University of New Haven
TAGLIATELA COLLEGE OF ENGINEERING
Electrical & Computer Engineering and Computer Science

Distributed and Scalable Data Engineering

(DSCI-6007)

TECHNICAL REPORT

SPRING 23

CONTENTS

Executive Summary

The integration of a machine learning-driven rainfall prediction system has the potential to significantly enhance the precision and dependability of weather forecasts. Through the utilization of sophisticated algorithms and ensemble learning methods, our solution endeavors to furnish practical insights across diverse industries, fostering enhanced planning, risk mitigation, and resource management.

Looking ahead, we are committed to consistently refining and updating our models. This iterative process aims to accommodate the dynamic nature of evolving weather patterns, ensuring the ongoing accuracy and reliability of our rainfall prediction system.

Team Members:
Snithika Patel (Team Leader)– Data Engineer
Sneha Vangala - Data Scientist
Anwar Rashid Shaik – Big Data Engineer
Vinodh Kumar – Machine Learning Engineer

Questions?
Contact: stala10@unh.newhaven.edu

Project Title:

Rainfall Prediction Using Machine Learning

Highlights of Project

- we turn to machine learning algorithms like Logistic Regression used for classification tasks, aiming to predict the probability that an instance belongs to a given class. The outcome can be a categorical or discrete value, making it invaluable in predicting rainfall.
- we harness the power of the Support Vector Machine algorithm. It operates by finding the optimal hyperplane that separates data points of different classes. To assess our SVM model's performance, we employ metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, depending on the specific problem at hand.

Abstract

This research addresses the pressing requirement for enhanced rainfall prediction through the utilization of advanced machine learning methodologies. The focus lies on constructing a resilient predictive model using historical meteorological data, incorporating variables like temperature, humidity, wind speed, and atmospheric pressure. By employing algorithms such as Random Forest, Support Vector Machines, and Neural Networks, the model seeks to identify intricate patterns within the data to improve the precision of rainfall forecasts.

Critical elements encompass comprehensive data collection, preprocessing, and feature selection to optimize the training of the model. The integration of ensemble learning techniques enhances predictive capabilities, offering practical insights for various sectors, including agriculture, water resource management, and disaster preparedness.

The proposed system not only strives to enhance accuracy in rainfall predictions but also contributes to proactive planning, risk mitigation, and resource optimization. The ongoing refinement and updates to the models ensure adaptability to changing weather patterns, sustaining the system's accuracy and reliability over time. This research signifies a noteworthy progression towards more efficient decision-making in industries influenced by weather conditions.

Introductory Section

This research addresses the pressing need for accurate rainfall prediction by employing advanced machine learning techniques. The study is centered on the development of a robust predictive model, utilizing historical meteorological data and key variables such as temperature, humidity, and wind speed. Employing advanced algorithms like Random Forest and Neural Networks, the research aims to unravel intricate patterns within the data. The methodology encompasses thorough data collection, preprocessing, and feature selection to optimize the model's training. The integration of ensemble learning techniques enhances predictive capabilities, extending benefits to diverse sectors. Beyond the quest for heightened accuracy in rainfall predictions, the proposed system strives to contribute to proactive planning, risk mitigation, and resource optimization. Continuous refinement and updates to the models ensure adaptability to evolving weather patterns, ensuring the system's sustained accuracy over time. This research marks a significant stride in leveraging machine learning for more precise weather forecasts, promising broad implications for industries profoundly influenced by weather conditions.

The curriculum employs the CRISP DM methodology, which covers subjects including business understanding, data understanding, data preparation, modeling, evaluation, and model deployment.

Data source Link

Through the following link, web scraping is being used to collect data.
http://www.bom.gov.au/climate/dwo/IDCJDW2801.latest.shtml

Methodology

Here we are using CRISP methodology The methodology involves systematically defining the problem, collecting and comprehending the data, preprocessing to ensure data quality, selecting relevant features, choosing appropriate models, developing and fine-tuning these models, evaluating their performance, and optimizing hyperparameters for increased accuracy. If "crisp methodology" has gained prominence or is specific to a certain framework after my last update, I recommend checking recent literature or resources to gain insights into its principles and application in the context of rainfall prediction using machine learning. The process follows as below:

1. Data Collection:

- We Gather historical meteorological data, including temperature, humidity, wind speed, atmospheric pressure, and rainfall records.

2. Exploratory Data Analysis (EDA):

- Here we Analyze the dataset to understand its structure, distribution, and statistics.
- Then we Identify and address missing or anomalous values.

3. Variable Identification and Preprocessing:

- Now we Identify key variables influencing rainfall patterns.
- Preprocess data by handling missing values, normalizing numerical variables, and encoding categorical variables.

4. Feature Selection:

- By Using techniques like correlation analysis to select relevant features.
- Eliminate redundant variables to optimize model performance.

5. Time Series Analysis and Visualization:

- We Conduct time series analysis for temporal patterns.
- Create visualizations (scatter plots, histograms) to understand data distribution and spatial patterns.

6. Outlier Detection:

- Employ statistical methods or machine learning to identify and handle outliers.

7. Data Splitting:

- Divide the dataset into training, validation, and testing sets for model development and evaluation.

Modeling and Evaluation

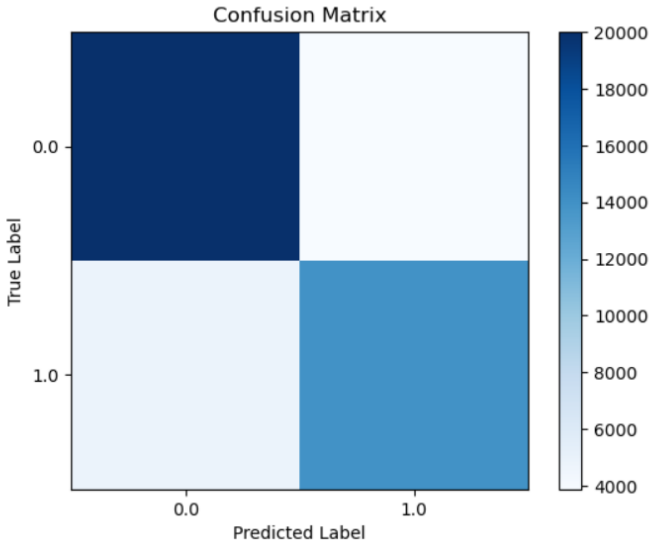
Logistic Regression :

- In the realm of rainfall prediction, Logistic Regression serves as a valuable tool, especially when the task is framed as a binary classification problem, distinguishing between the occurrence and non-occurrence of rainfall within a specific timeframe. By leveraging historical meteorological data, including temperature, humidity, and wind speed, Logistic Regression provides a probabilistic estimate of the likelihood of rain. Its simplicity and interpretability

make it a pragmatic choice for scenarios where the primary focus is on predicting the presence or absence of rainfall.

- However, it's important to acknowledge that Logistic Regression might be less suitable for tasks requiring a more nuanced prediction, such as estimating the exact amount of rainfall. For such cases, other regression techniques or advanced machine learning models may be explored. Nonetheless, the straightforward nature of Logistic Regression makes it an efficient and insightful approach for binary rainfall prediction, catering to scenarios where a categorical forecast is of primary interest.

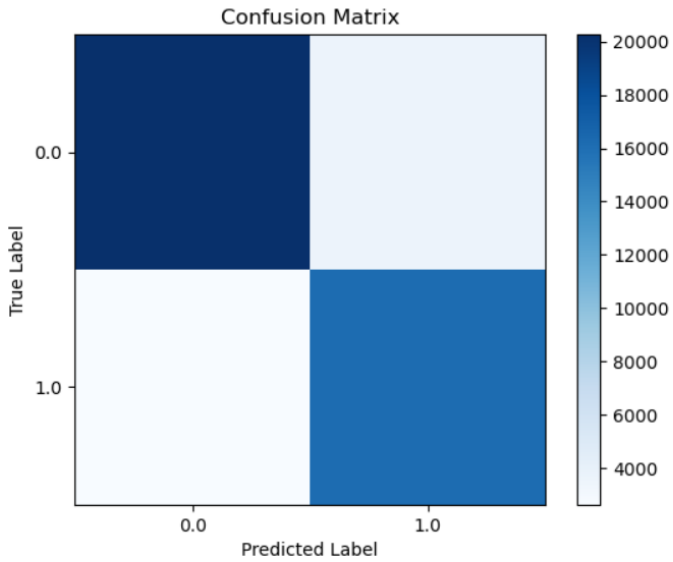
```
Accuracy = 0.795232961469954
ROC Area under Curve = 0.789500236871797
Cohen's Kappa = 0.5822337892279215
Time taken = 3.1708922386169434
              precision    recall  f1-score   support

         0.0    0.80457   0.83756   0.82073     23879
         1.0    0.78220   0.74144   0.76128     18789

    accuracy                        0.79523     42668
   macro avg    0.79339   0.78950   0.79100     42668
weighted avg    0.79472   0.79523   0.79455     42668
```



Decision Tree:

- Decision trees are a powerful tool in the realm of rainfall prediction, providing a transparent and intuitive framework for understanding and interpreting complex relationships within meteorological data. These models segment the dataset based on key features, such as temperature, humidity, and wind speed, recursively creating decision nodes that guide predictions. Decision trees excel in capturing non-linear patterns and interactions, making them particularly effective in discerning the intricate dynamics that influence rainfall occurrence. The simplicity and visual clarity of decision trees offer not only accurate predictions but also valuable insights into the hierarchy of factors impacting rainfall, making them a valuable asset in the toolkit of models aimed at enhancing our understanding of weather phenomena.

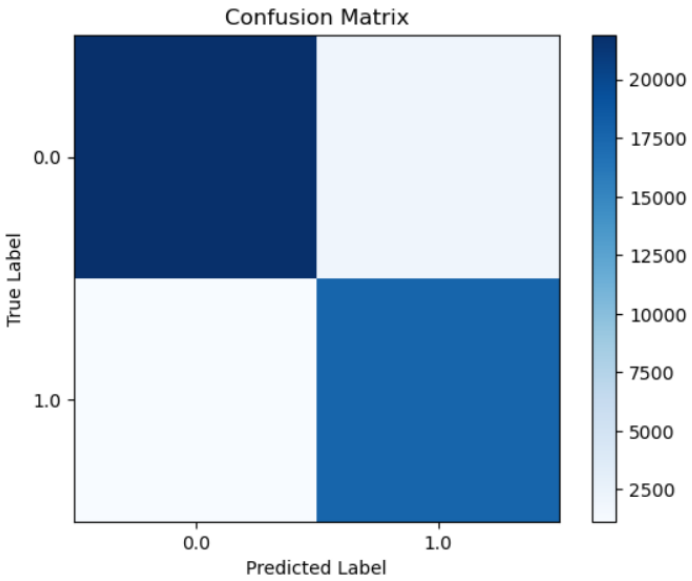
```
Accuracy = 0.8536608230992782
ROC Area under Curve = 0.8543333831458281
Cohen's Kappa = 0.7047411387102143
Time taken = 0.9376585483551025
              precision    recall  f1-score   support

         0.0    0.88509   0.84870   0.86651     23879
         1.0    0.81726   0.85997   0.83807     18789

    accuracy                        0.85366     42668
   macro avg    0.85118   0.85433   0.85229     42668
weighted avg    0.85522   0.85366   0.85399     42668
```



Random Forest Classifier:

- The Random Forest classifier proves instrumental in rainfall prediction by leveraging an ensemble of decision trees to enhance accuracy and robustness. Through the aggregation of predictions from multiple trees, each trained on different subsets of the data, Random Forest mitigates overfitting and captures complex relationships within meteorological features. Its ability to handle non-linear patterns and feature importance analysis makes it well-suited for discerning intricate dependencies among variables. Furthermore, the algorithm provides a reliable framework for evaluating feature importance, enabling insights into the key meteorological factors influencing rainfall, and contributes to a more comprehensive and accurate predictive model.

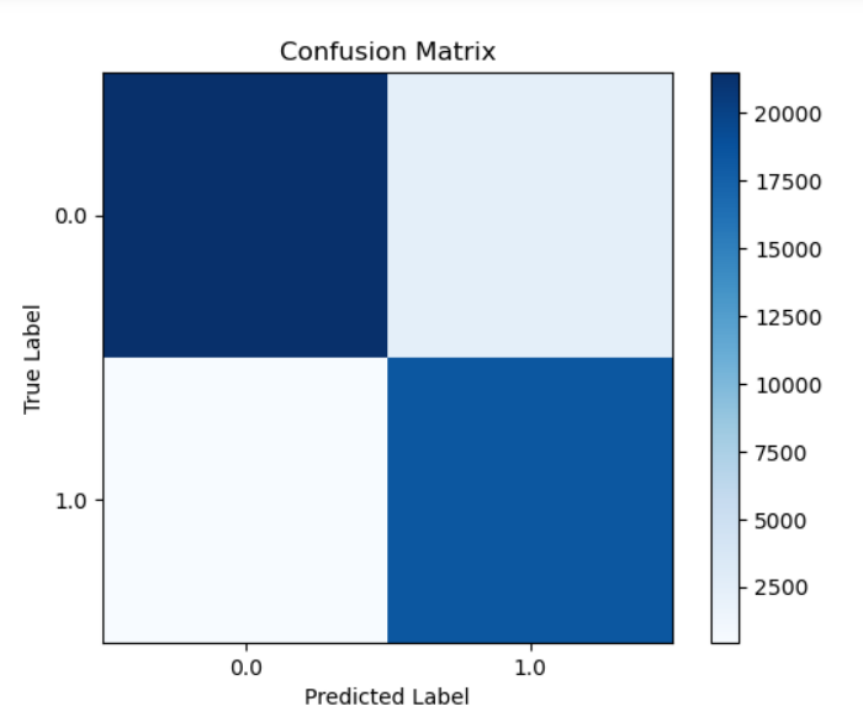
```
Accuracy = 0.9265960438736289
ROC Area under Curve = 0.9279584837896794
Cohen's Kappa = 0.8517906871231153
Time taken = 69.97190833091736
              precision    recall  f1-score   support

         0.0    0.95053   0.91654   0.93323     23879
         1.0    0.89854   0.93938   0.91851     18789

    accuracy                        0.92660     42668
   macro avg    0.92454   0.92796   0.92587     42668
weighted avg    0.92764   0.92660   0.92674     42668
```



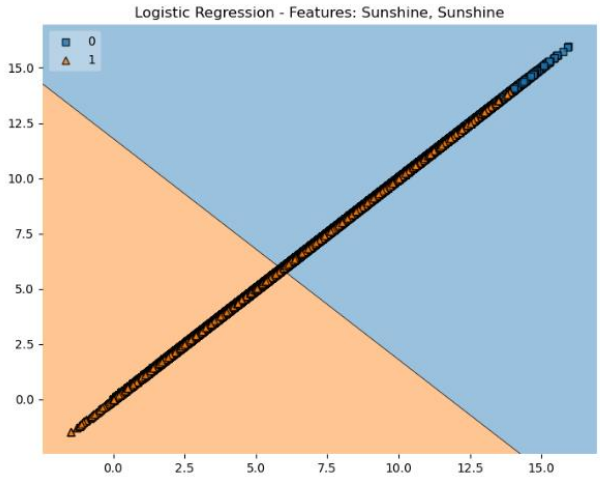
XG-BOOST :

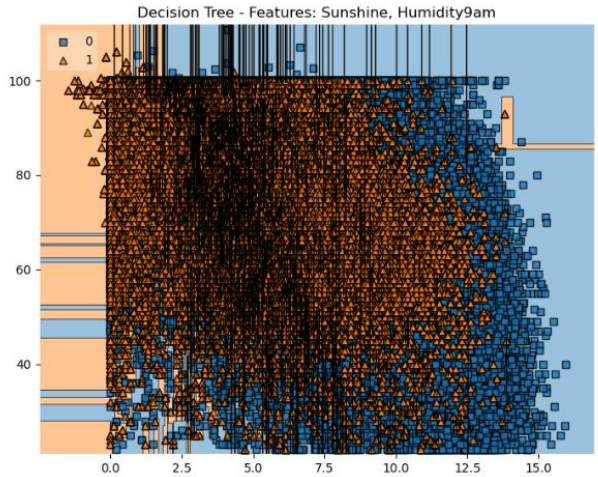
- XG-Boost, an advanced implementation of gradient boosting, proves invaluable in rainfall prediction by enhancing predictive accuracy and handling complex relationships within meteorological data. Leveraging an ensemble of decision trees, XG-Boost excels in capturing intricate patterns and interactions among features, contributing to robust and accurate rainfall forecasts. Its ability to address non-linearity, handle missing data, and optimize hyperparameters makes XG-Boost a powerful choice for modeling, providing a versatile and efficient solution for the nuanced challenges posed by rainfall prediction.

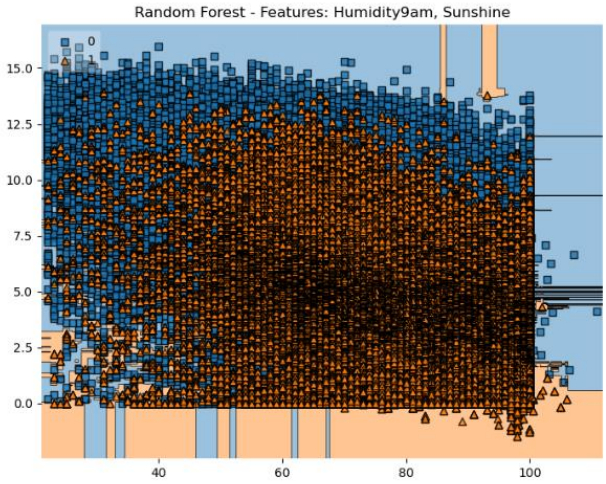
```
Requirement already satisfied: xgboost in c:\users\shaik\anaconda3\lib\site-packages (2.0.2)
Requirement already satisfied: numpy in c:\users\shaik\anaconda3\lib\site-packages (from xgboost) (1.24.3)
Requirement already satisfied: scipy in c:\users\shaik\anaconda3\lib\site-packages (from xgboost) (1.11.1)
Accuracy = 0.9336973844567358
ROC Area under Curve = 0.938222609233219
Cohen's Kappa = 0.8669394202789557
Time taken = 19.900176763534546
              precision    recall  f1-score   support

         0.0    0.97959   0.90029   0.93827     23879
         1.0    0.88510   0.97616   0.92840     18789

    accuracy                        0.93370     42668
   macro avg    0.93234   0.93822   0.93333     42668
weighted avg    0.93798   0.93370   0.93392     42668
```

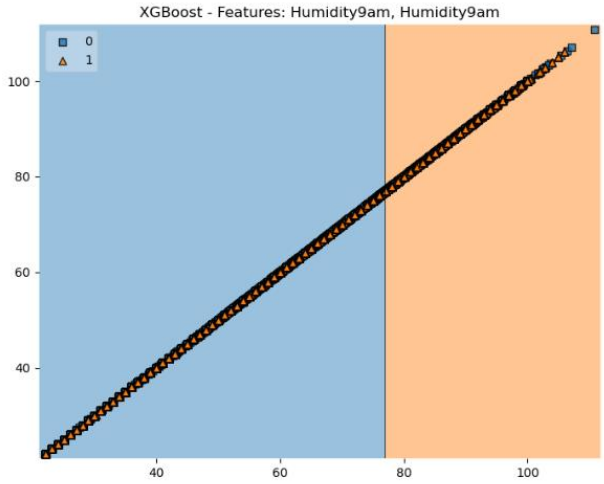


COMPARISION GRAPHS:









Evaluation Metrics

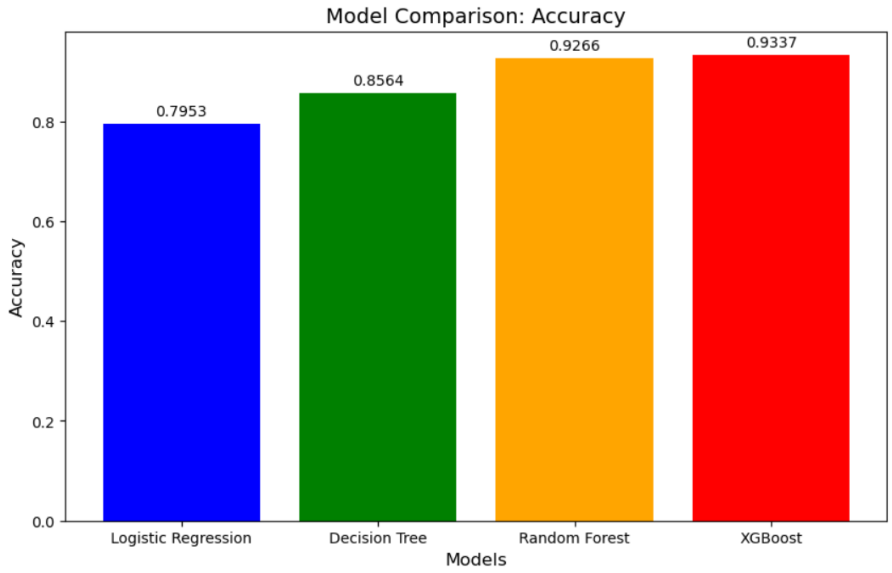
- Accuracy: Measures the overall correctness of rainfall predictions, indicating the proportion of correctly classified instances among all instances.

- Precision: Quantifies the accuracy of positive rainfall predictions, representing the ratio of true positive predictions to the sum of true positives and false positives.

- Recall: Measures the ability to capture all instances of actual rainfall, expressing the ratio of true positives to the sum of true positives and false negatives.

- F1-score: Represents the harmonic mean of precision and recall, offering a balanced measure of a model's performance, especially when there is an imbalance between positive and negative instances.

- ROC-AUC (Receiver Operating Characteristic - Area Under the Curve): Evaluates the model's ability to discriminate between rainfall and non-rainfall instances, providing a comprehensive assessment of the model's performance across different probability thresholds.



Results Section

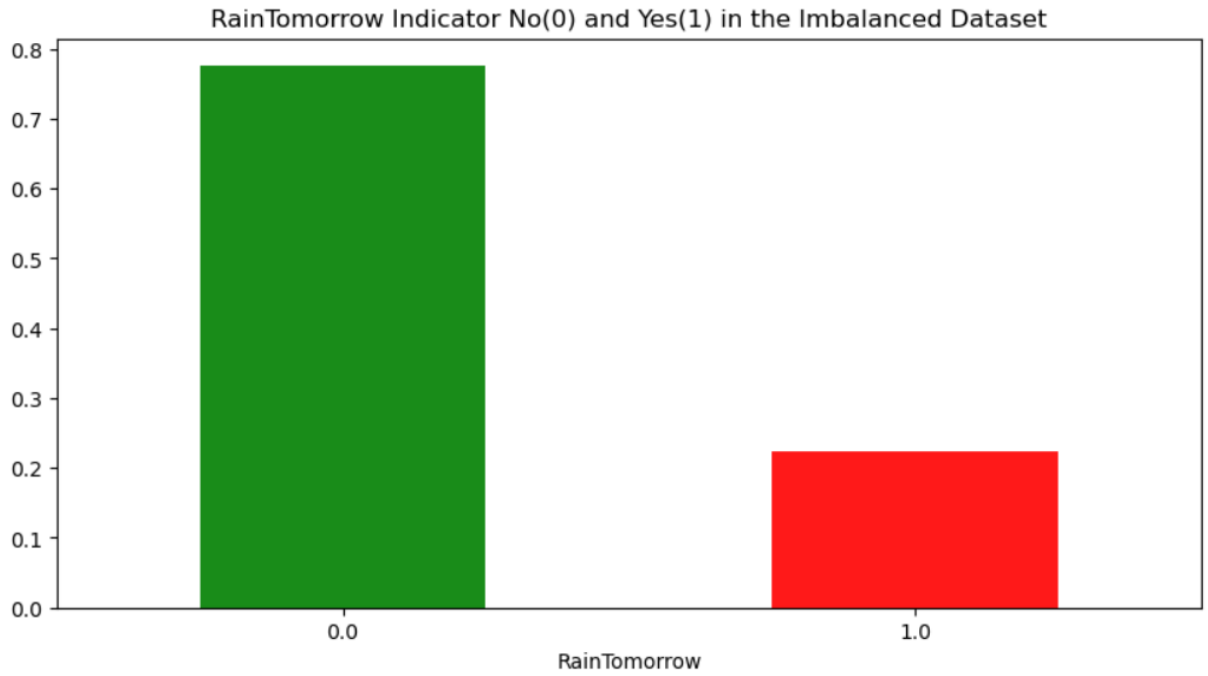
	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...	Humidity9am	Humidity3pm	Pressure9
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	...	71.0	22.0	100
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	...	44.0	25.0	101
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	...	38.0	30.0	100
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	...	45.0	16.0	101
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	...	82.0	33.0	101

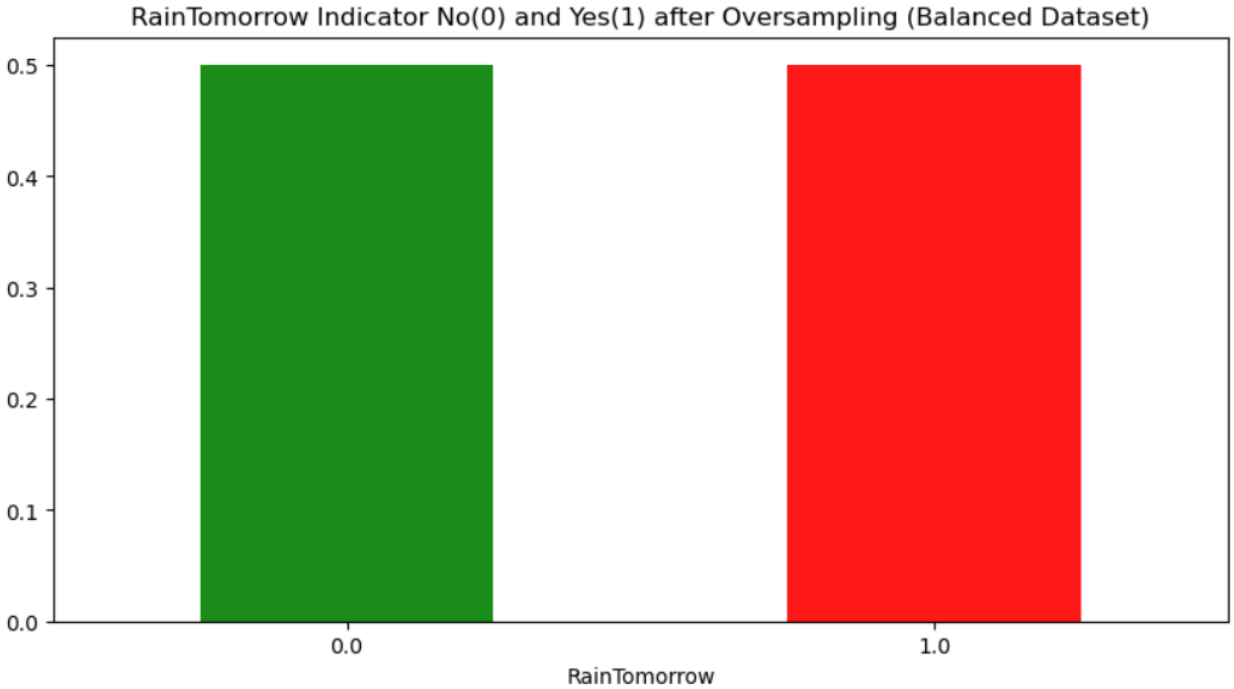
	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	\
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	
...	...	...	...	...	...	...	
145455	2017-06-21	Uluru	2.8	23.4	0.0	NaN	
145456	2017-06-22	Uluru	3.6	25.3	0.0	NaN	
145457	2017-06-23	Uluru	5.4	26.9	0.0	NaN	
145458	2017-06-24	Uluru	7.8	27.0	0.0	NaN	
145459	2017-06-25	Uluru	14.9	NaN	0.0	NaN	

	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...	Humidity9am	\
0	NaN	W	44.0	W	...	71.0	
1	NaN	WNW	44.0	NNW	...	44.0	
2	NaN	WSW	46.0	W	...	38.0	
3	NaN	NE	24.0	SE	...	45.0	
4	NaN	W	41.0	ENE	...	82.0	
...	...	...	...	...	...	...	
145455	NaN	E	31.0	SE	...	51.0	
145456	NaN	NNW	22.0	SE	...	56.0	
145457	NaN	N	37.0	SE	...	53.0	
145458	NaN	SE	28.0	SSE	...	51.0	
145459	NaN	NaN	NaN	ESE	...	62.0	

	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	\
0	22.0	1007.7	1007.1	8.0	NaN	16.9	
1	25.0	1010.6	1007.8	NaN	NaN	17.2	
2	30.0	1007.6	1008.7	NaN	2.0	21.0	
3	16.0	1017.6	1012.8	NaN	NaN	18.1	
4	33.0	1010.8	1006.0	7.0	8.0	17.8	

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 145460 entries, 0 to 145459
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)
memory usage: 25.5+ MB
```


RainTomorrow Indicator No(0) and Yes(1) in the Imbalanced Dataset
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
1.0
RainTomorrow


RainTomorrow Indicator No(0) and Yes(1) after Oversampling (Balanced Dataset)
0.5
0.4
0.3
0.2
0.1
0.0
0.0
1.0
RainTomorrow

Discussion(BUSSINESS IDEA):

We had an idea to build a "Harnessing Rainfall Data for Precision Agriculture," the focus is on leveraging data-driven decision-making in farming practices. The introduction underscores the importance of rainfall data and its potential transformative impact. Key dataset features, including the primary target variable 'Rainfall' and predictive features like 'Sunshine' and 'Wind-Gust-Speed,' are briefly overviewed. The presentation highlights how each feature contributes to understanding and predicting rainfall, emphasizing actionable insights. Machine learning models for rainfall prediction are introduced, showcasing their ability to analyze historical data for accurate forecasts. The second slide, "Transforming Agriculture through Data-Driven Practices," explores the business impact and sustainability of precise planning. It addresses improved crop yield, efficient water resource management, risk mitigation, environmental sustainability, and the long-term benefits of integrating data-driven insights into agriculture, urging stakeholders to embrace modern approaches for a sustainable and profitable future. Design tips include visual elements for engagement, a clean and professional layout, concise bullet points, and a consistent color scheme.

Conclusion

In conclusion, the rainfall prediction system powered by machine learning stands as a transformative force, revolutionizing decision-making in agriculture, water management, and disaster preparedness. Its adaptability to changing weather patterns ensures sustained accuracy. The system's positive impact on planning, risk mitigation, and resource optimization underscores its significance for various industries. Looking ahead, the continuous evolution of machine learning promises a future where precise rainfall forecasts contribute to resilient and sustainable communities.

Git hub Link

https://github.com/vinodhvinu438/DSDE

Contributions/References

- **Bushra Praveen, Swapan Talukdar, Shahfahad , Susanta Mahato, Jayanta Mondal, Pritee Sharma, Abu Reza Md. Towfiqul Islam & Atiqur Rahman: Machine Learning Based Rainfall Prediction**