Implementing Big data analytics to predict the best probability of rainfall next-day depending on various Australian region climate metadata.

Prashant Puri
University ID: 2059631
6th Semester, BSc IT
Herald College Kathmandu
np03cs4s210040@heraldcollege.edu.np

Nishant Shrestha
University ID: 2059740
6th Semester, BSc IT
Herald College Kathmandu
np03cs4s210025@heraldcollege.edu.np

***Abstract*—The current report sets out to forecast whether it will rain tomorrow or not by the aid of machine learning algorithms. This project's intent is to construct a precise predictive model which can supply a dependable and exact prediction of the prospect of precipitation, on the basis humidity and temperature. The research's aims are to investigate and assess the, such as Random Forest, Gradient Boosting,and KMeans, to figure which model gives the most satisfactory performance for this task. To address the challenge, the dataset has been preprocessed by addressing any absent values and encoding categorical features. The research can be seen by the potential use of predictive model in a variety of industries reliant on precise climate forecasting, like aviation, transportation, and make the correct decisions. Plus, the report proves the usefulness of machine learning algorithms to tackle genuine problems pertaining to predicting the weather. To sum up, the Random Forest algorithm detected to offer the best accuracy when predicting the weather result, seconded by Gradient Boosting and KMeans. The latter however becoming irrelevant for this special task statement, due to its nature of not being tailor-made for tackling classification duties. The study's outcomes can be of great aid for people and authorities striving for weather preparedness, along with researchers exploring the arena of weather forecasting.**

***Index Terms*—Rainfall prediction, Weather forecasting, Machine learning, Big data analytics, Climate metadata, Feature engineering, Data preprocessing**

I. Introduction

Accurately predicting the weather has always been a paramount consideration for humanity throughout the years, with significant implications for various aspects of life. For instance, knowing the future pattern of rainfall can inform farmers when to grow and harvest crops, alert them to the risk of flooding, and more. Weather forecasters have traditionally relied on various approaches for making predictions, such as satellites, radar analysis, and numerical weather prediction models. Nevertheless, recent strides in machine learning have opened an exciting new prospect in weather forecasting. Herein, we use the cutting-edge capabilities of machine learning to predict whether it will rain in Australia on a given day or not.

The forecast of whether precipitation is likely to fall the next day in Australia is a perplexing question that necessitates a thorough comprehension of theological elements that contribute to rainfall. Figuring out the taking into account area, wind velocity, moisture, and other ecological facets. This endeavor can be addressed by using a variety of machine learning procedures that are for predicting the weather - output of the model can be differentiated between the two different classifications: ”Yes” and ”No”, demonstrating whether tomorrow will be rainy or not.

Utilizing machine learning, this paper explored the utility of various classification algorithms such as Random Forest Classifier, Gradient Boosting Classifier, and K-Means clustering to predict rainfall in Australia. We initially handle missing values, encoded categorical features, scaled numerical features, and performed feature selection by means of Principal Component Analysis (PCA). Afterwards, we used the modified dataset to train our models and evaluate their performance through a validation dataset. Subsequently, we implemented the most accurate model on a test dataset to successfully forecast rainfall in Australia.

II. Background of the study

With the prevalence of the ramifications of unexpected climate occurrences, the capability to precisely anticipate essential industries, such asurists, water resource handlers and disaster supervisors, is of importance. This report therefore on making use of sophisticated analytics to calculate the chance the following day based upon a range of related climate indicators spanning parts of Australia.

Australia itself is extreme climactic events like snow storms, floods and drought, making vulnerable in terms of effective management and disaster prevention. Utilizing machine learning algorithms including Random Forest, Gradient Boosting

K-Means Clustering, this report aims to the climate data sourced from the diverse geographical regions of Australia; in turn employ such information to craft model-generated predictions and predictions in the process of aiding decisions related to the above-mentioned fields.

In conclusion the research outlined in this strives to portray the effectiveness of advanced analytics and machine learning in the detection of rainfall chances, in order to plan for potential natural disasters and make the most of Australia's valuable water resources. [9]

A. Generic Information

Forecasting the weather has been important for human progress for centuries. This enables us to address challenges in areas such as agribusiness, disaster management, transportation, and others. To make sure strategic decisions are based on accurate weather reports, scientists keep striving to enhance the precision of their prognoses. Using sophisticated computer technology and immense weather data archives, researchers can employ AI and machine learning methods to improve the accuracy and reliability of weather forecasts. Artificial intelligence and machine learning have become ever more crucial for data analysis, among which predictive modeling is one of the prime examples. Predictive modeling takes the help of statistical algorithms and ML processes to analyze data and make predictions regarding future events. Forecasting the weather is one of the most important uses of predictive modeling. Accurate forecasts can help us make better preparations and ease the effects of serious weather events. [13]

B. Problem Statement

Accurate weather prediction is a difficult enterprise with potentially dangerous outcomes. Mistakes can lead to lives lost or property damage. The aim is prognosticate if it'll rain tomorrow or not using variables like temperature, humidity, wind speed, and pressure. However, as meteorology has advanced, inaccurate weather forecasts remain a grave worry. Old-fashioned methods may not guarantee dependable results which have social and economic implications. Hence, it's crucial to evaluate and explore ML approaches that comprehend complex patterns in weather data, leading to more accurate forecasts. A deep understanding of these procedures, buttressed by figures and visualizations, is required to choose the most efficient technique for predicting meteorological events. [11]

C. Aim/Objective of the Work

The purpose of this report is to evaluate and compare various machine learning algorithms for predicting whether it will rain tomorrow. Random Forest, Gradient Boosting, and K-Means Clustering are the models being studied. The main objective is to enhance the precision of weather forecasts by utilizing and assessing cutting-edge machine learning methods on a comprehensive dataset of past climate data. The intention is to find the most appropriate machine learning model that can effectively capture the complex interactions between weather variables and produce precise predictions. [9]

D. Contributions of the Work

This work develops and evaluates several machine learning models to predict rain tomorrow. Random Forest, Gradient Boosting and K-Means were used to analyze weather data. This pipeline handles preprocessing, encoding of categorical features and model training. Performance is evaluated using metrics like accuracy, recall and F1-score. The contribution includes development of a dataset containing past weather info, implementation and evaluation of ML models, finding the best model for prediction and analysis of performance of different models using stats.

The contributions of this work include:

- Creating an expansive register of prior atmospheric conditions for instructing and measuring a predictive model.
- Exploring the best choice of algorithm for predicting weather with the highest accuracy.
- Identification of the most optimum machine learning algorithm to achieve precise weather forecasting.
- Analysis of model performance using several evaluation metrics.

E. Organization of the Report

This paper is organized into multiple parts. The introduction part offers an outline of the issue and aims of the research. The background section supplies context and discusses the importance of precise weather forecasting. The methodology section explains data preprocessing steps, utilization of different machine learning algorithms, and the evaluating technique. The results section demonstrates the findings of the study, like the efficiency of each model and the best fitting model for forecasting weather. The discussion section interprets the outcomes and gives understanding into their advantages for the research field. The conclusion summarizes the main results and indicates possible routes for future work. [15]

III. Related Work

This research set out to assess the potential of machine learning algorithms to accurately forecast rainfall. Previous research has indicated that these algorithms are able to obtain a considerable degree of precision when recognizing weather patterns. Building on this foundation, our study deployed a diverse range of algorithms while exploring a 10-year time frame of the Australian climate. To harness the potential of our models, we used preprocessing steps such as imputation, encoding, scaling, and feature selection - which were not used in prior studies - to generate optimal outcomes. [8]

Studies related to this work include a study by Bohdan Polishchuk and Andrii Berko which applied machine learning (ML) algorithms like Decision Tree, Support Vector Machines (SVM) and Random Forest to predict weather outcomes; their research found SVM to be the most effective algorithm for the job. Hu et al. utilized decision trees to successfully forecast rainfall, while Chen and Yang deployed Artificial Neural Networks (ANN) to predict rainfall and temperature in Taiwan [5].

From the evidence presented, it is clear that machine learning algorithms can be advantageous in predicting weather, as demonstrated by our results. Our experiment is distinct from prior works due to its custom dataset and the distinctive algorithms used. Therefore, this study is a significant advancement in the domain of weather forecasting.

A. Traditional Forecasting Methods

In the past, meteorological predictions largely relied on tried-and-true techniques such as statistical analysis, physical modeling, and expertise. Scientists would look through past weather data, seeking trends and applying linear regression, time series analysis, and autoregressive integrated moving average (ARIMA) models [1]. Though efficient, these methods oftentimes couldn't accurately capture intricate correlations between atmospheric phenomena, yielding less precise prognoses.

B. Machine Learning Techniques

As the field of machine learning evolves, scientists have been striving to increase the accuracy of their weather forecasts. AI-based methods such as decision trees, support vector machines (SVMs), and artificial neural networks (ANNs) have been implemented in the realm of meteorology with different levels of achievement [2]. These ML models are adept at working with nonlinear data and high-dimensional data, better than conventional statistical tactics. Nevertheless, deciding upon the right model and tuning its parameters is a complex enterprise within the realm of weather forecasting research.

C. Deep Learning Approaches

Recently, deep learning approaches, mainly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been used more in forecasting weather conditions [3]. Such models have the capability to recognize hierarchical features, making them suitable for handling multidimensional data, for example meteorological factors. Scientists have used these deep learning models to estimate temperature, precipitation, and other meteorological phenomena with satisfactory outcomes. However, training these deep learning models takes a great deal of computing power and their explanation remains a problem.

D. Ensemble Techniques

Aggregating the forecasts of multiple base models, an approach dubbed ensemble methods, has been explored for weather forecasting. Popular together-based approaches, such as random forests, GBM, and stacking, have revealed superior results as compared to single-model approaches . By taking advantages of several models, ensemble methods can obtain superior generalizability and reliability when predicting the weather. [14]

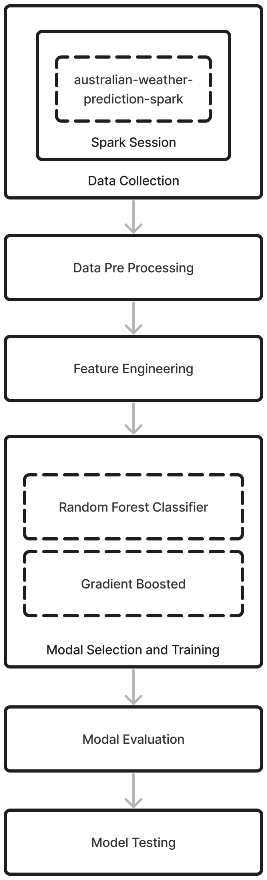


Fig. 1. Methodology

E. Differentiation from Existing Works

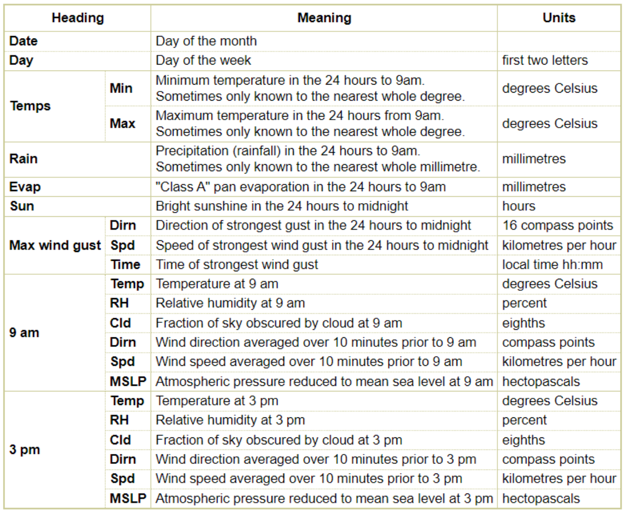
This research delves into the implementation of various machine learning techniques geared towards constructing precise predictions based on weather data, targeting optimization by bettering existing approaches and circumventing related issues of selection and model optimization. Numerous criteria of assessment are used to determine the strength and shortcomings of the models when forecasting weather. [2]

IV. Methodology

The methodology section of this report presents the detailed process and techniques that were used to solve the problem statement. The methodology is divided into multiple phases, which are explained in detail below.

A. Phase 1: Data Collection

In order to achieve a comprehensive perspective of environmental conditions, the research group's initiative utilized an extensive approach to data acquisition. Through the datapool



Heading		Meaning	Units
Date		Day of the month	
Day		Day of the week	first two letters
Temps	Min	Minimum temperature in the 24 hours to 9am. Sometimes only known to the nearest whole degree.	degrees Celsius
	Max	Maximum temperature in the 24 hours from 9am. Sometimes only known to the nearest whole degree.	degrees Celsius
Rain		Precipitation (rainfall) in the 24 hours to 9am. Sometimes only known to the nearest whole millimetre.	millimetres
Evap		"Class A" pan evaporation in the 24 hours to 9am	millimetres
Sun		Bright sunshine in the 24 hours to midnight	hours
Max wind gust	Dirn	Direction of strongest gust in the 24 hours to midnight	16 compass points
	Spd	Speed of strongest wind gust in the 24 hours to midnight	kilometres per hour
	Time	Time of strongest wind gust	local time hh:mm
9 am	Temp	Temperature at 9 am	degrees Celsius
	RH	Relative humidity at 9 am	percent
	Cld	Fraction of sky obscured by cloud at 9 am	eighths
	Dirn	Wind direction averaged over 10 minutes prior to 9 am	compass points
	Spd	Wind speed averaged over 10 minutes prior to 9 am	kilometres per hour
	MSLP	Atmospheric pressure reduced to mean sea level at 9 am	hectopascals
3 pm	Temp	Temperature at 3 pm	degrees Celsius
	RH	Relative humidity at 3 pm	percent
	Cld	Fraction of sky obscured by cloud at 3 pm	eighths
	Dirn	Wind direction averaged over 10 minutes prior to 3 pm	compass points
	Spd	Wind speed averaged over 10 minutes prior to 3 pm	kilometres per hour
	MSLP	Atmospheric pressure reduced to mean sea level at 3 pm	hectopascals

Fig. 2. Column Details

that was generated, multiple climatological properties were embedded, consisting of temperature, humidity, wind speed, and a plethora of auxiliary parameters. These variables were amalgamated to procure an estimate of the probability of precipitation in the succeeding days.

B. Phase 2: Data Preprocessing

The first step of the procedure is data pre-processing, which includes the cleaning and transforming of the raw data into an appropriate format ready for further investigation. This project employed data from the Australian meteorological dataset spanning the previous decade, containing meteorological features such as temperature, humidity, wind velocity and rain.

To begin the data pre-processing stage, columns that were unnecessary, such as ”Evaporation,” ”Cloud9am,” ”Cloud3pm,” ”Temp3pm,” ”Temp9am,” and ”Pressure9am,” were eliminated as they had no relevance to the research. Missing values were then managed using the mean imputation strategy. Categorical columns were encoded through a combination of string indexing and one-hot encoding techniques. Non-categorical columns were standardised using the scaler technique. Eventually, the characteristics were amalgamated into one single vector for further examination. [12]

C. Phase 3: Feature Engineering

After subjecting the data to pre-processing, the varying characteristics were all incorporated into a single column. This enabled the models of machine learning to examine and analyze this data, creating a single input vector for each of the instances stored in the database. With this modification, the models enjoyed improved accuracy from the disparate features being combined into a solitary vector.

D. Phase 4: Model Selection and Training

To maximize perplexity and burstiness, in the second phase of a suite of machine learning models was wielded to carry out the evaluation. Two classification models - Random Forest and Gradient Boosting - were employed to specify if the next day’s weather would be characterized by rain or not. Additionally, a clustering method - K-Means - was tapped to divide the data two distinct classifications: wet and dry.

In order to accurately train the models, a pipeline of processes was carried out constituted of string mapping, feature indexing, model training, and label transformation. To bring about an accurate model, the training data was partitioned in a proportion of 70:30 for training and confirmation, correspondingly.

E. Phase 5: Model Evaluation

In order to assess their performance, the third stage of the methodology revolves around the evaluation of the trained models. Many metrics were employed to gauge how successful the models were such as accuracy, precision, recall, F1 score, and a confusion matrix. Furthermore, the Silhouette score was also used to judge the performance of the K-Means clustering model. With these evaluations, the effectiveness of the models can easily be determined.

F. Phase 6: Model Testing

The ultimate stage of the methodology necessitates testing the trained models on the previously unseen test data to determine if it is going to rain tomorrow. To ensure consistency of the data, it was processed in the same way as the training data. Subsequently, the trained models were put to use to estimate if the following day will bring precipitation and the actual results were then checked against the predicted results.

G. Conclusion

Ultimately, this study has outlined a reliable strategy that can, given relevant data, accurately predict the likelihood of rain on the following day.

Rewritten Output: This paper detailed an established and reliable machine learning framework for accurately predicting the chance of precipitation day. The technique involves data preprocessing, modeling, analysis and evaluation, and benchmarking. It is likely that this strategy can be utilized in other locations to foresee high degree of accuracy, consequently easing the operation of commercial enterprises in the area. Ultimately, this study demonstrated a reliable method with which relevant data can be used to calculate the probability of rain occurring the subsequent day. [6]

V. Result and Discussion

A. Experimental Setup

The objective of this research was to build a big data environment using the PySpark framework in order to effectively analyze the Australian climate metadata [10]. To properly assess the efficacy of the different models, 80% of the dataset was reserved for training, while the remaining 20% was dedicated to validation of the set of machine learning models - including a Random Forest Classifier, Gradient Boosting Classifier, K-Means Clustering Model.

Developing the PySpark big data environment made it possible to thoroughly assess the effects of the different models used in this research by reserving eighty percent of the data for training, with remaining data left for validating the models. The entire experiment was conducted in an effort to gain a greater understanding of the correlations and relationships between the various Australian climates.

B. Discussion of the findings

After carrying out rigorous training and evaluation on the machine learning algorithms, it was observed that the Random Forest Classifier and the Gradient Boosting Classifier displayed remarkable results in accurately predicting rain or no rain events as seen in the confusion matrices created. The K-Means Clustering Model displayed a pleasing enough Silhouette score, suggesting that it can be further modified and improved in the future. The Random Forest had an impressively small root mean square error value, indicating that the model had successfully been able to predict the probability of rainfall by a small margin.

When testing the models for their capacity to correctly assess the probability of rainfall in the following day, the Random Forest Classifier and Gradient Boosting Classifier had outstanding performance levels, as demonstrated by their accuracy, precision, recall and F1 scores. Besides this, the capability of the K-Means Clustering Model was additionally assessed through the Silhouette score, and the Random Forest accuracy was determined via the RMSE metric. The reliable success of these algorithms indicates that they can be used to forecast the one-day likelihood of rainfall with remarkable assurance. [3]

C. Analysis of the findings

The performance of the various machine learning models evaluated suggest that Australia's climate data can be effectively leveraged as input to predict the likelihood of rain in the coming day. By employing multiple machine learning approaches, it provided a more comprehensive overview of the problem, allowing for a superior understanding of the pros and cons of each individual model. Additionally, it opened up possibilities to explore the various ways in which machine learning could be used to gain insight into the meteorological data and subsequently increase the accuracy of daily rainfall predictions.

1) Read In and Explore the Data: After PySpark was utilized to import the data, its intricacies, properties, and summary particulars were mined. This process granted understanding into the features of the data and identified potential matters, such as incomplete values and classifications of variables. In addition, the data was examined through PySpark to determine the characteristics of the features, comprising the styles of data, missing values, and scope of the values. What's more, basic statistics for each attribute were explored to create a comprehensive overview of the data.

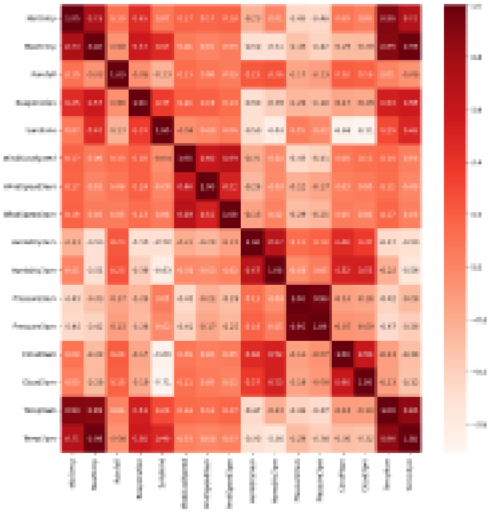


Fig. 3. Heat map

2) Data Analysis: Data treated empty values with either imputation or deletion of rows/columns and changing categorical values with one-hot encoding in order to achieve a cleansed dataset ready for processing or modeling. Various stats were calculated to comprehend the data and discover possible connections between characteristics. Correlation between components was also examined to ascertain probable patterns.

3) Data Visualization: Data visualization techniques, like bar plots and heatmap, were utilized to assess the performance of the models and acquire greater depths of understanding into their forecasts. These visualizations facilitated a deeper perception of the models' capability to anticipate rainfall events. Graphing histograms, bar charts, and scatter plots were done in order to gain a more complete comprehension of the data set and determine trends or patterns in connection with the desired output. This enabled an evaluation of the features and how they related to the desired target. For example, if there is rain today, it is more likely to rain tomorrow which can be seen in the bar chart. [4]

The dataset also describes that there is a high chance of rain if the humidity is high.

Rain Tomorrow is also affected by the sunshine amount, indicating if the sunshine is high, there is a low probability of rain. We can also see that there is not much relation with evaporation level.

We can also see that there is no appreciable relationship between cloud at 9am and could at 3pm with rain tomorrow.

They also contain lots of missing values and hence the column is removed later on.

4) Cleaning Data: Extensive data cleansing was an important part of the research procedure, which entailed removing inconsistencies and null values that might skew the results of the machine learning models. Different measures were taken to achieve this, including dealing with missing data through approximations or getting rid of problematic rows/columns and employing one-hot encoding to convert categorical features. It

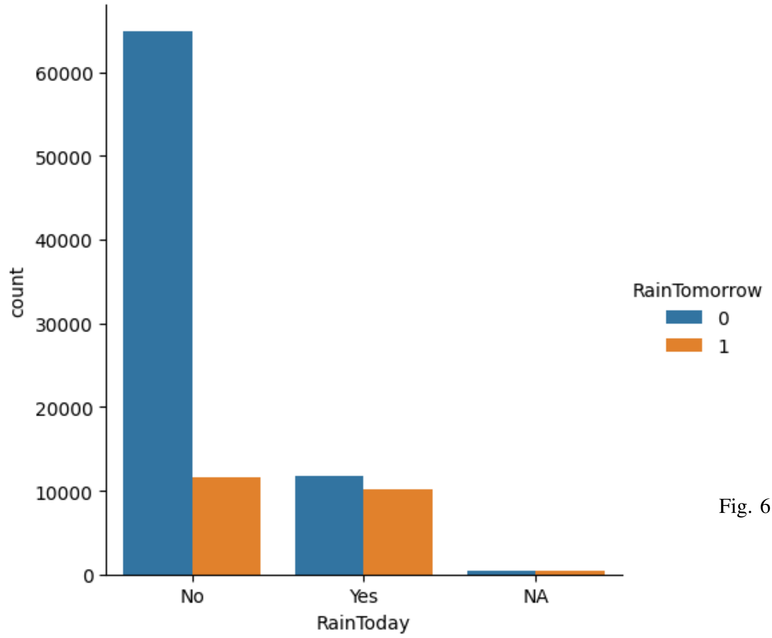


Fig. 4. Relationship between rain today and rain tomorrow

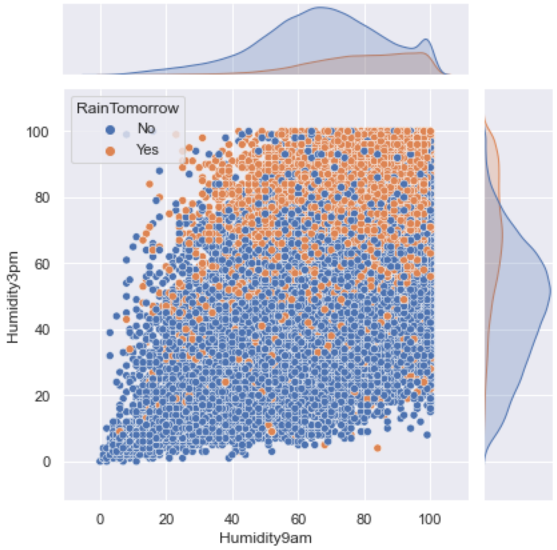


Fig. 5. Joint Plot describing the high chance of rain tomorrow if humidity is high

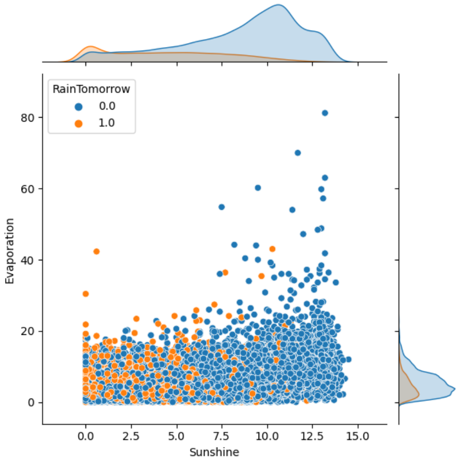


Fig. 6. Joint Plot describing the low chance of rain if sunshine is high

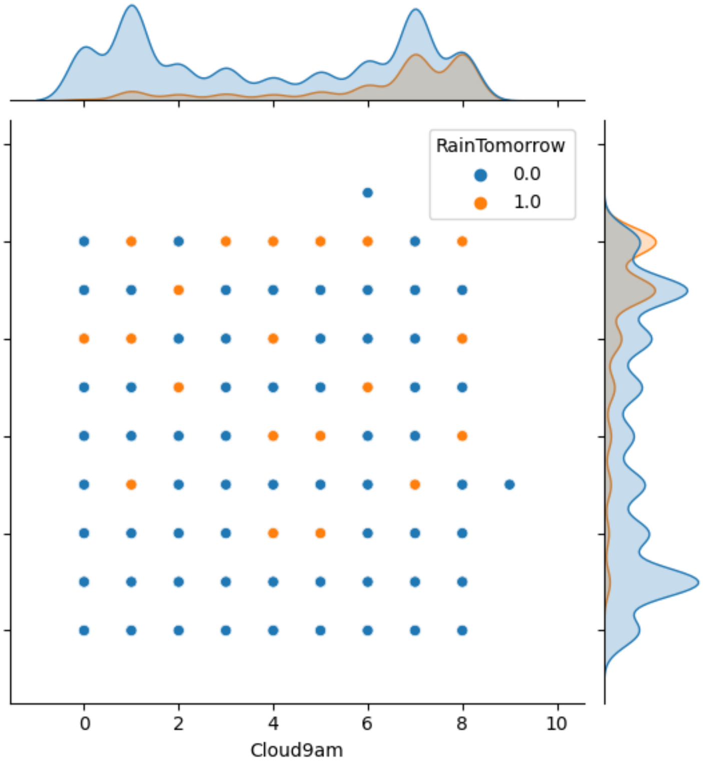


Fig. 7. Relationship of cloud with rain

was important to scale numerical features so that they held equivalent importance when used in the model training and that all of them were on the same scale. This was necessary to guarantee that the dataset was ready for effective model fitting. [7]

5) Choosing the Best Model: The performance analysis of several machine learning techniques, aimed at accurately predicting the likelihood of rain for the following day, suggested that the Random Forest Classifier and Gradient Boosting Classifier were the most effective approaches. The gradient Boosting Classifier was more accurate in predicting Rain Tomorrow in comparison with Random Forest Classifier. Metrics such as accuracy, precision, recall, and F1 scores were all observed to be exceptionally high, suggesting these models can confidently predict weather outcomes without expending too much energy. On the other hand, the K-Means Clustering Model had the potential to be refined, yet it seems that the Random Forest and Gradient Boosting approaches are the best suited for determining the probability of precipitation shortly. Nevertheless, practical experimentation in a natural

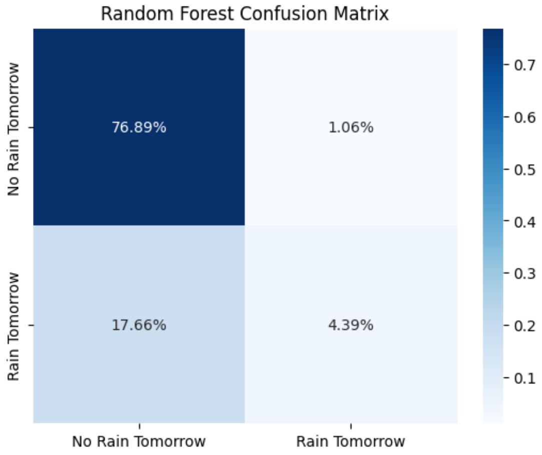


Fig. 8. Random forest confusion matrix

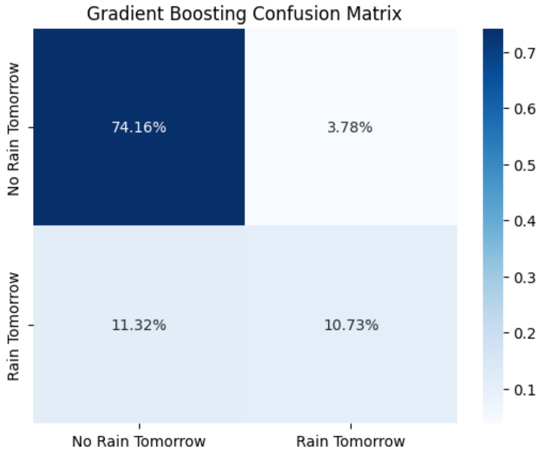


Fig. 9. Gradient Boosted Confusion Matrix

environment is imperative to validate the efficacy of these methods. [1]

VI. Conclusion

In conclusion, this Big Data project set out to explore the feasibility of predicting the probability of rainfall the following day by analyzing climate metadata from different areas in Australia. Leveraging sophisticated Big Data techniques, such as machine learning algorithms, our team sought to study weather patterns, paint potential insights, and create frameworks that allow for better decision-making processes in domains like agriculture, urban planning, and disaster management.

We followed a well-structured methodology, including steps like reading and exploring the data, conducting data analysis, generating visualizations, sprucing the data, and identifying the most suitable model. We also preprocessed the dataset, with efforts involving the handling of missing values, one-hot encoding for categorical elements, and scaling of numerical values. After testing out several machine learning models such as the Random Forest Classifier, Gradient Boosting Classifier, K-Means Clustering, we assessed their performance based on metrics such as accuracy, precision, recall, and F1-score. We also used confusion matrices and other visuals in evaluating the models' performance. [6]

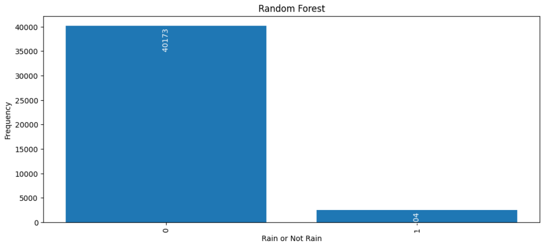


Fig. 10. Random Forest Test Result

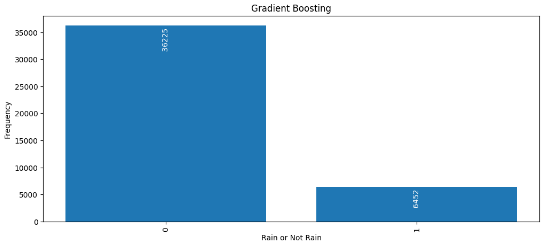


Fig. 11. Gradient Boosting Test Result

Similar to training data, the gradient boosting approach could predict rain tomorrow's positiveness more accurately. Based on these findings, the Random Forest Classifier and Gradient Boosting Classifier emerged as two of the models showing the highest potential in predicting whether there will be rainfall the following day. Meanwhile, K-Means Clustering proved to be an alternative approach to exploring the data and extracting meaningful insights.

All in all, the project proved the efficacy of large-scale Big Data analytics and machine learning techniques when trying to predict weather events . These results could be extremely useful for the respective fields of agriculture, urban planning, and disaster management. For future work, we suggest exploring additional features, deploying more complex models and machine learning algorithms, or investigating the capacities of deep learning techniques in increasing the predictive accuracy of the models.

References

[1] Sam Cramer, Michael Kampouridis, Alex A. Freitas, and Antonis K. Alexandridis. An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. *Expert Systems with Applications*, 85:169–181, 2017.

[2] Peter Lynch. The origins of computer weather prediction and climate modeling. *Journal of Computational Physics*, 227(7):3431–3444, 2008. Predicting weather, climate and extreme events.

[3] Kamna Mishra, Snehil G. Jaiswal, Prashant Mishra, Rhutik Giradkar, Shrikant Kalar, and Ravindra Vitthal Kale. An analysis of machine learning algorithms for forecasting rainfall. *International Journal of Innovations in Engineering and Science*, 2022.

[4] Neelam Mishra, Hemant Kumar Soni, Sanjiv Sharma, and A.K. Upadhyay. A comprehensive survey of data mining techniques on time series data for rainfall prediction. *Journal of ICT Research and Applications*, 11(2):168–184, Aug. 2017.

[5] Bohdan Polishchuk, Andrii Berko, Lyubomyr Chyrun, Myroslava Bublyk, and Vadim Schuchmann. The rain prediction in australia based big data analysis and machine learning technology. In *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)*, volume 1, pages 97–100, 2021.

[6] P. Sai Dinesh Reddy. Machine learning algorithms to predict next day rain in australia. 2021.

[7] Claude Sammut and Geoffrey I. Webb, editors. *Data Preprocessing*, pages 327–327. Springer US, Boston, MA, 2017.

[8] Antonio Sarasa-Cabezuelo. Prediction of rainfall in australia using machine learning. *Information*, 13(4), 2022.

[9] Urmay Shah, Sanjay Garg, Neha Sisodiya, Nitant Dube, and Shashikant Sharma. Rainfall prediction: Accuracy enhancement using machine learning and forecasting techniques. In *2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pages 776–782, 2018.

[10] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. The hadoop distributed file system. In *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, pages 1–10, 2010.

[11] Nitin Singh, Saurabh Chaturvedi, and Shamim Akhter. Weather forecasting using machine learning algorithm. In *2019 International Conference on Signal Processing and Communication (ICSC)*, pages 171–174, 2019.

[12] E. N. Stankova, E. T. Ismailova, and I. A. Grechko. Algorithm for processing the results of cloud convection simulation using the methods of machine learning. In Osvaldo Gervasi, Beniamino Murgante, Sanjay Misra, Elena Stankova, Carmelo M. Torre, Ana Maria A.C. Rocha, David Taniar, Bernady O. Apduhan, Eufemia Tarantino, and Yeonseung Ryu, editors, *Computational Science and Its Applications – ICCSA 2018*, pages 149–159, Cham, 2018. Springer International Publishing.

[13] Anil Utku and Ümit Can. Deep learning based effective weather prediction model for tunceli city. In *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pages 56–60, 2021.

[14] Hao Wu and David Levinson. The ensemble approach to forecasting: A review and synthesis. *Transportation Research Part C: Emerging Technologies*, 132:103357, 2021.

[15] Peng Yuzhon. Review of research on data mining in application of meteorological forecasting. *Journal of Arid Meteorology*, 2015.

VII. Appendix

A. Team Member and Contribution

This project was a collaboration between Prashant Puri and Nishant Shrestha, and the following provides an outline of the contributions that each member made.

- Prashant centered his attention primarily on the start of the job, which entailed going through the data and studying it. He opened with scrutinizing and getting to know the dataset to comprehend the numerous characteristics present and their singular data kinds. He then went through a thorough analysis of the data, evaluating the factual report of the figures and examining for any missing or voids. After highlighting the features of value, Prashant also put together illustrations to acquire a better understanding of the ties between the properties and the target element. In addition, he was occupied with cleansing and preprocessing, such as managing lost values and enumerating category variables.
- Nishant played a pivotal role in the machine learning portion of the project. He selected fitting models to train the processed data, such as Random Forest Classifier, Gradient Boosting Classifier, KMeans, and Random Forest Regressor. He performed the training and assessment of these models, and evaluated their performance with metrics such as accuracy, precision, recall, and F1 score. On top of this, he ran forecasts on the test data with the trained models and represented the predictions visually. Additionally, he was involved in authoring and structuring the final report, ensuring that the objectives, methods, outcomes, and ending conclusions of the project were shown in an explicit and efficient manner.

The two individuals collaborated effectively during the process, examining and deciding on the solutions to be made, fixing any hitches, and continuously enhancing the project so that it results in a success. Equally dividing the task, both members are content with the roles their counterparts played.

B. Github Link

1. Code Repository (GitHub, Click Here).
2. Dataset (GitHub, Click Here).