohen’s Kappa, and F1-score. Due to its superior performance, **CatBoost was selected as the primary model for weather prediction**

Model	Random Forest	Catboost	XGBoost
Accuracy	0.6565	0.748	0.61
ROC AUC	0.6556	0.75	0.61
Cohens Kappa	0.312	0.49	0.23

System Design for Rain Prediction System