Project Report for Probabilistic Machine Learning
Benchmarking classification models on weather data

Konstantin Abe
Department of Computer Science
University Leipzig
ki31osid@studserv.uni-leipzig.de

Abstract

This written report for the module "Probabilistic Machine Learning" benchmarks three different classification models on weather data from Australia. The three models, namely, Logistic Regression, XGBoost, and Bayesian Logistic Regression are used to predicting next-day rainfall in Australia based on historical weather observations. The prediction are compared based on commom metrics and validated by a statistical test.

1 Introduction

Predicting weather patterns, especially rainfall, is crucial for various sectors, including agriculture, transportation, and disaster management. Accurate predictions can help in resource allocation, planning, and mitigating the impacts of adverse weather conditions. This project aims to leverage probabilistic machine learning techniques to build a model that can predict whether it will rain tomorrow based on historical weather data.

The prediction of rain is one of the most recognized and visible predictions task in the broader public, largely due to the presence in modern media, TV and weather apps. Almost any news show wouldn't be complete without a weather forecast, and the public is used to seeing these predictions. But also academia is interested in leveraging machine learning techniques to improve the accuracy of weather predictions. Given the growing complexity of the climate system and the limitations of traditional forecasting methods, artificial intelligence has emerged as a transformative tool. These methods are increasingly used not only to improve predictive accuracy, but also to enhance the detection, attribution, and communication of extreme events. Their ability to integrate heterogeneous data sources and uncover complex spatio-temporal patterns makes them especially suited to this task. As highlighted in Camps-Valls et al. [2025], developing reliable and explainable machine learning models is a critical step toward strengthening early warning systems and building trust in risk communication and decision-making processes.

2 Research Question and Hypotheses

Due to the complexity of the task, the project will only focus on comparing different methods to each other in the context of the given dataset. The goal is to find the best performing model, even though state-of-the-art methods are already more advanced at this point in time.

Deriving from that, the project will focus on the following research questions: What is the best performing model for predicting rain tomorrow based on the given dataset?

To follow a scientific approach, the project will be structured along the following hypotheses:

- Hypothesis 1: The XGBoost model will outperform a Logistic Regression model in predicting whether it will rain tomorrow based on the historical weather data.

- Hypothesis 2: The Logistic Regression model will outperform both the Bayesian Logistic Regression in terms of prediction accuracy.
- Hypothesis 3: XGBoost model will outperform both the Bayesian Logistic Regression in terms of prediction accuracy.

3 Data Description

The dataset used in this analysis originates from historical weather observations collected across various locations in Australia. It comprises a total of 145,460 records and 23 variables, including both continuous measurements and categorical features. The data is typically used for modeling weather-related outcomes, especially for predicting whether it will rain the following day. The target variable for predictive modeling is RainTomorrow, which indicates whether it rained the next day. This makes the dataset particularly suitable for binary classification tasks in the context of weather forecasting. The variables available in the dataset are e.g. *MaxTemp*, the maximum temperature recorded that day or *Sunshine*, which refers to the hours of sunshine on this day. All the variables and their respective meanings can be seen in the appendix 1.

4 Data Cleaning and Exploration

4.1 Data Cleaning

To enhance the performance of the machine learning models and ensure the extraction of reliable insights, the dataset undergoes a structured cleaning and transformation process. Features with more than 30% missing values are removed, and the variable Date is excluded due to its lack of relevance to the prediction task. After these steps, the dataset contains 18 columns. Observations with missing values in the target variable *RainTomorrow* are then discarded. To enable the use of *RainTomorrow* in the modeling stage, its categorical values ("No", "Yes") are mapped to a binary format (0, 1). Furthermore, all remaining categorical features are encoded into numeric representations to facilitate their use in machine learning algorithms. Finally, the data is standardized to ensure training stability and improve model convergence.

4.2 Data Exploration

In the column "Location" we can find the location of the weather station, which is used to collect the data. The dataset contains 49 different locations, which are shown in the following figure. Australia has six different climate zones, which not only differ by temperature but also by humidity and wind strengths, according to the Australian Bureau of Meteorology [2023].

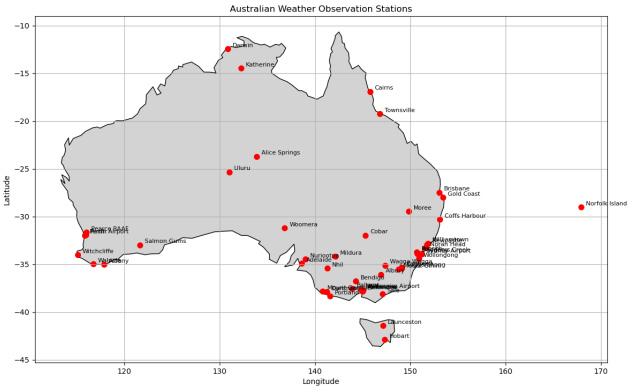

Figure 1: Spatial Distribution of the weather stations

The dates range from 01.11.2007 to 25.06.2017, which are almost 10 years of data. For the 145,460 recorded observations during the years, rainfall was registered on 31,880 days, corresponding to approximately 22% of all records. This indicates a marked class imbalance, as illustrated on the

left side of Figure 2. The right side of the figure presents a histogram, overlaid with a density curve, illustrating the distribution of daily rainfall amounts across the dataset. The distribution exhibits a pronounced peak at low precipitation levels (0–2 mm), with frequencies exceeding 120,000 occurrences. As rainfall amounts increase, the frequency declines sharply, indicating that substantial precipitation events are comparatively rare. Beyond approximately 10 mm of daily rainfall, such events become infrequent, while measurements exceeding 30 mm represent extreme outliers within the dataset. An examination of the *Rainfall* variable reveals that many of the outliers can be associated with flood events in Australia. The highest recorded rainfall occurred in Coffs Harbour, New South Wales, on 17 February 2010 as reported in ABC News [2009], with a total of 371.0 mm. Another notable peak corresponds to the major flooding events in Queensland during 2010 and 2011 as described in the report of the Australian Institute for Disaster Resilience (AIDR) [2011].

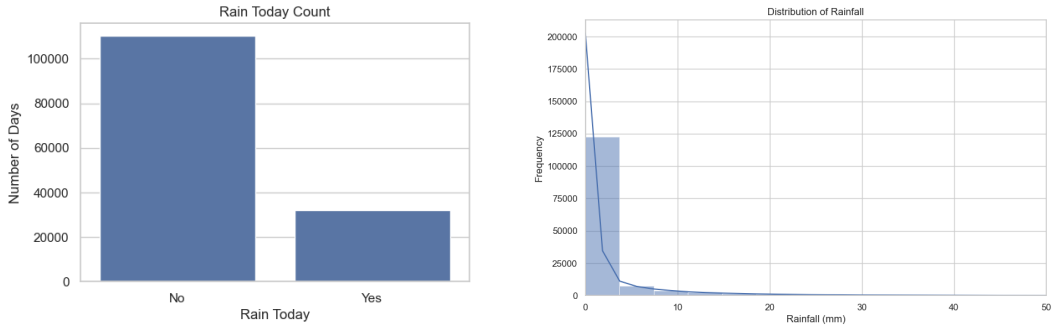

Figure 2: Left: Rain day count; Right: Distribution of Rainfall

Exploring the monthly average rainfalls revealed seasonal variability. Rainfall peaks in late summer and early autumn, with February and March showing the highest averages, followed by a secondary peak in June. The driest period occurs in late winter and early spring, particularly in September and October. This pattern reflects Australia's climatic diversity, where northern regions experience wet summers from monsoonal influences, while southern regions receive more winter rainfall from frontal systems. The corresponding plot can be seen in the appendix 3.

The correlation analysis of the numeric weather variables shows several strong linear relationships. Morning and afternoon temperatures are highly correlated with the daily minimum and maximum temperatures, respectively, while morning and afternoon pressure measurements exhibit almost perfect correlation, reflecting daily stability. Humidity levels show moderate correlation between morning and afternoon, and wind gust speeds are moderately correlated with afternoon wind speeds. Negative correlations between temperature and humidity indicate that warmer conditions are generally associated with lower relative humidity. Rainfall shows weak correlations with other features, suggesting it depends on more complex interactions rather than individual variables. The corresponding plot can be seen in the appendix 4.

5 Modeling

For the benchmarking three different methods are used. Firstly the Logistic Regression, secondly the XGBoost and lastly the Bayesian Logistic Regression. The methods were used via the Python Packages *Sklearn*, *xgboost* and *pymc*. In the following, the methods and their suitability for the task are shortly explained.

5.1 Logistic Regression

Logistic regression is a generalized linear model used for modeling binary dependent variables. It is designed to estimate the probability that a particular event occurs as a function of one or more independent variables. This model is particularly suitable for research questions where the outcome is categorical and dichotomous, as in the given context of the research question. Logistic regression is chosen over linear regression because the assumptions of linearity, constant variance (homoscedasticity), and normality of residuals are violated when modeling binary outcomes. Furthermore, logistic

regression guarantees that the predicted probabilities remain in the [0, 1] interval, which is a fundamental requirement for probabilistic interpretation. The logistic regression model transforms a linear combination of predictors through the logit link function, allowing the model to express probabilities of a binary outcome:

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Where:

- p is the probability of the outcome occurring (i.e., $Pr(Y = 1)$),
- β_0 is the intercept,
- $\beta_1, ..., \beta_k$ are the coefficients for the independent variables $x_1, ..., x_k$

This can be rewritten to express the predicted probability directly as:

$$p = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k\right)}$$

The model parameters are estimated using Maximum Likelihood Estimation, which identifies the set of coefficients that maximize the likelihood of observing the given sample data [James et al., 2014, p. 133ff.]. With the explained properties of the method, it suits the binary classification task with multiple independent features well. The outputs of the method can be interpreted as probabilistic, even though the method itself is not. The threshold at which the next day will be predicted as a rain day is 0.5.

5.2 XGBoost

XGBoost represents a general framework for scalable, regularized gradient boosting over decision trees. It integrates algorithmic and systems-level innovations to improve learning efficiency, model generalization, and scalability. Central to its design are a regularized objective that balances model fit and complexity, a principled mechanism for handling sparsity, and an approximate optimization strategy based on weighted quantile sketching. Combined with optimized memory access patterns and parallelization techniques, the method enables efficient training on large-scale and high-dimensional datasets, making it suitable for a broad range of predictive modeling tasks [Chen and Guestrin, 2016].

In the case of binary weather forecasting, XGBoost is suitable, since it's not only applicable to regression tasks but also to classification. Further investigation of the method in the given context shows that XGBoost can help improve short-term precipitation forecasts. In the paper of Dong et al. [2023] the XGBoost model significantly enhances forecast accuracy across various climatic zones in China by integrating multi-variable inputs and applying bias correction simultaneously. This shows the relevance of the method in the context of weather prediction beyond the scope of this project.

Through its tree-based method, XGBoost captures non-linear interactions between features without requiring explicit feature engineering. It also improves the common property of overfitting of tree-based methods by the built-in regularization. Even though XGBoost can produce probabilistic outputs, the method is not inherently probabilistic since it is a deterministic gradient boosting framework that builds an ensemble of decision trees.

5.3 Bayesian Logistic Regression

Bayesian logistic regression applies Bayesian inference to the logistic regression model, where class probabilities are modeled using the logistic sigmoid function. A Gaussian prior is placed over the weight parameters, and the posterior is proportional to the product of this prior and the likelihood. Exact inference is intractable because the likelihood involves a product of sigmoid terms across all data points. To make the problem tractable, a Laplace approximation is used. This involves finding the maximum a posteriori (MAP) estimate of the weights and approximating the posterior distribution by a Gaussian centered at this estimate, with covariance given by the inverse Hessian of the negative log-posterior [Bishop, 2007, 217f.].

Predictions for new inputs are obtained by integrating over the approximate posterior distribution. Since this integration is not analytically solvable, it is approximated using a probit function to exploit its analytical convolution properties with a Gaussian. This yields an efficient approximation to the predictive distribution, enabling calibrated probability estimates and improved uncertainty quantification. While the decision boundary under equal class priors is unchanged from the MAP solution, the Bayesian treatment enhances probabilistic predictions and accounts for parameter uncertainty [Bishop, 2007, 217f.].

Bayesian logistic regression is well-suited for predicting whether it will rain tomorrow because it directly models the probability of a binary outcome using the logistic function. Unlike purely deterministic classifiers, it provides probabilistic predictions, which are particularly valuable in meteorology where uncertainty assessment is crucial. The Bayesian framework incorporates parameter uncertainty into the predictive distribution, resulting in better forecasts, especially in small or noisy datasets. Furthermore, prior distributions allow the integration of domain knowledge, such as known meteorological relationships, while the inherent regularization effect of the prior mitigates overfitting and improves generalization to unseen weather conditions.

6 Model Training and Evaluation

Before the training of the model, the data is split into a training and a test set. The training set contains 80% of the initial data, while the test set contains the remaining 20%. This split is done to prevent the methods from overfitting on the data. It also allows for a better estimation of the classification errors, since the models have not seen the test data in training. In order to be able to compare the methods with each other, the following metrics are used:

- Precision
- Recall
- F1-Measure

Since the metrics were explained in the lecture, no further explanation is required. To validate the findings the McNemar test is used to check the results for significance.

6.1 McNemar test

McNemar's test [McNemar, 1947] is a non-parametric statistical procedure designed to evaluate whether two classifiers exhibit a statistically significant difference in predictive performance when applied to the same dataset. As discussed in Dietterich [1998] comparative study of statistical tests for supervised classification, McNemar's test is particularly well-suited to scenarios in which each classifier is trained and evaluated once on a fixed training–test split, such as in benchmark competitions or when model training is computationally expensive.

The test operates on paired binary outcomes that record, for each instance, whether each classifier's prediction was correct or incorrect. These outcomes are organized into a 2×2 contingency table, with the critical information residing in the off-diagonal elements, n_{01} and n_{10} , representing instances correctly classified by one model and misclassified by the other, and vice versa. The null hypothesis of the test states that the probability of an instance falling into n_{10} is equal to the probability of falling into n_{01}, i.e., that both classifiers have the same error rate. The test statistic is commonly evaluated using a chi-square approximation with one degree of freedom, optionally applying Yates' continuity correction, or via an exact binomial test when $n_{10} + n_{01}$ is small [Dietterich, 1998, p. 1902 ff.].

To improve the statistical significance of the results of this benchmarking, the following three McNemar test are computed:

- XGBoost vs. Logistic Regression
- XGBoost vs. Bayesian Logistic Regression
- Logistic Regression vs. Bayesian Logistic Regression

The test results can be seen in the appendix 2. To reject the null hypothesis a p-value level of 0.05 is chosen. For the first test, the calculated binomial p-value of $3.718 * 10^{-23}$ is well below the

chosen significance level. Consequently, the null hypothesis that the XGBoost and logistic regression classifiers have identical error proportions is rejected. Based on the contingency table results, the difference in performance is statistically significant, with XGBoost demonstrating superior accuracy.

For the second test, the calculated binomial p-value of 0 is well below the chosen significance level. Consequently, the null hypothesis that the XGBoost and Bayesian Logistic regression classifiers have identical error proportions is rejected. Based on the contingency table results, the difference in performance is statistically significant, with XGBoost demonstrating superior accuracy.

For the last test, the calculated binomial p-value of 0 is well below the chosen significance level. Consequently, the null hypothesis that the Logistic Regression and Bayesian Logistic regression classifiers have identical error proportions is rejected. Based on the contingency table results, the difference in performance is statistically significant, with Logistic Regression demonstrating higher accuracy.

7 Results

Using the results of the applied metrics and the McNemar test, the hypotheses from section 2 can be addressed as follows.

7.1 Hypothesis 1

The hypothesis stated that:

- The XGBoost model will outperform a Logistic Regression model in predicting whether it will rain tomorrow based on the historical weather data.

With the statistically significant results of the McNemar test and the better results in respect to the precision ($75.20\% > 72,84\%$), recall ($55.17\% > 47,53\%$) and the F1-Measure ($63.64\% > 57,52\%$), Hypothesis 1 can be approved. The XGBoost does perform better in predicting if it will rain the next day or not, based on the data of the current day.

7.2 Hypothesis 2

The hypothesis stated that:

- The Logistic Regression model will outperform both the Bayesian Logistic Regression in terms of prediction accuracy.

With the statistically significant results of the McNemar test between the Logistic Regression and the Bayesian Logistic Regression and the better results in respect to the precision ($72,84\% > 27.40\%$), recall ($47,53\% < 58.20\%$) and the F1-Measure ($57.52\% > 37,20\%$), Hypothesis 2 can be approved, even though the recall of the Bayesian Logistic Regression is better. The Logistic Regression does perform better in predicting if it will rain the next day or not, based on the data of the current day.

7.3 Hypothesis 3

- XGBoost model will outperform both the Bayesian Logistic Regression in terms of prediction accuracy.

With the statistically significant results of all of the first 2 and last McNemar test 9 and the better results in respect to the precision, recall and the F1-Measure, Hypothesis 3 can be approved, even though the recall of the Bayesian Logistic Regression is better. The XGBoost does perform better in predicting if it will rain the next day or not, based on the data of the current day.

8 Discussion

Based on the benchmarking results, XGBoost emerged as the best-performing method for the given prediction task. However, this outcome should be interpreted with caution, as the model's performance, while superior to the alternatives, is not flawless. XGBoost achieved a precision

of 75.20%, indicating that most of the days it predicted as rainy were correctly classified. Its recall of 55.17% shows that it successfully identified more than half of the actual rain days in the dataset. The resulting F1-score of 63.64% reflects a balanced trade-off between precision and recall, demonstrating robust performance in both detecting rain events and limiting false alarms. Overall, these results suggest that XGBoost is highly effective for this binary classification task, outperforming the baseline models by maintaining strong accuracy while capturing a substantial proportion of true rain occurrences.

When comparing Logistic Regression to Bayesian Logistic Regression, a notable trade-off between recall and precision becomes apparent. The Bayesian variant achieves a higher recall (58.20% vs. 47.53%), indicating that it correctly identifies a larger proportion of actual rain days. This improvement can be attributed to the Bayesian model's incorporation of parameter uncertainty, which produces softer decision boundaries and more liberal positive predictions. The effect is amplified by the dataset's class imbalance, where rain days represent a clear minority. In such settings, standard Logistic Regression tends to bias toward the majority class ("no rain"), which increases overall accuracy but suppresses recall for the minority class. Bayesian Logistic Regression, by contrast, is less conservative in its predictions, thereby detecting more rain days but at the expense of precision, as the same tendency increases the number of false positives. This reflects a fundamental precision–recall trade-off shaped in part by the imbalance in the target variable.

Another factor to consider is the lack of systematic hyperparameter tuning for the Bayesian Logistic Regression. While the other two methods achieved acceptable results with their default configurations, the Bayesian model would likely benefit from more extensive parameter optimization. However, given the substantial computational cost of the method (approximately one hour per run), a comprehensive hyperparameter search was not conducted. It is therefore plausible that, with appropriate tuning, the Bayesian Logistic Regression could achieve performance comparable to that of standard Logistic Regression in terms of accuracy.

To identify a state-of-the-art method for predicting whether the next day will be rainy, more advanced approaches need to be explored. Consequently, this research project offers only a limited knowledge gain, which should be expanded. Given that day-to-day rain prediction may not be of highest priority, redefining the research objective could enhance its relevance. As extreme weather events are expected to become more frequent due to climate change, focusing on predicting such events over a one to two-week horizon could present a challenging yet highly impactful goal in terms of usability and real-world benefit. Another relevant research question could examine the factors underlying XGBoost's superior performance compared to the other methods, as well as identify potential modifications to these methods that could improve their predictive accuracy.

9 Conclusion

This project report for the module *Probabilistic Machine Learning* evaluated three classification methods, namely, Logistic Regression, XGBoost, and Bayesian Logistic Regression for predicting next-day rainfall in Australia based on historical weather observations. The results demonstrate that XGBoost achieved the highest overall performance, with superior precision, recall, and F1-score, and statistically significant improvements over both baseline models according to McNemar's test. Logistic Regression outperformed Bayesian Logistic Regression in terms of accuracy and F1-score, though the Bayesian approach achieved higher recall, reflecting its greater sensitivity to rain events at the expense of precision. The findings highlight the suitability of gradient boosting methods for capturing complex, non-linear relationships in meteorological data, while also underscoring the trade-offs between recall and precision inherent in different modeling paradigms. Nevertheless, the absolute performance levels indicate room for improvement, suggesting that future work should explore systematic hyperparameter tuning, additional feature engineering, and more advanced algorithms. Moreover, shifting the focus toward predicting extreme weather events over longer lead times could enhance the practical relevance and societal impact of such predictive systems.

References

ABC News. Flooded coffs region declared disaster area, 2009. URL https://www.abc.net.au/news/2009-11-08/flooded-coffs-region-declared-disaster-area/1133200. Accessed: 2025-07-31.

Australian Bureau of Meteorology. Climate classification, 2023. URL http://www.bom.gov.au/climate/maps/averages/climate-classification/. Accessed: 2025-07-31.

Australian Institute for Disaster Resilience (AIDR). Queensland flood, 2010–2011, 2011. URL https://knowledge.aidr.org.au/resources/flood-queensland-2010-2011/. Accessed: 2025-07-31.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007. ISBN 0387310738.

Gustau Camps-Valls, María Ángeles Fernández-Torres, Klaus H. Cohrs, et al. Artificial intelligence for modeling and understanding extreme weather and climate events. *Nature Communications*, 16:1919, 2025. doi: 10.1038/s41467-025-56573-8. URL https://doi.org/10.1038/s41467-025-56573-8.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM, August 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 10 1998.

Jianhua Dong, Wenzhi Zeng, Lifeng Wu, Jiesheng Huang, Thomas Gaiser, and Amit Kumar Srivastava. Enhancing short-term forecasting of daily precipitation using numerical weather prediction bias correcting with xgboost in different regions of china. *Engineering Applications of Artificial Intelligence*, 117:105579, 2023. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2022.105579. URL https://www.sciencedirect.com/science/article/pii/S0952197622005693.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN 1461471370.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996.

Appendix

Table 1: Features in the weatherAUS Dataset

Feature	Description
Date	Date of the observation
Location	Location of the weather station
MinTemp	Minimum temperature (°C) recorded that day
MaxTemp	Maximum temperature (°C) recorded that day
Rainfall	Amount of rainfall (mm) in the 24 hours
Evaporation	Evaporation (mm) in the 24 hours to 9am
Sunshine	Sunshine (hours) of bright sunshine in the day
WindGustDir	Direction of strongest wind gust
WindGustSpeed	Speed (km/h) of strongest wind gust
WindDir9am	Wind direction at 9am
WindDir3pm	Wind direction at 3pm
WindSpeed9am	Wind speed (km/h) at 9am
WindSpeed3pm	Wind speed (km/h) at 3pm
Humidity9am	Relative humidity at 9am
Humidity3pm	Relative humidity at 3pm
Pressure9am	Atmospheric pressure (hPa) at 9am
Pressure3pm	Atmospheric pressure (hPa) at 3pm
Cloud9am	Fraction of sky obscured by cloud at 9am (0–8 scale)
Cloud3pm	Fraction of sky obscured by cloud at 3pm (0–8 scale)
Temp9am	Temperature (°C) at 9am
Temp3pm	Temperature (°C) at 3pm
RainToday	Whether there was rain today (Yes/No)
RainTomorrow	Whether there will be rain tomorrow (target variable)

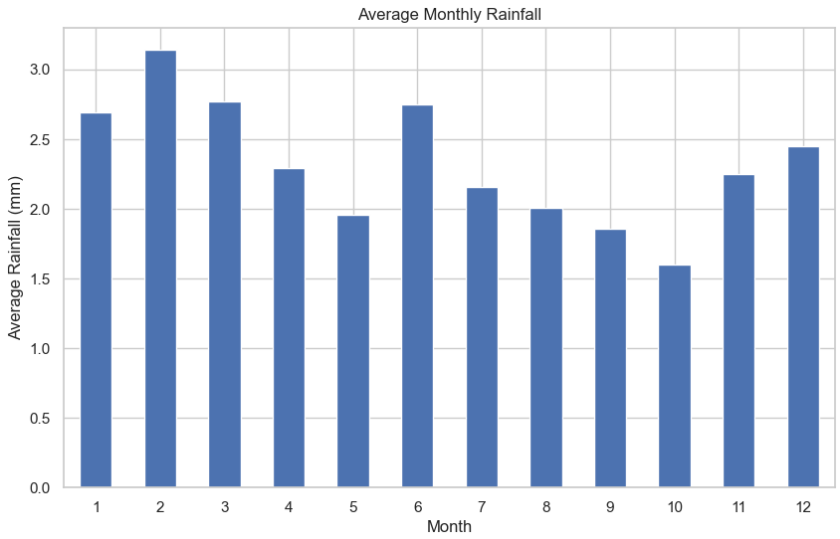

Figure 3: Seasonal rainfall in australia

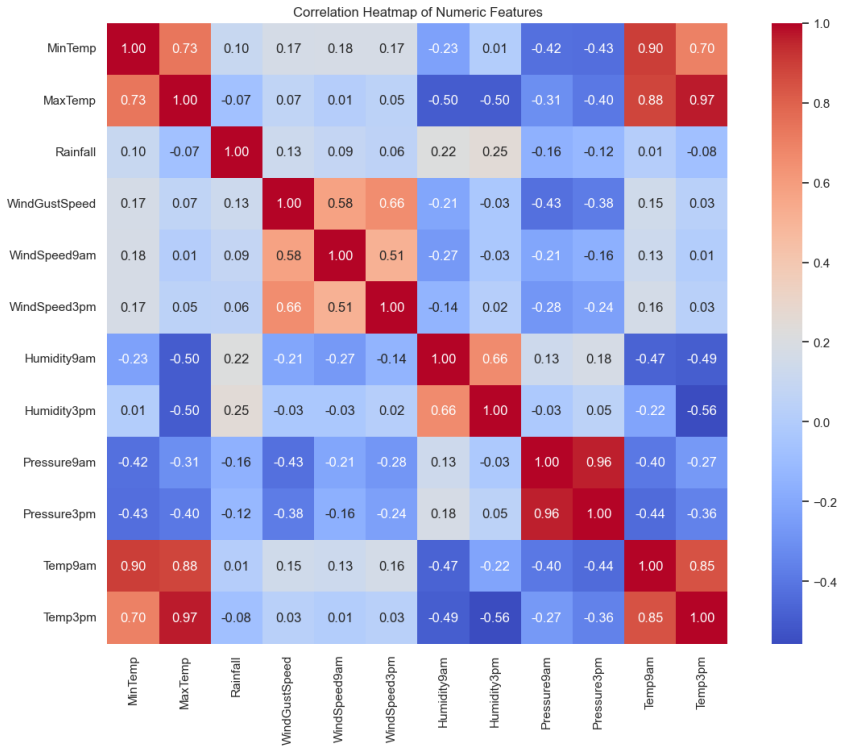

Figure 4: Correlation matrix of numerical features

Table 2: McNemar's test results comparing XGBoost and Logistic Regression. n_{10} = XGBoost correct, Logistic Regression wrong; n_{01} = XGBoost wrong, Logistic Regression correct.

	Logistic Regression correct	**Logistic Regression wrong**	**Total**
XGBoost correct	23,126	$n_{10} = 1{,}295$	24,421
XGBoost wrong	$n_{01} = 838$	3,180	4,018
Total	23,964	4,475	28,439

Test statistic: $\chi^2 = 97.49$ (Yates corrected)
Asymptotic p-value: 5.426×10^{-23}
Exact binomial p-value: 3.718×10^{-23}
Interpretation: Significant difference; XGBoost outperforms Logistic Regression.

Table 3: McNemar's test results comparing XGBoost and Bayesian Logistic Regression. n_{10} = XGBoost correct, Bayesian Logistic Regression wrong; n_{01} = XGBoost wrong, Bayesian Logistic Regression correct.

	Bay. Logistic Regression correct	**Bay. Logistic Regression wrong**	**Total**
XGBoost correct	13,892	$n_{10} = 10{,}529$	24,421
XGBoost wrong	$n_{01} = 2{,}046$	1,972	4,018
Total	15,938	12,501	28,439

Test statistic: $\chi^2 = 5721.22$ (Yates corrected)
Asymptotic p-value: 0
Exact binomial p-value: 0
Interpretation: Significant difference; XGBoost outperforms Bayesian Logistic Regression.

Table 4: McNemar's test results comparing Logistic Regression and Bayesian Logistic Regression. n_{10} = Logistic Regression correct, Bayesian Logistic Regression wrong; n_{01} = Logistic Regression wrong, Bayesian Logistic Regression correct.

	Bay. Logistic Regression correct	**Bay. Logistic Regression wrong**	**Total**
Logistic Regression correct	13,567	$n_{10} = 10{,}397$	23,964
Logistic Regression wrong	$n_{01} = 2{,}371$	2,104	4,475
Total	15,938	12,501	28,439

Test statistic: $\chi^2 = 5043.91$ (Yates corrected)
Asymptotic p-value: 0
Exact binomial p-value: 0
Interpretation: Significant difference; Logistic Regression outperforms Bayesian Logistic Regression.

Table 5: Comparison of the used methods

Performance Metrics Comparison

Model	**Precision (%)**	**Recall (%)**	**F1-score (%)**
XGBoost	75.20	55.17	63.64
Logistic Regression	72.84	47.53	57.52
Bayesian Logistic Regression	27.40	58.20	37.20