DATA ANALYTICS PROJECT REPORT

on

"Rainy Day Prediction"

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN
COMPUTER SCIENCE & ENGINEERING

BY

Sreejata Banerjee	2105754
Sayan Chattopadhyay	2105748
Debarghya Roy	21051047
Semanti Das	2105149
Sachi Verma	21052917

UNDER THE GUIDANCE OF

Dr. Sricheta Parui

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

March 2024

DATA ANALYTICS PROJECT REPORT

on

"Rainy Day Prediction"

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN
COMPUTER SCIENCE & ENGINEERING

BY

Sreejata Banerjee	2105754
Sayan Chattopadhyay	2105748
Debarghya Roy	21051047
Semanti Das	2105149
Sachi Verma	21052917

UNDER THE GUIDANCE OF
Dr. Sricheta Parui

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

March 2024

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

"RAINY DAY PREDICTION"

submitted by

Sreejata Banerjee	2105754
Sayan Chattopadhyay	2105748
Debarghya Roy	21051047
Semanti Das	2105149
Sachi Verma	21052917

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering OR Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during the year 2024-2025, under our guidance.

Date: / /

Sricheta Parui
Project Guide

Acknowledgements

We are profoundly grateful to **Sricheta Parui** of **Affiliation** for her expert guidance and continuous encouragement throughout to see that this project meets its target since its commencement to its completion. ....

Sreejata Banerjee
Sayan Chattopadhyay
Debarghya Roy
Semanti Das
Sachi Verma

ABSTRACT

The "Rainy Day Prediction" project focuses on predicting whether it will rain tomorrow by utilizing the Rainfall in Australia dataset. Various machine learning models such as Gaussian Naive Bayes, KNN, Random Forest,Logistic Regression and Support Vector Classifier, among others, were tested, with Random Forest performing exceptionally well, achieving an accuracy of 0.85 and a ROC score far superior to other models. Due to system compatibility limitations, hyperparameter tuning was not conducted, although it is strongly recommended for optimal performance.

The project's tech stack includes HTML, CSS, Bootstrap for the front-end, Flask for the back-end, and development carried out in Jupyter notebook and PyCharm. To run the app, users need to create a virtual environment, activate it, install necessary packages, and then execute the app. The workflow involved data collection from the Rainfall Prediction in Australia dataset, and preprocessing steps like handling missing values, categorical values, outliers, feature selection, and scaling. Imbalanced datasets were addressed using SMOTE.

Model creation encompassed testing various models like Random Forest, Logistic Regression, Support Vector Machines, KNN, and Naive Bayes, with Logistic Regression, Random Forest, and Support Vector Machines emerging as the top performers. The project's conclusions were drawn based on classification metrics, ROC curve analysis, and AUC scores, highlighting the effectiveness of the predictive models in determining rainfall patterns

Keywords: Preprocessing, Machine Learning, Data Analytics, Hyperparameters, Classifiers - Random Forest, Logistic Regression, Gaussian Naive Bayes

Contents:

List of Figures

Chapter 1

Introduction

In the realm of weather forecasting, accurate prediction of rainfall plays a pivotal role in various sectors including agriculture, transportation, and disaster management. The ability to forecast whether it will rain tomorrow with precision is crucial for making informed decisions and mitigating potential risks associated with unexpected weather events. Traditional methods of rainfall prediction often rely on simplistic models and historical data, which may not capture the complex dynamics of weather patterns effectively. In recent years, advancements in Machine Learning (ML) and Data Analytics (DA) have offered promising avenues for improving the accuracy and reliability of rainfall forecasts.

This report introduces a novel approach to rainfall prediction that harnesses the power of ML and DA techniques. By leveraging historical weather data, real-time atmospheric observations, and a range of relevant meteorological parameters, our model aims to provide more accurate and timely predictions of rainfall. Furthermore, we introduce a PowerBI Dashboard as part of our framework, enabling users to visualize and interact with the forecasted data in a user-friendly manner.

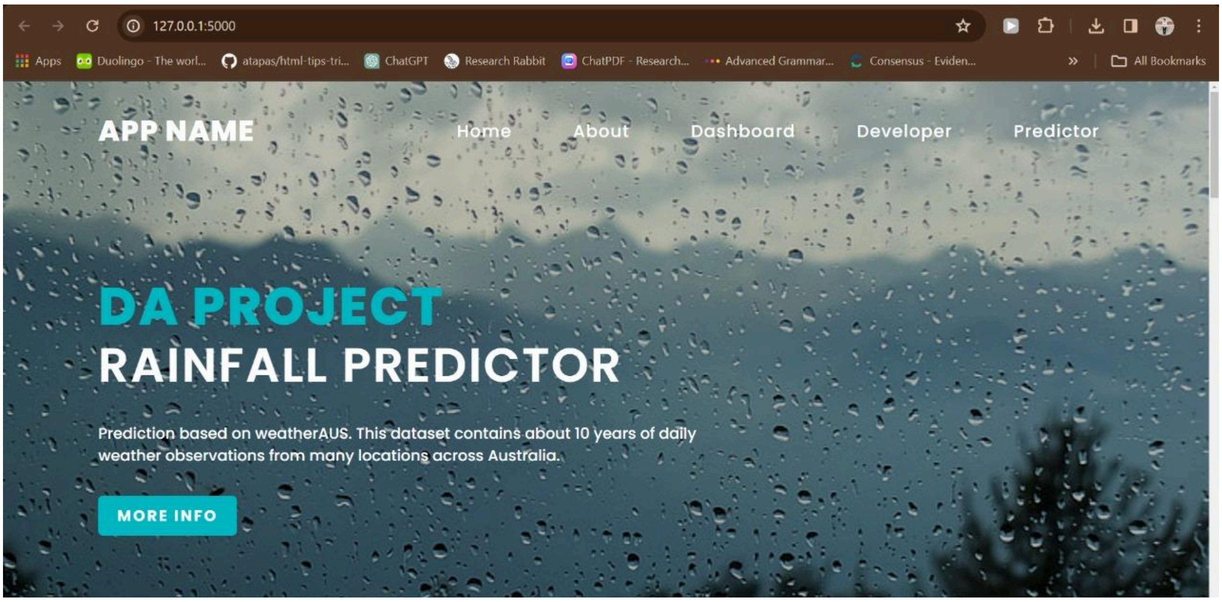


Figure 1.1: Home Page of Website

Chapter 2

Basic Concepts/ Literature Review

Previous research has shown promising results with regard to using machine learning techniques in improving agricultural practice and weather prediction. There have been studies from Australia that use Logistic Regression and Random Forest classifiers for improving rainfall forecasting. By doing this, stakeholders will be able to manage resources better and have risk reduction strategies in place.

Another study in India employed a machine learning approach based on Random Forest Regression for crop yield prediction. The model's accuracy was at 0.95 and is evident to be of high value to enable farmers to plan against possible surplus or shortages, and hence optimize production. Another study used the Weka toolkit to forecast varied weather conditions, using machine learning algorithms. Temperature, wind, and precipitation data were able to develop models that have high accuracy when predicting weather states like rain, sun, and snow. These findings can be translated into user-friendly dashboards, giving relevant information in decision-making, for instance in weather-sensitive industries. In general, such findings indicate the great power of machine learning to help change agricultural practice and weather prediction, eventually changing improved decision-making across sectors.

Chapter 3

Problem Statement / Requirement Specifications

3.1 Problem Statement

Accurate weather forecasting is thus critical for various domains, including agriculture, water resource management, and disaster preparedness. Traditional weather forecasting methods are very significant in informing the user. For instance, the weather forecasts inform farmers about when they should plant, how much they need to irrigate, and when they should harvest. When these forecasts are wrong, the crops are lost, the yields are lower, and economic losses are incurred.

Comparable practices apply to water resource managers who get only a general idea of weather forecasts to determine the best way to run their units and fight against water scarcity. Unexpected changes in the weather can result in the withdrawal of water supplies and create a ripple effect on various sectors. Disaster preparedness is also based on the ability to predict severe weather conditions, and these predictions have to be timely and precise so that communities can take proactive measures such as issuing evacuation orders, deploying emergency response teams, and securing infrastructure.

To summarize, traditional forecasting methods do not usually provide enough detail or accuracy that could be used for proper decision-making. Therefore, there is a strong imperative to develop more sophisticated and data-oriented approaches to weather prediction. This user-centric approach helps users to get clear and actionable weather forecasts tailored to their specific needs. This not only predicts the binary outcome of either sunny or rainy but the user-defined details in relation to meteorological information, such as temperature, rainfall, wind speed, humidity, sunshine, pressure, and cloud cover. Through such a strategy, users can learn from their historical weather data and also have the weather prediction generated by a machine learning model. For example, the forecast can contain the chances of rainfall and its intensity as light rain, downpour, or the probability distribution of different kinds of weather conditions (sunny, cloudy, or chance of showers). This very granular level of details will help the user in taking decisions in various weather-related contexts.

3.2 Requirements:

3.2.1 Data Acquisition:

Historical weather data encompassing various weather parameters like temperature, rainfall, wind speed, humidity, sunshine, pressure, and cloud cover is essential for training the machine learning model. This data should be reliable and representative of the target prediction regions. We have taken the dataset from Kaggle named “weatherAUS.csv”. This dataset contains about 10 years of daily weather observations from many locations across Australia.

3.2.2 Machine Learning Model:

A pre-trained Random Forest classifier model, proven effective in weather prediction, will be employed. Alternatively, other well-suited machine learning algorithms have been explored and compared for optimal performance. The model thus is trained and validated on the acquired historical weather data to establish a robust mapping between input features and weather conditions (sunny or rainy).

3.2.3 User Interface:

Flask Web Application: This will utilize a web form to gather user input, including location and various weather-related data points. The application will then process this information, make predictions using the trained model, and display the results (sunny or rainy) on a dedicated web page.

Streamlined Interface: This streamlined interface will collect user input for location and weather data points directly within the application. It should display the predicted weather conditions (sunny or rainy) in a clear and visually appealing format.

Chapter 4

Implementation

In this section, present the implementation done by you during the project development.

4.1 Methodology

Regression analysis:
Regression is a powerful statistical tool used to understand how one variable (the independent variable) affects another variable (the dependent variable). Imagine you're studying the relationship between house size (independent) and price (dependent). Regression helps you uncover this connection and predict how changes in house size might influence price.

Linear regression:
Linear Regression assumes a linear relationship between the dependent variable (house price) and the independent variables (factors like square footage, number of bedrooms, and location). It utilizes an equation to model this relationship:

$$\mathbf{y = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n}$$

y: Predicted house price (dependent variable)
$\mathbf{b_0}$: Bias or intercept term, representing the house price when all independent variables are zero (often not a realistic scenario, but helps account for a base value). b_i (i = 1 to n): Coefficients for each independent variable (x_i). These coefficients represent the impact of a unit change in the corresponding independent variable on the predicted house price.
$\mathbf{x_i}$ (i = 1 to n): Independent variables representing house features like square footage or number of bedrooms. The core idea is to find the values for b_0 and b_i that minimize the difference between the predicted house prices (y) and the actual house prices in the training data. This minimization is typically achieved using techniques like gradient descent.

4.2 Testing OR Verification Plan

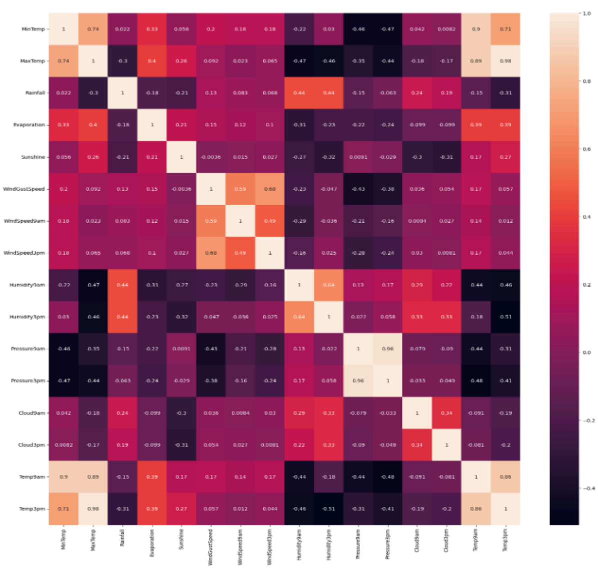
The application should be deployed onto a platform like a cloud server to make it accessible to users. The application is evaluated using metrics such as accuracy, precision, and recall, to assess how well the predicted weather conditions match the actual ones. Real-world data can be used to test whether the predictive power of the model is effective in getting actual weather patterns over time.

Metric	Rouge Type	English	Indonesian	Dutch	Vietnamese
R	ROUGE-1	0.500	0.231	0.378	0.500
	ROUGE-2	0.100	0.037	0.147	0.075
	ROUGE-L	0.267	0.192	0.333	0.367
P	ROUGE-1	0.197	0.059	0.293	0.214
	ROUGE-2	0.043	0.007	0.124	0.032
	ROUGE-L	0.105	0.050	0.259	0.157
F	ROUGE-1	0.283	0.094	0.330	0.300
	ROUGE-2	0.060	0.012	0.134	0.045
	ROUGE-L	0.151	0.079	0.291	0.220

Fig 4.2 Test Case Analysis

4.3 Result Analysis OR Screenshots:

The data used in the analysis is a survey report of Australia This dataset consists of about 10 years' worth of daily weather observations from many weather stations.



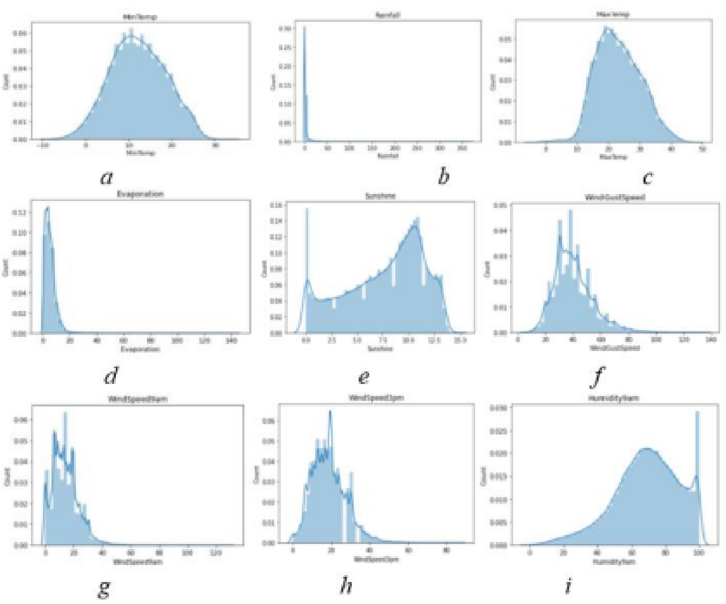


Fig 4.3.1 Confusion Matrix

Fig 4.3.2 Distribution Plots

Chapter 5

Standards Adopted

5.1 Design Standards

While developing the Rainy Day Prediction project, stringent principles and methodologies were followed to craft a robust Rainy Day Detection system, leveraging both Machine Learning (ML) and Data Analytics (DA) techniques, with Power BI employed for visualization. The design process adhered to a systematic methodology encompassing requirements elicitation, conceptualization, detailed design, implementation, and testing phases. The system architecture was meticulously structured, integrating ML models and DA algorithms within a scalable and modular framework.

Design choices were informed by established architectural patterns, ensuring flexibility, maintainability, and interoperability. Throughout the design process, emphasis was placed on optimizing system performance while accommodating evolving user needs and technological advancements. The integration of Power BI facilitated intuitive data visualization, enhancing the system's usability and providing actionable insights for stakeholders.

We adhered to rigorous design standards in developing ML and DA-based systems, ultimately leading to a reliable and effective Rainy Day Detection solution. By systematically following design methodologies and leveraging appropriate technologies such as Power BI, the resulting system is well-equipped to meet the challenges of detecting rainy conditions with accuracy and efficiency, while also providing a foundation for future enhancements and scalability.

5.2 Coding Standards

Strict attention to coding standards was essential during the Rainy Day Detection project's development in order to guarantee the codebase's scalability, readability, and maintainability. The project adhered to accepted coding standards such as uniform name standards, use of proper comments ,indentation patterns, and code structure guidelines to improve code readability and promote developer cooperation. Developers could quickly comprehend and traverse the codebase by adhering to a consistent coding style throughout the project, which decreased the possibility of mistakes and increased overall productivity.

In order to efficiently organize code contributions and monitor changes, the project also made use of version control systems like Git. This allowed smooth cooperation between teammates and also made bug fixing easier.

In addition, the project gave performance and efficiency first priority while developing its coding standards, optimizing data structures and algorithms to reduce computational overhead and enhance resource usage.

5.3 Testing Standards

Strict testing guidelines were used in the Rainy Day Detection project to guarantee the correctness and dependability of the system. The project used a thorough testing

methodology that included phases for system, integration, and unit testing. While integration tests confirmed how various modules interacted with one another, unit tests were used to confirm the operation of individual components.

To ensure fair comparisons and reproducibility, a systematic approach was adopted, encompassing data acquisition, preprocessing, and partitioning for training and testing. Baseline models were utilized for comparative analysis, providing a benchmark against which the proposed approach was evaluated.Throughout testing, stringent procedures were followed to maintain consistency and reliability, with challenges addressed transparently. Results were meticulously analyzed, shedding light on the system's performance and providing insights into its reliability and generalizability. By automating testing procedures and accelerating feedback loops, the project made use of continuous integration (CI) and continuous deployment (CD) techniques to make sure that any problems were quickly found and fixed throughout the development lifecycle.

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

In conclusion, this study delved into the application of machine learning techniques for predicting next-day rainfall in Australia using historical weather data. Through extensive experimentation with several models including SVM, KNN, Random Forest, and Support Vector Classifier, it was observed that outperformed the others with an impressive AUC score approaching 0.89 and a ROC score of 89. These results indicate the promising potential of machine learning in accurately forecasting rainfall patterns.

However, it's important to note that due to hardware limitations, the full extent of hyperparameter tuning was not feasible within the scope of this study. Thus, there remains a possibility of further enhancing the predictive performance of the models through fine-tuning of hyperparameters.

Despite this limitation, the findings of this research contribute significantly to the domain of weather prediction, particularly in the context of Australia. The ability to forecast rainfall with high accuracy can have profound implications for various sectors including agriculture, water resource management, and disaster preparedness.

Moving forward, future research endeavors could focus on overcoming the hardware constraints to perform comprehensive hyperparameter tuning, thereby potentially refining the predictive capabilities of the models. Additionally, exploring the incorporation of more advanced features and data sources could offer deeper insights into the complex dynamics of rainfall patterns.

In essence, this study underscores the efficacy of machine learning methodologies in addressing meteorological forecasting challenges and underscores the importance of continued exploration and refinement in this area for the betterment of society.

6.2 Future Scope

The future scope of the rain prediction model project could include several avenues for further exploration and improvement:

1. Hyperparameter Tuning: Since hardware limitations restricted hyperparameter tuning in the initial project, this could be a key area for future work. Implementing techniques such as grid search, random search, or Bayesian optimization could help fine-tune model parameters and potentially enhance predictive performance.

2. Feature Engineering: Exploring additional features or refining existing ones could lead to improved model accuracy. This might involve incorporating domain-specific knowledge or utilizing advanced techniques like time-series decomposition, lag features, or weather pattern analysis.
3. Ensemble Methods: Investigating ensemble learning techniques such as stacking, blending, or boosting could be beneficial. Combining predictions from multiple models might further boost predictive performance and robustness.
4. Advanced Models: Experimenting with more sophisticated machine learning models or deep learning architectures tailored for sequence data (e.g., recurrent neural networks, convolutional neural networks) could be worthwhile. These models might capture complex temporal dependencies inherent in weather data more effectively.
5. Spatial Analysis: Considering geographical factors and spatial relationships between weather stations could provide valuable insights. Spatial analysis techniques such as kriging or spatial interpolation could be used to enhance the model's understanding of regional weather patterns.
6. Real-Time Prediction: Developing a real-time rainfall prediction system could have practical applications. Integrating the model with live weather data feeds and deploying it in a cloud environment could enable timely and accurate rainfall forecasts.
7. Seasonal Variation Analysis: Investigating the model's performance across different seasons and climatic conditions could provide valuable insights. Understanding how the model generalizes to various weather scenarios and adapting its behavior accordingly could improve its reliability.
8. User Interface Development: Building a user-friendly interface or dashboard to visualize predictions and provide actionable insights to stakeholders (e.g., farmers, emergency responders, urban planners) could enhance the project's practical utility.
9. Collaborative Research: Collaborating with meteorologists, climatologists, or researchers in related fields could enrich the project with domain expertise and open up opportunities for interdisciplinary research.
10. Data Augmentation: Exploring techniques for data augmentation, such as synthetic data generation or incorporating external datasets, could help address data scarcity issues and improve model generalization.

By pursuing these avenues, the project could evolve into a more robust and comprehensive rainfall prediction system with broader applicability and impact.

References

https://ieeexplore.ieee.org/search/searchresult.jsp?queryText=rain%20prediction&highlight=true&returnFacets=ALL&returnType=SEARCH&matchPubs=true&sortType=newest

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10486681&tag=1

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10486733

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10486160

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10485045

INDIVIDUAL CONTRIBUTION REPORT

RAINY DAY PREDICTION MODEL

DEBARGHYA ROY

21051047

Abstract: This project outlines a rainfall prediction website employing machine learning (ML), data analytics (DA), and PowerBI Dashboard integration. Utilizing historical weather data, ML algorithms forecast whether it will rain tomorrow in a given location. DA techniques analyze meteorological parameters, while PowerBI Dashboard enhances data visualization. The website offers real-time updates and customizable settings, facilitating informed decision-making across sectors reliant on weather forecasts.

Individual contribution and findings: I was responsible for data collection, preprocessing, and exploratory data analysis. I was involved in conducting feature engineering to extract relevant features from the dataset. Applied statistical techniques to analyze historical weather data and identify patterns related to rainy days. Selected appropriate machine learning algorithms and models for training and evaluation.

Individual contribution to project report preparation: Chapter 6: Conclusion and Finding out the future prospect of the project.

Full Signature of Supervisor: Full signature of the student:

Debarghya Roy

……………………………. ……………………………..

SREEJATA BANERJEE

2105754

Abstract: This project outlines a rainfall prediction website employing machine learning (ML), data analytics (DA), and PowerBI Dashboard integration. Utilizing historical weather data, ML algorithms forecast whether it will rain tomorrow in a given location. DA techniques analyze meteorological parameters, while PowerBI Dashboard enhances data visualization. The website offers real-time updates and customizable settings, facilitating informed decision-making across sectors reliant on weather forecasts.

Individual contribution and findings: I undertook the implementation process, encompassing layout design, data analysis, model training, and Power BI visualization. Specifically, I led the creation of a user-friendly interface, conducted thorough data analysis including cleaning and feature extraction, selected and fine-tuned machine learning algorithms for model training, and developed interactive dashboards using Power BI for intuitive data presentation.

Individual contribution to project report preparation: I was involved in providing contents for various sections in the report and error correction.

Full Signature of Supervisor: Full signature of the student:

Sreejata Bannerjee

……………………………   ……………………………..

SEMANTI DAS

2105149

Abstract: This project outlines a rainfall prediction website employing machine learning (ML), data analytics (DA), and PowerBI Dashboard integration. Utilizing historical weather data, ML algorithms forecast whether it will rain tomorrow in a given location. DA techniques analyze meteorological parameters, while PowerBI Dashboard enhances data visualization. The website offers real-time updates and customizable settings, facilitating informed decision-making across sectors reliant on weather forecasts.

Individual contribution and findings: My contributions to this project spanned several areas. I leveraged my research skills to analyze relevant prior research papers, ensuring a comprehensive understanding of the field. Furthermore, I actively participated in writing and formatting the report, showcasing my technical writing and communication abilities. My grasp of the code and core concepts like the implementation allowed me to provide valuable insights throughout the project.

Individual contribution to project report preparation: Chapter 4: Implementation of the project and various error correction and formating of the report.

Full Signature of Supervisor:	Full signature of the student:
	Semanti Das
……………………………	……………………………..

SACHI VERMA

21052917

Abstract: This project outlines a rainfall prediction website employing machine learning (ML), data analytics (DA), and PowerBI Dashboard integration. Utilizing historical weather data, ML algorithms forecast whether it will rain tomorrow in a given location. DA techniques analyze meteorological parameters, while PowerBI Dashboard enhances data visualization. The website offers real-time updates and customizable settings, facilitating informed decision-making across sectors reliant on weather forecasts.

Individual contribution and findings: Played a multifaceted role in this project. My research efforts involved analyzing academic papers to build a solid knowledge base. I also contributed to writing and formatting the report, demonstrating my understanding of the code and relevant concepts like the literature review and adopted standards. Additionally, learning with peers was a great opportunity which I used to enhance my soft skills. At the end of the project, compiling the report while keeping the plagiarism in check was also one of my biggest challenges that I overcame.

Individual contribution to project report preparation: Contributed to Literature review; basic concepts of implementation and formatting of documentation.

Full Signature of Supervisor:	Full signature of the student:
	Sachi Verma
…………………………….	……………………………..

SAYAN CHATTOPADHYAY

2105748

Abstract: This project outlines a rainfall prediction website employing machine learning (ML), data analytics (DA), and PowerBI Dashboard integration. Utilizing historical weather data, ML algorithms forecast whether it will rain tomorrow in a given location. DA techniques analyze meteorological parameters, while PowerBI Dashboard enhances data visualization. The website offers real-time updates and customizable settings, facilitating informed decision-making across sectors reliant on weather forecasts.

Individual contribution and findings: I spearheaded project planning activities, and participated in brainstorming sessions to outline objectives, milestones, and responsibilities.
I also contributed by reading some research papers to extract useful information and also participated in writing and formatting the report. Additionally, I collaborated with team members to allocate resources efficiently, ensuring alignment with our goals and facilitating a cohesive approach to propel the project forward.

Individual contribution to project report preparation: Chapter 1: Introduction, Chapter 5: Standards Adopted, Report layout correction and error corrections across the pages.

Full Signature of Supervisor: Full signature of the student:

Sayan Chattopadhyay

……………………………… ………………………………..

TURNITIN PLAGIARISM REPORT

Data Analytics Rainy Day (1).docx

ORIGINALITY REPORT

14%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.coursehero.com Internet	257 words — 8%
2	filedata.kiit.ac.in Internet	42 words — 1%
3	www.monstersjournal.net Internet	24 words — 1%
4	Faisal Baig, Luqman Ali, Muhammad Abrar Faiz, Haonan Chen, Mohsen Sherif. "How accurate are the machine learning models in improving monthly rainfall prediction in hyper arid environment?", Journal of Hydrology, 2024 Crossref	22 words — 1%
5	Pablo Ferri Borredà. "Deeep Continual Multimodal Multitask Modells for Out-of-Hospital Emergency Medical Call Incidents Triage Support in the Presence of Dataset Shifts", Universitat Politecnica de Valencia, 2024 Crossref Posted Content	15 words — < 1%
6	Abhay B. Upadhyay, Saurin R. Shah, Rajesh A. Thakkar. "Theoretical Assessment for Weather Nowcasting Using Deep Learning Methods", Archives of Computational Methods in Engineering, 2024 Crossref	11 words — < 1%